Print Edition

Friday 14 August 2026

EDITION NO. 22 · 2 PIECES · 10 PAGES

IN THIS EDITION

AI Content Must Now Carry a Label. Cameras Are Next.

BY AI FRONTIERS · AI FRONTIERS · 13 AUGUST 2026

Eddan Katz, Head of Policy & Gov Affairs at Encypher — August 13, 2026

Two days before Slovakia's 2023 parliamentary election, a fabricated audio clip spread widely on social media, appearing to catch a leading candidate and a prominent journalist discussing how to rig the vote. The following year, nonconsensual pornographic deepfakes of Taylor Swift amassed over 45 million views on X in less than 24 hours. In summer 2025, an elderly woman in Ontario lost her life savings to a cryptocurrency scam, which began with a Facebook advertisement featuring a deepfake of Prime Minister Mark Carney.

In each of these cases, harm might have been reduced by **content provenance metadata**: information attached to content that declares how it was produced. (Generally, the term "provenance" refers to a record of where something came from and what happened to it along the way.) Content provenance tools can't directly prove whether a bare image or video is real or synthetic; no technology reliably can. Instead, they serve as an evidentiary layer of AI governance, producing records that other protec-

tions build on top of. As signed provenance becomes the norm for content captured on cameras and microphones, fabricated material—with no or suspicious provenance—will attract more scrutiny.

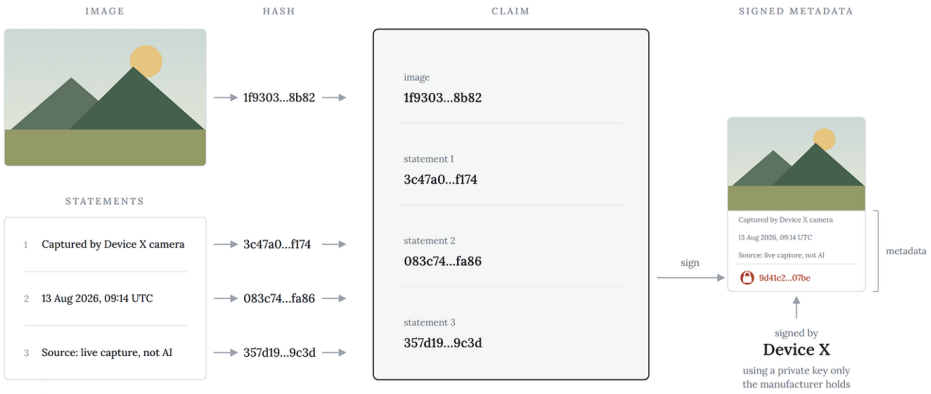
New EU and California laws mandate content provenance. New legal requirements, rolled out in two major markets on the same day, will significantly increase adoption of provenance techniques across the AI ecosystem. On August 2, 2026, both the EU AI Act's Article 50 and California's AI Transparency Act took effect. These regulations expand the creation and preservation of digital records that make it easier to trace a given piece of content back to the cameras, AI models, and other tools that created or captured it.

This essay explains what content provenance is, how it can be implemented, and its benefits and drawbacks. Then, I'll cover the two laws that have taken effect, and their global impact. As a disclosure: I lead policy at Encypher, which develops infrastructure for content provenance.

What Content Provenance Is and How It Works

The new EU and California laws require some companies—including camera manufacturers and AI companies—to attach certain metadata to their products' outputs. Generally speaking, that metadata falls into three main categories: first, origination details describe where the output came from, such as which camera photographed it. Second, chain-of-custody information describes which platforms or tools handled the content after it was made. Third, a modification history tracks any changes made, such as cropping or face-swapping.

The compliance ecosystem is converging on a technological standard: C2PA. The Coalition for Content Provenance and Authenticity (C2PA), whose user-facing implementation is known as Content Credentials, maintains an open, royalty-free standard that many companies will likely use to comply with the EU and California laws. Neither law requires C2PA by name, but its interoperability and growing adoption make it the leading compliance option. Anthropic, Google, and OpenAI have all indicated they will adopt C2PA.

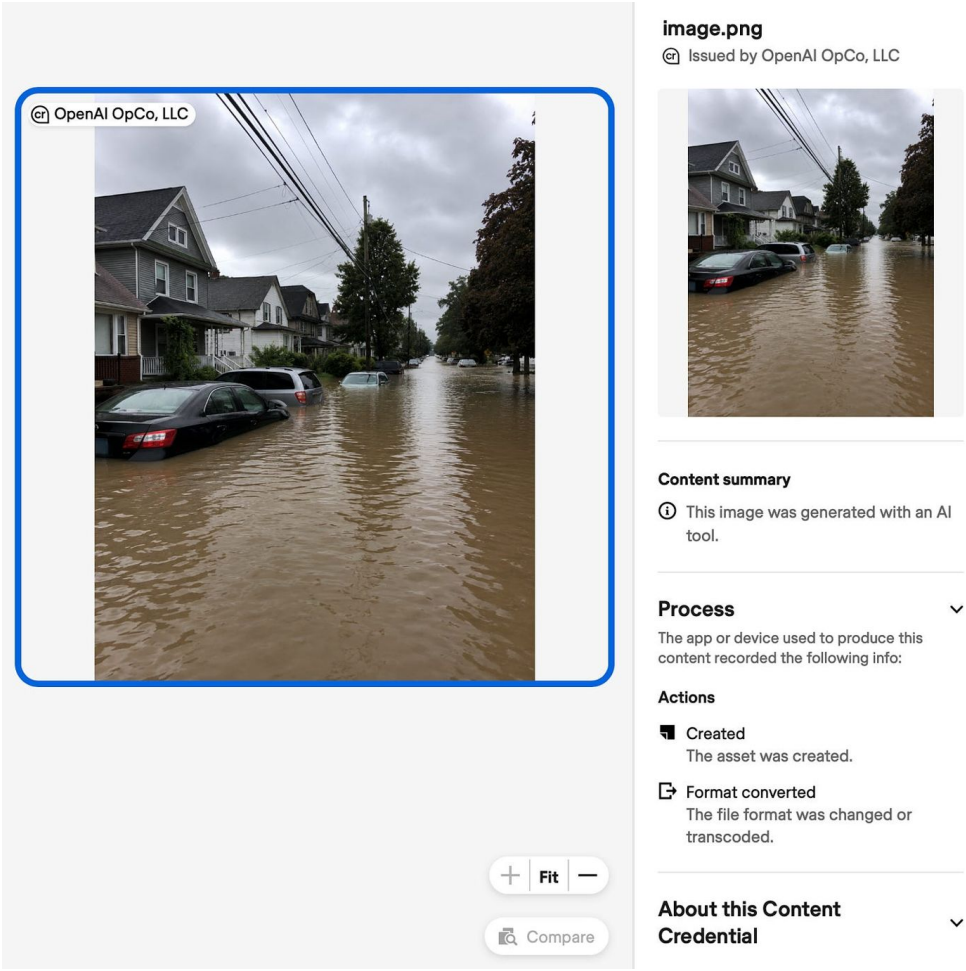


A simplified representation of how Content Credentials metadata is generated. The original data, such as a raw image, and statements about it are hashed and combined into a set of numbers called a "claim." Hardware manufacturers embed trusted keys in their devices to "sign" claims. With Content Credentials, outside observers can easily verify that the data and statements were signed by the hardware manufacturer's key.

C2PA creates provenance records using cryptographic techniques. For example, suppose that someone takes a photo with their smartphone. First, the phone "hashes" the photo's pixels and provenance metadata, generating numerical tags uniquely tied to the data. Then it generates a cryptographic "signature" linked to these hashes and to the smartphone manufacturer. It stores the hashes and signature in the photo's metadata. As a result, tampering typically leaves clear evidence: if anyone modifies the photo or its provenance metadata, a quick computational check can detect that a change has occurred. These same methods apply to audio and video files.

Edits create new provenance records that reference earlier records. In C2PA parlance, each version of the content gets its own "manifest"—a cryptographically signed provenance record describing the latest changes. Continuing the example above, if someone later modifies the photo in Adobe Photoshop, then Adobe can create a new manifest describing the edits it applied, storing this manifest in the photo's metadata. Before signing, Adobe hashes the prior manifest and adds that hash to the new manifest, so a quick computational check can detect tampering with the prior manifest.

A provenance record can declare that content is AI-generated. Under C2PA, provenance metadata contains a field labeled "digitalSourceType," which states whether the content was captured by a camera or microphone, generated by AI, or a combination of the two (such as a photograph with an AI-generated background element inserted in Photoshop). The EU and California requirements lean heavily on this information.



A tool provided by the Content Authenticity Initiative displays C2PA metadata from an image generated in ChatGPT, revealing its true provenance.

Content Credentials show up as small, familiar cues. Creators will see a simple toggle to "attach Content Credentials" in Photoshop or Firefly, or their camera will sign photos automatically. Viewers may see a small "cr" badge on the content, which they can click to see the metadata. Platforms will see machine-readable data they can turn into on-screen labels.

Limitations of Content Credentials

While Content Credentials are a promising way of fulfilling the new EU and California requirements, they are not infallible. They ultimately need to be paired with technical and governance measures to ensure that metadata is accurate and remains attached to its content.

Technical vulnerabilities mean Content Credentials can be removed or falsified. The most common failure of Content Credentials is *metadata stripping*, in which the process of uploading files or converting file formats discards manifests automatically. The second is the *analog hole*: if a user screenshots or rerecords content, the original manifest doesn't carry over to the new file. The third issue is *signing-through-camera*: point a camera that signs images automatically at a deepfake on a screen and it will honestly sign that it captured the image. The fourth is *forged manifests*: skilled attackers can build manifests that contain false assertions. Addi-

tionally, researchers continue to probe whether C2PA's cryptographic validation itself can be defeated.

Mitigations can help reduce these vulnerabilities. To address metadata stripping, two technical fixes are to include invisible watermarks such as Google's SynthID and to use "digital fingerprints," which let a stripped file be rematched to its manifest. When these measures are combined with Content Credentials, it becomes far harder to sever the record from the content it describes. The California AI Transparency Act also requires that, beginning in 2027, large platforms not strip standards-compliant provenance data.

The analog hole problem, meanwhile, can be mitigated as norms change. While there's no way to prevent users from stripping provenance metadata by rerecording content, material without provenance metadata will eventually be seen as less trustworthy for this reason. As camera and microphone manufacturers widely adopt C2PA, bare content will stand out, reducing the incentive to strip metadata.

For signing-through-camera, emerging technical solutions such as Sony's Camera Authenticity Solution embed 3D depth information in metadata, revealing whether the photo captured a real scene or a flat screen. The threat of forged manifests, meanwhile, requires both technical and institutional responses. For example, conformance programs can verify that products adhere to the Content Credentials technical specification and that their signatures cannot be manipulated. The specification itself can also be revised to address vulnerabilities.

Sony's Camera Authenticity Solution makes it hard to pass off synthetic images as genuine by photographing them off a screen, addressing the signing-through-camera vulnerability. Source: Sony.

Private and public institutions can help determine which signers are trustworthy. As discussed above, C2PA uses cryptographic signatures for security, and each signature traces back to a particular signer (e.g., a smartphone manufacturer). A signature answers which credential signed a claim, but who decides which signers are trustworthy? Currently, certificate authorities vet signers, and the C2PA Trust List governs which of those authorities receive official approval. In this certification system, private actors police themselves; major firms building the tools that create and sign content, such as Adobe, also set the conformance rules and decide which signers are trustworthy. Public memory institutions such as libraries, archives, and museums could act as independent, mission-driven custodians in this process.

There is a long-standing tension between transparency and privacy. While provenance brings benefits, it also carries the risk of misuse. Provenance trails can sometimes reveal who filmed a video, where, and when. Journalists who wish to prove the authenticity of a piece of media might also struggle to protect dissidents and whistleblowers. While there is an inherent trade-off between transparency and privacy, C2PA tries to

strike a balance. Most C2PA assertions are optional; sensitive fields can be redacted; and certificates can be pseudonymous.

Content Credentials have been adopted widely but unevenly. Several generative AI tools already sign content by default, including Adobe Firefly and OpenAI's DALL·E 3 and Sora. Meanwhile, Midjourney still embeds no manifest and has not committed to a timeline for implementing one. Adoption in hardware is also growing. Leica's M11-P was the world's first camera to have Content Credentials built in. Now, Sony and Canon have followed suit, alongside Google's Pixel 10. The most consequential gap is Apple, which hasn't yet committed to implementing C2PA or given a public explanation as to why. California's capture-device rule will increase the pressure on Apple and other major manufacturers to adopt standards-compliant provenance beginning in 2028.

What the EU and California Acts Require

The EU and California laws create several different obligations. Some center on generating and preserving provenance records; others focus on using those records to create appropriate labels and notices.

The EU obliges generative AI providers to create provenance records. In the EU AI Act, the single most important obligation is Article 50(2): providers of generative systems must ensure those systems automatically mark synthetic audio, image, video, and text output in machine-readable form. This obligation is the key to an evidentiary layer that can automatically detect material with trustworthy provenance.

The EU's other two duties focus on disclosure. With some exceptions, Article 50(4) requires deployers of AI systems to label deepfakes and to disclose AI-generated text on matters of public interest. Article 50(1) places a similar requirement on AI providers to tell people when they are interacting with an AI system.

California's law goes further by requiring online platforms to retain provenance data. The California AI Transparency Act applies to providers of generative AI systems that have more than a million monthly users and are accessible in the state. The headline obligations on these providers, effective August 2, 2026, are to include embedded disclosures and offer a free tool that the public can use to surface content provenance data. These requirements act as California's equivalents to the EU's marking-and-disclosure duties, and they require the embedded data to be difficult to remove. Starting on January 1, 2027, large online platforms must detect and surface standards-compliant provenance data attached to content, and—critically—refrain from stripping it. This rule will greatly increase the durability of provenance metadata, making the whole content provenance ecosystem more useful.

California's law requires many hardware manufacturers to create provenance records. The final component of California's law, effective January 1, 2028, extends provenance requirements beyond AI-generated content to content recorded by devices such as smartphones, cameras, and voice recorders. These devices must create provenance records by default, though manufacturers may give users an option to turn this feature off.

Enforcement can add up to enormous sums. Under the EU AI Act, national regulators can impose fines of up to €15 million or 3% of worldwide annual turnover, whichever is higher, for breaching the Article 50 transparency duties. The percentage-of-global-revenue model scales with the size of the offender; for a large firm, fines could grow to hundreds of millions of dollars. In California, the Attorney General, city attorneys, and county counsel can seek civil penalties of $5,000 per violation, which look modest by comparison. However, each day of noncompliance is treated as a new violation. The fine for a single noncompliant product, multiplied across days, can quickly add up.

The EU and California Laws' Global Impact

The principle of transparency around AI-generated content has long been discussed in AI governance and has appeared in international agreements in recent years. However, the laws enforced in the EU and California will, for the first time, set an effective global baseline for transparency requirements.

The notion of content provenance is not new, but requirements vary widely. AI governance documents from the OECD, UNESCO, the G7, and the Council of Europe have already articulated content-provenance norms at the international level, although they are mostly nonbinding. The United States still has no general federal content provenance law. At the state level, both Utah and Washington have enacted their own content provenance laws aligned with California's AI Transparency Act.

China's laws also fuel efforts to build provenance infrastructure. Beyond the EU and the US, the most prescriptive approach is China's labeling regime. Its Deep Synthesis Provisions, Interim Generative AI Measures, Measures for Labeling of AI-Generated Synthetic Content, GB 45438-2025 national standard, and other regulations and institutions already require AI-generated content to be marked. However, China has not yet extended comparable requirements to hardware manufacturers.

The EU and California laws apply to content that ends up within their borders. Neither the EU law nor the California law hinges on where a company is incorporated or headquartered, or where its servers sit. Instead, both laws center on market access and effects—whether the system is placed on the market or publicly accessible in the jurisdiction, and whether its outputs are used by people there—with the California AI Transparency Act adding a one-million-monthly-user threshold.

The laws' footprints are effectively global. Digital content does not respect borders, and no generative system can guarantee that its outputs will never appear in the EU or California—two of the largest and most affluent markets in the world. Attempting to run a compliant environment for these markets while running a noncompliant one everywhere else could be inconvenient and expensive. Instead, companies may apply the stricter standard to all output by default. As a result, these region-specific mandates are set to have an international impact.

How Provenance Mandates Support Courts in Tackling Deepfake Evidence

The primary effect of the new EU and California mandates is to require the creation of signed, time-stamped, tamper-evident records that document data origin and custody. Such records will be generated at scale, as a matter of legal obligation. Records are evidence, which raises a question: how should courts use this evidence?

Advisory bodies are considering responses to deepfakes presented as evidence. The Advisory Committee on Evidence Rules, which proposes amendments to the US Federal Rules of Evidence, drafted a working Rule 901(c) to handle evidence fabricated by generative AI. Consider a case in which a party claims that a piece of evidence presented against them is a deepfake. Under the working draft, that party would first have to produce evidence sufficient to support a finding of fabrication—a lower bar than proving falsification outright. Only then must the item's proponent show the judge that it is more likely than not authentic. If such a rule were adopted, validated provenance records would be powerful evidence for meeting its test.

Provenance records can help courts weigh authenticity, though their absence must be read with care. The machinery of 901(c) turns on whether an item was generated or altered by AI, something courts have had no dependable way of establishing. Validated provenance records offer an important factor to consider. A missing manifest, by contrast, proves little by itself: not all cameras create one, and platforms routinely strip provenance information. Absence becomes significant only where the facts of the case suggest that provenance information should be present—a situation that will grow more common as California's and the EU's provenance mandates take effect. In such situations, an unexplained gap begins to raise a question about the party that could have signed and did not, much as courts already draw inferences when a party fails to preserve evidence it had a duty to keep.

New technologies have often presented courts with novel challenges when it comes to identifying authentic evidence. In the 1800s, courts spent decades working out how to treat photographic evidence. The following century, fingerprinting and DNA became more reliable and useful in court as institutions—registries, labs, and standards—grew to support the collection and analysis of such evidence. Whether C2PA will become a global default for verifying the status of images, audio, and other data remains an open question. But with the arrival of EU and California laws, courts will at least have a better record of how a digital item came to be and how it was altered. Content provenance is the evidentiary layer the AI era has lacked.

Work on this essay was supported by the Summer Research Fellowship at the Institute for Law & AI.

See things differently? AI Frontiers welcomes expert insights, thoughtful critiques, and fresh perspectives. Send us your pitch.

Eddan Katz is the Head of Policy & Government Affairs at Encypher. He is also a Research Fellow with the LexLab at UC Law San Francisco, writing on and teaching AI liability. Previously, he was Executive Director of Yale Law School's Information Society Project (ISP), was the International Affairs Director at the Electronic Frontier Foundation (EFF), the Project Lead for the AI Governance platform at the World Economic Forum's Centre for the Fourth Industrial Revolution network, and worked on the AI Policy team at Meta.

The problem with disjunctive existential risk

BY STEFAN SCHUBERT · THE UPDATE · 13 AUGUST 2026

Suppose that Alice and Bob have different beliefs about existential risk over the next 50 years. While Alice thinks it comes overwhelmingly from a single threat, Bob views it as a disjunction of four independent threats.

Alice:

- Threat A: 20%

Bob:

- Threat A: 6%
- Threat B: 6%
- Threat C: 5%
- Threat D: 5%
- Combined risk (given independence): ~20%

Setting aside your beliefs about actual threats such as AI misalignment or nuclear war, who looks more reasonable to you?

I think many people would say Bob. While Alice's focus on a single big threat can come across as fanatical, Bob's broader portfolio of lower credences has a more balanced and humble look.

But I take the opposite view. Serious existential threats seem to arise rarely: it's debatable whether we even have one yet. This makes it decidedly unlikely that we have evidence of four near-term risks of 5 per cent or more. A single big risk is much more credible (if still unlikely), given the plausible hypothesis that the risk distribution is heavy-tailed.

Existential risk communication has often given comparable attention to several unrelated threats, creating the impression that they pose risks of roughly the same order. But if the near-term risk is indeed high, it's more likely to be concentrated in one or two threats – AI and possibly synthetic biology. So it's better to emphasise those threats than to use broader framings that make existential risk seem more disjunctive than it is.