The Social License to Lie: When Synthetic Media Destroys the Fourth Estate's Credibility

Weekly Analysis — https://ainews.social

There is a particular kind of power that comes not from being believed but from making belief itself unaffordable. The deepfake's deepest threat was never that a forged video of a senator would fool us; it was that the *existence* of forgery would give every genuine recording an exit ramp. Once anyone can manufacture a convincing lie, anyone caught in a true recording can claim it was manufactured. This is what political scientists have started calling the liar's dividend, and it is already paying out. The 2024 AI Index Report warned that deepfake tools had improved dramatically since 2020 and that large-scale synthetic disinformation could "undermine trust in democratic institutions, manipulate public opinion, and polarize public discussions" [2]. What that flat institutional prose understates is the redistribution underneath it: the cost of establishing what is real is being pushed downward and outward, off the balance sheets of the firms that built the generative tools and onto the rest of us.

[2] AI Index Report 2024

This essay is about that redistribution, and about a question the discourse around synthetic media keeps declining to ask directly: who has the power here, and who pays for it? The standard framing treats deepfakes as a technical problem awaiting a technical fix—better detectors, watermarks, provenance metadata. But the technical framing conceals a political one. The burden of detection is being socialized while the capacity to deceive is being privatized and sold. Newsrooms must now authenticate every image in real time. Courts must adjudicate evidence whose chain of custody no existing doctrine anticipated. And ordinary citizens, told to become amateur forensic analysts, are being handed responsibility for a verification labor that the companies profiting from generative AI have no intention of performing themselves. The collapse of shared evidentiary standards is not a side effect. It is the predictable result of who got to build, who gets to profit, and who gets left holding the question *is this real?*

The Burden Nobody Volunteered For

Start with the tools sold as the solution, because they reveal the asymmetry most cleanly. When institutions panicked about machine-generated text, a market for detection software appeared almost overnight, promising to restore certainty. It did the opposite. Across American school districts, detectors flagged so many genuine submissions as fake that administrators began pulling the tools entirely, because the false-positive rate was tainting innocent people faster than it caught real fraud [13]. One writer described the experience from the accused side: software declares a person guilty of fabrication, the person insists they are innocent, and there is no neutral arbiter to settle it—only an algorithm's confidence score against a human's word [15]. Whole universities concluded the detectors were unfit for use and banned them outright [17].

[13] Schools are racing to catch AI-written homework — but the detectors flag so many innocent students that some districts are banning the tools outright

[15] The software says my student cheated using AI. They say they're not. Who do I believe?

[17] Universities That Banned AI Detectors: 2026 Full List

Hold onto what that episode demonstrates, because it scales far beyond any classroom. The detection apparatus we are being told will protect public truth is itself unreliable, and its unreliability does not fall evenly. A false positive is not a neutral error; it is an accusation. And the discourse around these tools rarely centers the accused—the person who must now prove a negative, prove that the words or the image or the recording are authentically theirs. This is the structure of the whole synthetic-media crisis in miniature: the power to assert "fake" is cheap and widely distributed, while the power to *establish* "real" is expensive, slow, and concentrated. The burden of detection was never agreed to by the people now carrying it. It was assigned to them by default, because the alternative—holding the producers of the technology responsible for its outputs—was never seriously on the table.

The verification labor that detection actually requires, when it is done well, is human and it is hidden. Behind the promise of automated authenticity sits an enormous, deliberately invisible workforce of people who label data, moderate content, and clean up the outputs of systems marketed as autonomous. The AI Ethics volume in the MIT Press series is blunt about this: human labor is "hidden behind the scenes—miners, workers on ships, click workers who label data sets, all in the service of capital accumulation by very few people," and the users of these systems "are also more vulnerable than others" [2]. When we say the burden of detection is being socialized, we should be precise about which part of society absorbs it—and notice that the people doing the hardest verification work are the least audible in the conversation about it.

[2] AI Ethics - The MIT Press Essential Knowledge series

The Speed Trap and the News Cycle

The Fourth Estate sits at the exact pressure point where this redistribution bites hardest, because journalism's value proposition is speed *and* accuracy, and synthetic media forces a trade between them. A newsroom that must now authenticate every photograph, every clip, every leaked recording before publication has absorbed a cost that did not exist a decade ago—and absorbed it precisely as the economics of news were already gutted. The result is a structural squeeze: verify thoroughly and lose the timeliness that defines the news cycle, or move at the old speed and risk amplifying a forgery that will, when exposed, be used to discredit everything else the outlet has ever published.

This is not hypothetical. The AI Index documented an automated propaganda pipeline in which a system generates an article attributed to a fake journalist, a second system writes comments to simulate organic engagement, and a third searches social media for relevant posts and replies in the guise of ordinary users—an entire synthetic ecosystem of manufactured credibility, running without a human in the loop [2]. The point of such a system is not any single lie. It is to flood the channel until the marginal cost of distinguishing signal from noise exceeds what any reader, or any newsroom, can pay. And the firms that build the generative models that make this possible bear none of that downstream cost. They externalize it onto the institutions whose credibility is the actual target.

[2] AI Index Report 2024

Notice who shapes the conversation about solutions. The companies amplified in the discourse are the same ones selling the next layer of the stack—the detection service, the provenance API, the watermarking standard. We have arrived, through technology, at something close to the dynamic Edward Herman and Noam Chomsky described decades ago, in which the flow of news is shaped less by what is true than by the filters of ownership, advertising, and sourcing [2]. Synthetic media does not break that model; it supercharges it, by making the sourcing filter—who gets treated as a credible source—into a technical arms race that favors whoever can afford the most verification infrastructure. The independent outlet, the local reporter, the journalist in a country without a domestic AI industry: these are the voices that go quiet when authentication becomes a capital expense. The structural silence here is geographic and economic. The discourse about deepfakes and journalism is conducted almost entirely in the languages and institutions of the wealthy North, while the newsrooms least able to afford real-time authentication—and most exposed to state-sponsored synthetic disinformation—are barely present in the room.

[2] Manufacturing Consent

What makes this especially corrosive is that the credibility of synthetic content is now demonstrably good enough to pass expert review. In one striking case, an AI system produced a complete scientific study, and the resulting article was actually evaluated by human researchers before the deception was caught [16]. If the apparatus of peer review—the gold standard of institutional verification—can be fooled, the casual reader scrolling a feed has no chance. The expectation that individuals will simply "be more critical" is not a solution; it is an abdication dressed as empowerment.

[16] Une IA a produit une étude scientifique complète, et son article a même été évalué par des chercheurs

What the Courtroom Cannot Yet See

If journalism is where synthetic media corrodes public belief, the courtroom is where it corrodes something even more foundational: the legal system's assumption that evidence can be trusted to mean what it appears to mean. Photographs, audio recordings, and video have functioned in law as a kind of frozen testimony—mechanically captured, presumptively reliable, harder to impeach than a witness's memory. Synthetic media dissolves that presumption. Chain-of-custody doctrines were built to track the *handling* of evidence, to ensure a recording was not tampered with between seizure and trial. They were never designed to answer whether the recording depicts an event that occurred at all.

This is a power vacuum, and power vacuums get filled by whoever moves first. In the absence of updated standards, the advantage flows to the party that can afford expert forensic testimony and to the party willing to weaponize doubt. The liar's dividend operates with brutal efficiency in an adversarial setting: a defendant captured on genuine video can now gesture at the mere possibility of deepfakery and shift the burden onto the prosecution to prove a negative. The *accusation* of fakery costs nothing; the *refutation* costs a forensic budget. We have already seen, in the detection-software controversies, how an algorithm's verdict can override a human's sworn account of their own conduct [15]. Transpose that dynamic into a setting where liberty is at stake, and the stakes of getting verification wrong stop being abstract.

[15] The software says my student cheated using AI. They say they're not. Who do I believe?

There is a deeper inequality embedded here, because the forensic tools courts will rely on inherit the biases of the systems that built them. We know that AI systems encode and amplify the prejudices of their training data; the largest study to date of AI hiring algorithms found "clear racial disparities," with more than a quarter of Black applicants disadvantaged by bias [9]. Research from across Latin America documents how these systems reproduce gender, racial, and

[9] Largest study of AI hiring algorithms to date finds 'clear racial disparities' — over 25% of Black applicants tainted by bias

xenophobic bias as a structural feature rather than a bug [7]. There is no reason to expect deepfake-detection tools to be exempt. A forensic authentication system that performs worse on certain faces, accents, or languages would import discrimination directly into the determination of what counts as real evidence—and it would do so under a veneer of mechanical objectivity that makes the bias harder to contest. The people most likely to be misjudged by such a system are, predictably, the people least represented in the rooms where it is designed.

[7] Género, racismo y xenofobia: así son los sesgos de la Inteligencia Artificial en Latinoamérica

What *After Shock* captures about this moment is the epistemic vertigo underneath the procedural questions: challenges to our ability to discern what is authentic "fundamentally impact our understanding of the world and the future" [2]. A legal system runs on a shared agreement that certain kinds of evidence settle certain kinds of disputes. Remove that agreement and you do not get a more careful jurisprudence; you get a contest of resources, in which the capacity to manufacture and to refute doubt becomes another asset that the powerful hold and the powerless lack.

[2] After shock

Who Builds the Lie, Who Takes the Blame

Here is the question the discourse works hardest to avoid: when synthetic media degrades public trust, who is held accountable? The answer, almost always, is everyone except the firms that built and profit from the generative tools. The blame is distributed to "misinformation," to "bad actors," to gullible audiences who failed to think critically, to platforms that failed to moderate—a diffuse cloud of responsibility that conveniently never settles on the companies whose products made industrial-scale forgery cheap and accessible. Kate Crawford names this evasion precisely: tech companies "rarely suffer serious financial penalties when their AI systems violate the law and even fewer consequences when their ethical principles are violated" [2]. The accountability gap is not an oversight. It is the achieved outcome of a discourse that frames synthetic media as a force of nature rather than a product with manufacturers.

[2] The Atlas of AI - Power, Politics, and the Planetary Costs

Follow the labor and the asymmetry becomes a map. The generative systems that produce convincing fakes, and the moderation systems meant to catch them, both rest on a workforce concentrated in the Global South and paid a fraction of what the value they create is worth. Kenyan workers who took AI-related jobs believing they had "tickets to the future" instead found themselves traumatized and underpaid, doing the psychological dirty work of training and filtering these systems [8]. Swiss public broadcasting documented the

[8] Kenyan workers with AI jobs thought they had tickets to the future. Then the work took a darker turn

same pattern across the industry: the ”exploits” of AI rest on invisible workers exploited in poor countries [4]. Analysts of data labor in the Global South describe a structure in which the cognitive work of making AI function is extracted from the periphery while the profits and the decision-making accrete at the center [3]. The Brookings Institution has called for reimagining this arrangement precisely because, as currently built, it reproduces a colonial division of labor under a digital veneer [12].

[4] Derrière les prouesses de l'IA, l'exploitation de travailleurs invisibles dans les pays pauvres

[3] Data Labeling in the Global South

[12] Reimagining the future of data and AI labor in the Global South

That word—colonial—is not rhetorical excess. A growing body of analysis frames the global AI economy as digital colonialism: a systemic dependency in which the infrastructure, the standards, and the value flow toward a handful of firms and nations while the rest of the world supplies raw labor and raw data [5]. Africa's position in the AI supply chain has been analyzed as a question of infrastructure control, with sovereignty resting less on rhetoric than on who owns the physical and computational layers [1]. French-language analysis makes the same point about machine learning broadly: the age of AI is reproducing colonial relations of extraction and dependence [10].

[5] El colonialismo digital en la era de la IA: siete dimensiones de una dependencia sistémica

[1] Africa's Role in the AI Supply Chain: Why Infrastructure Control Matters More Than Digital Sovereignty

[10] Le colonialisme numérique à l'ère de l'IA et de l'apprentissage machine

Why does this matter for the credibility of the press and the courts? Because it explains *who gets to define authenticity for everyone else*. When the verification layer is built and controlled by the same concentration of power that built the deception layer, the people whose voices, faces, and recordings will be authenticated—or impeached—have no seat at the table where the standards are set. The structural silence is total: the workers labeling the data, the populations whose languages and faces are underrepresented in training sets, the newsrooms and courts of countries without a domestic AI industry—none of them shape the discourse, and all of them inherit its consequences. The dominance of ”ethical failures” as a topic in AI discussion is itself revealing. We document harm endlessly. The documentation has become a genre. But documentation is not accountability, and the gap between the lavish attention paid to cataloguing harms and the meager attention paid to building enforceable remedies is exactly where corporate power lives comfortably.

The Provenance Bargain and Its Price

So we arrive at the proposed grand solution, the one offered as the way out of the collapse: institutional verification layers, built to scale—provenance metadata, cryptographic content authentication, a chain of custody for reality itself. The honest version of the stakes admits what this requires. To prove that a piece of media is authentic,

you generally have to prove *where it came from*, which means attaching identity and origin to content at the moment of capture. Scaling verification means scaling provenance, and scaling provenance means, at some point, ceding privacy. This is the bargain hiding inside every clean technical proposal: the price of provenance transparency is paid in surveillance.

We should be exceptionally skeptical of who is offering this bargain and who will be bound by it. The same architecture that authenticates a journalist's footage can track a dissident's movements; the same metadata that lets a court trust a recording lets a state map who was where, when, and with whom. We have a preview of how these systems behave when deployed on the relatively powerless. Surveillance infrastructure marketed as protective has already reshaped public institutions: monitoring technology installed in the name of safety has turned public schools into sites of pervasive observation, with private vendors watching students continuously [11]. Students and families have gone to court alleging that this digital monitoring is unconstitutional and conducted without meaningful consent [14]. Researchers examining emotion-recognition systems deployed to monitor people's affective states have warned that such tools cross ethical lines while resting on shaky scientific ground [6]. The pattern is consistent: surveillance infrastructure introduced to solve a legitimate problem becomes permanent, expands beyond its stated purpose, and falls heaviest on those with the least power to refuse it.

[11] Public Schools, Private Eyes: How EdTech Monitoring Is Reshaping Public Schools

[14] Students allege continued unconstitutional AI digital monitoring and violations of Open Records Act in school district lawsuit

[6] Emotion AI in the classroom: ethics of monitoring student affect

A provenance regime for all media would be surveillance infrastructure of unprecedented reach, and the question of who controls it is the whole game. If the verification layer is owned by the same firms that escape accountability for the deception layer, we will have ceded our privacy to the very actors who created the crisis the privacy is meant to resolve—paying twice, once in trust and once in liberty, while they collect on both ends. *After Shock* points at the consumer-profiling logic that already drives so much of this: the project of building "personas" that "neatly describe large categories of people in terms of their beliefs, passions, and motivations" [2]. A universal provenance system is, among other things, a universal profiling system. Whether it serves the public or the platform depends entirely on a governance question that the technical discourse is structured to never quite reach.

[2] After shock

None of this is an argument against verification. The collapse of shared evidentiary standards is real, and a society that cannot agree on what is real cannot deliberate, adjudicate, or govern. The argument is that the *terms* of verification are a political settlement disguised as an engineering specification, and that the settlement currently being drafted hands the most power to the actors with the least claim to it.

A provenance layer built and governed democratically—transparent, contestable, with the people whose authenticity is being adjudicated holding real authority over the standards—would be one thing. A provenance layer built by and for the firms that profit from both the lie and its cure is another. The discourse treats them as the same proposal. They are not.

The Standard We Are About to Lose

Step back from the specific institutions—the newsroom, the courtroom, the verification API—and the shape of the whole becomes visible. Synthetic media's threat to the Fourth Estate is not fundamentally a threat of being deceived. It is a threat of redistribution: of pushing the cost of establishing truth downward onto individuals and institutions least able to bear it, while the capacity to manufacture falsehood is concentrated, productized, and sold by actors who face no consequences for the corrosion they enable [2]. The "social license to lie" is not a license everyone holds equally. It is held most securely by those who can afford to manufacture doubt and to refute it, and it is held least by everyone now drafted into the unpaid, untrained, unending work of figuring out what is real.

[2] The Atlas of AI - Power, Politics, and the Planetary Costs

Watch the moves as the discourse develops, because they recur. When a deepfake causes harm, notice how quickly responsibility diffuses into the passive voice—"misinformation spread," "trust eroded"—and how rarely it lands on a named company with a balance sheet. Notice that the people asked to become amateur forensic analysts are never the people who built the tools that made forensics necessary. Notice that the workers who actually perform the world's verification labor, in Nairobi and Manila and across the Global South, are absent from every panel about the future of trust [8]. Notice that the proposed cures arrive bundled with surveillance, and that the surveillance is always pointed at the governed rather than the governing [11].

[8] Kenyan workers with AI jobs thought they had tickets to the future. Then the work took a darker turn

[11] Public Schools, Private Eyes: How EdTech Monitoring Is Reshaping Public Schools

The deepest casualty is not any particular truth but the shared standard that lets strangers agree on what counts as evidence—the quiet civic agreement that a recording, a document, a photograph can settle a dispute between people who otherwise trust nothing about each other. *After Shock* is right that challenges to our ability to discern the authentic "fundamentally impact our understanding of the world and the future" [2]. That standard was never a natural fact; it was an institutional achievement, maintained by the credibility of a press that did the verification work and the integrity of courts that

[2] After shock

adjudicated it. Synthetic media does not merely attack that achievement. It privatizes the means of repairing it, and then offers to sell the repair back to us at the price of our privacy and on terms set by the firms that profited from the damage.

The educated reader's task, then, is not to become a better lie-detector—that is the abdication, the burden quietly transferred. It is to keep asking the question the technical framing is built to suppress: who decided that the cost of truth would fall here, on us, and not there, on them? The collapse of evidentiary standards can be answered by institutional verification that scales. But whether that verification serves a democratic public or entrenches the power of those who already shape the discourse is not a question of engineering. It is a question of who gets to speak, who is kept silent, and who is made to pay—and on current trajectory, the answer to all three is being decided by the people with the least reason to decide it well.

References

1. Africa's Role in the AI Supply Chain: Why Infrastructure Control Matters More Than Digital Sovereignty
2. After shock
3. Data Labeling in the Global South
4. Derrière les prouesses de l'IA, l'exploitation de travailleurs invisibles dans les pays pauvres
5. El colonialismo digital en la era de la IA: siete dimensiones de una dependencia sistémica
6. Emotion AI in the classroom: ethics of monitoring student affect
7. Género, racismo y xenofobia: así son los sesgos de la Inteligencia Artificial en Latinoamérica
8. Kenyan workers with AI jobs thought they had tickets to the future. Then the work took a darker turn
9. Largest study of AI hiring algorithms to date finds 'clear racial disparities' — over 25% of Black applicants tainted by bias
10. Le colonialisme numérique à l'ère de l'IA et de l'apprentissage machine
11. Public Schools, Private Eyes: How EdTech Monitoring Is Reshaping Public Schools

12. Reimagining the future of data and AI labor in the Global South

13. Schools are racing to catch AI-written homework — but the detectors flag so many innocent students that some districts are banning the tools outright

14. Students allege continued unconstitutional AI digital monitoring and violations of Open Records Act in school district lawsuit

15. The software says my student cheated using AI. They say they're not. Who do I believe?

16. Une IA a produit une étude scientifique complète, et son article a même été évalué par des chercheurs

17. Universities That Banned AI Detectors: 2026 Full List