Curatone Art & Research Journal
Vol. 1, Issue 2 (2026) | ISSN 3054-6621
Published by Curatone.art | Berlin, Germany

RESEARCH ARTICLE · PEER REVIEWED · OPEN ACCESS

Less, but Verified: A Rams-Derived Framework for Evaluating Image Provenance in the Age of Generative AI

Weniger, aber verifiziert: Ein aus den Prinzipien von Dieter Rams abgeleitetes Rahmenwerk zur Bewertung der Bildprovenienz im Zeitalter generativer KI

Laxmi Narayan Behera
University of Maine at Presque Isle

Received: October 5, 2026 | Accepted: October 29, 2026 | Published: November 5, 2026
Editor-in-Chief: Elizaveta Akimova | Peer Review Board: Josephine Florens, Maria Raskin, Yiting Chen

ABSTRACT

Generative image models have severed the historical link between a picture and the hand or lens that produced it. In response, industry, regulators, and researchers have assembled a provenance stack of signed metadata, invisible watermarks, statistical detectors, and distributed ledgers. Each component is usually evaluated in isolation and against narrow technical metrics such as robustness or false-positive rate. This article argues that provenance is not only a security problem but a design problem, and that it should be judged by how well it serves the people who make images and the people who look at them. Drawing on Dieter Rams's ten principles of good design, I derive ten evaluation criteria for provenance systems and apply them to four families of mechanisms: C2PA Content Credentials, invisible watermarking, synthetic-image detection, and blockchain registries. The analysis is informed by a model borrowed from defense logistics, where the National Stock Number identifies what an item is but cannot, on its own, establish where a particular part came from. Four harms structure the evaluation: loss of artistic credit, deceptive synthetic media, loss of metadata in distribution, and provenance as surveillance. I find that no single mechanism satisfies the criteria, that robust designs are layered rather than monolithic, and that the artist, not the platform, should serve as the root of trust. The article closes with concrete recommendations for artists, platforms, standards bodies, regulators, and the institutions that exhibit and publish visual work.

Keywords: content provenance; C2PA; generative AI; design principles; digital art authentication; blockchain

1. Introduction

In 1936, Walter Benjamin argued that mechanical reproduction strips the artwork of its aura, the singular presence of a work in a particular time and place (Benjamin, 1936/2008). His worry was the copy. Nearly a century later the problem has inverted. A diffusion model can now produce a convincing photograph of an event that never occurred, or a painting in the manner of a living artist who never made it. Such an image has no here and now to lose. The question before visual culture is no longer whether a copy is faithful to its original, but whether an original exists at all, and who stands behind it.

The dominant response has been technical. The Coalition for Content Provenance and Authenticity (C2PA), founded in 2021 by Adobe, Arm, the BBC, Intel, Microsoft, and Truepic, publishes an open specification for cryptographically signed provenance records now in its fourth major revision (C2PA, 2026). The National Institute of Standards and Technology has mapped the field into provenance data tracking, watermarking, and synthetic content detection (NIST, 2024). In the European Union, the transparency obligations of Article 50 of the Artificial Intelligence Act became applicable on 2 August 2026, requiring providers of generative systems to mark outputs in a machine-readable way and requiring deployers to label deepfakes (European Commission, 2026; Regulation (EU) 2024/1689, 2024). These are serious developments, and they deserve serious engagement.

Yet the discussion is conducted almost entirely in the vocabulary of engineering: robustness, detection accuracy, false-positive rates, cryptographic binding. That vocabulary is necessary but incomplete. It omits the questions an artist actually asks. Does this system help me say, credibly, that I made this? Will the claim survive the journey from my studio to a stranger's screen? And can a viewer believe me without the system exposing more of me than I chose to reveal? These are design questions as much as security questions, and they call for a design standard.

This article proposes one. I write from the position of a practitioner in federal cybersecurity and information technology, a field in which provenance is not a metaphor. For more than a decade, my company has delivered cybersecurity and software services to government customers. In that work, software is only as trustworthy as its documented origin: what it is, who built it, and what happened to it before it reached the system it protects. In defense supply chains, a counterfeit component and a genuine one can be visually indistinguishable; what separates them is a documented chain of custody back to an authorized source. That perspective shapes my central claim: provenance for images should be engineered with the seriousness we bring to critical components, and designed with the restraint of good industrial design. The German industrial designer Dieter Rams summarized that restraint in a phrase that became the title of his 1995 book, Weniger, aber besser: less, but better (Rams, 1995; Lovell, 2011). For provenance, I propose a variant: less, but verified.

The article makes four contributions. First, it maps image provenance onto a supply-chain identity model, distinguishing the identity of a class of item from the pedigree of a specific instance. Second, it translates Rams's ten principles of good design into ten evaluation criteria for provenance systems. Third, it applies those criteria to four families of mechanisms against four harms, producing a structured qualitative assessment. Fourth, it offers concrete recommendations, organized around a single design commitment: the artist should be the root of trust. The work is conceptual and analytical rather than empirical; its ratings are argued judgments grounded in published evidence, not benchmark measurements, and I return to that limitation in Section 9.

2. Background

2.1 From Aura to Manifest

Benjamin's aura was a property of physical presence. Digital images never had it in his sense, but they acquired a functional substitute: context. A photograph published by a known newspaper, credited to a named photographer, and accompanied by a date and location carried a social warrant of authenticity even when its pixels were infinitely copyable. That warrant lived outside the image, in the institutions and captions around it.

Hito Steyerl's account of the poor image describes what happens to that context in circulation. Images on the network are compressed, cropped, re-encoded, and reposted until they become low-resolution copies detached from their origins (Steyerl, 2009). Steyerl treats the poor image partly as a democratizing force, and she is right that degraded circulation has expanded who can see and remix visual culture. But the same process that liberates the image also launders it. Each reposting is a point at which credit can fall away and at which a fabricated image can acquire the borrowed authority of a real one.

Provenance technology is an attempt to rebuild the lost warrant inside the file itself. Where the aura was presence and the newspaper caption was context, the C2PA record is a manifest: a structured, signed statement of where an asset came from and what was done to it. The shift from aura to manifest is the shift from authenticity as an experience to authenticity as a verifiable claim.

2.2 The Provenance Stack

NIST's taxonomy provides a useful map of the available mechanisms (NIST, 2024). Provenance data tracking records facts about content, either as metadata attached to the file or as a watermark embedded in the content itself. Synthetic content detection attempts to classify content as generated or not, typically with statistical models, when no reliable provenance record is present. To these I add a fourth family that sits

adjacent to the NIST categories: distributed ledgers that timestamp a record of an asset in an append-only public log.

The C2PA specification is the most developed instance of metadata-based provenance. Its central object, the Content Credential or manifest, binds a set of assertions about an asset (for example, the capture device, the editing actions performed, or the use of a generative model) into a claim that is cryptographically signed by the entity making it (C2PA, 2026). The binding is tamper-evident: altering the asset or the assertions without re-signing breaks verification. The specification's own design goals include privacy, scalability, and explicit attention to potential harms, including threats to human rights and disproportionate risks to vulnerable groups (C2PA, 2026).

Watermarking takes the opposite approach. Instead of attaching information to the file, it hides a signal in the pixels, ideally imperceptibly, so that it can survive the stripping of metadata. Detection does without any embedded signal and instead infers origin from statistical traces left by generators. Ledgers, finally, descend from a line of cryptographic work on timestamping that predates cryptocurrency: Haber and Stornetta (1991) showed how to certify that a digital document existed at a given time by chaining hashes, the same structural idea that later underpinned Bitcoin's public ledger (Nakamoto, 2008).

2.3 Authorship, Law, and Regulation

Provenance matters to artists not only as a matter of reputation but increasingly as a matter of law, because legal outcomes now turn on facts about process. In 2022, Jason Allen's image Théâtre D'opéra Spatial, produced with the Midjourney system, won the digital art category at the Colorado State Fair. When Allen sought copyright registration, he described entering at least 624 text prompts and revisions to arrive at the initial image. On 5 September 2023 the U.S. Copyright Office Review Board affirmed the refusal to register the work as submitted, holding that the AI-generated material had to be disclaimed, while leaving open protection for Allen's own subsequent edits (U.S. Copyright Office Review Board, 2023).

In March 2025, the U.S. Court of Appeals for the D.C. Circuit held in Thaler v. Perlmutter (2025) that the Copyright Act requires human authorship, rejecting an application that named an AI system as the sole author. The court was careful to note that the human authorship requirement does not bar protection for works made by or with the assistance of AI. The line the law draws, in other words, is drawn through the process: how much of the work reflects human creative choices.

European regulation reaches the same territory from the side of transparency rather than ownership. Article 50 of the AI Act requires providers to mark synthetic outputs in a machine-readable, detectable format and requires deployers to disclose deepfakes, with obligations applying from 2 August 2026 and a grace period to December 2026 for the marking of generative systems already on the market (European Commission, 2026). Taken together, these developments make provenance something more than a courtesy. It is becoming evidentiary infrastructure. The record of how an image was made is precisely the evidence that tribunals, regulators, and competition juries now need, and the party best placed to create that record is the person who made the image.

2.4 The National Security View

The U.S. national security community reached many of the same conclusions from the direction of threat. In September 2023, the National Security Agency, the Federal Bureau of Investigation, and the Cybersecurity and Infrastructure Security Agency jointly warned that synthetic media had become a practical tool for impersonating senior leaders and defrauding organizations, and recommended that organizations adopt real-time verification, passive detection, and the determination of multimedia provenance, alongside rehearsed response plans and personnel training (NSA, FBI, & CISA, 2023). In January 2025, the NSA joined the Australian Cyber Security Centre, the Canadian Centre for Cyber Security, and the United Kingdom's National Cyber Security Centre in guidance devoted specifically to Content Credentials (NSA et al., 2025).

Two points in that guidance anticipate the argument of this article. First, it is careful about what credentials prove: they record who created and edited media, when, and how, but they establish verifiability rather than truth. Second, it recognizes the fragility of metadata in distribution and recommends durable credentials, in

which watermarks and fingerprint matching allow a stripped credential to be recovered, together with an organizational commitment to preserve unaltered metadata across the media lifecycle (NSA et al., 2025). The intelligence community also offers an older and more general lesson about how evidence is read. Sherman Kent (1964) showed that analysts and policymakers attached widely different probabilities to the same estimative words, and Richards Heuer's study of intelligence analysis documented how even trained analysts over-weight vivid evidence and neglect the significance of evidence that is missing (Heuer, 1999). Both findings bear directly on how ordinary viewers will interpret a provenance badge, a detector's score, or the absence of either.

3. A Supply-Chain Lens: What the National Stock Number Teaches

Defense logistics offers an instructive model for thinking about provenance, because it has confronted the problem of authenticity at industrial scale for decades. Every item repeatedly bought, stocked, or issued by the U.S. Department of Defense receives a National Stock Number (NSN), a 13-digit code assigned through the Federal Logistics Information System, together with a record describing the item's name, part numbers, and characteristics (Defense Logistics Agency, 2022). The first four digits classify the item within a supply group; the remaining nine form the National Item Identification Number, which begins with a two-digit code for the national codification bureau that assigned it. The system is shared across NATO, where the same number is known as a NATO Stock Number.

The NSN is an extraordinarily effective answer to one question: what is this item? Any conforming component of a given specification shares a single number, regardless of which manufacturer produced it. That is precisely its strength for interchangeability, and precisely its weakness for authenticity. The NSN says nothing about which physical part, in which box, came from where. It identifies a class, not an instance.

The gap between class and instance is not hypothetical. In 2012 the U.S. Senate Committee on Armed Services reported approximately 1,800 cases of suspect counterfeit electronic parts in the defense supply chain, involving more than one million individual parts, including approximately 84,000 suspect parts traced to a single supplier; in each of the three cases the committee examined in depth, the Department of Defense had been unaware that counterfeit parts were installed in its systems (U.S. Senate Committee on Armed Services, 2012). The counterfeits did not lack identity. They carried valid part numbers and passed for the right class of item. What they lacked was legitimate pedigree. The policy response accordingly emphasized traceability to the original manufacturer and, through Item Unique Identification under Department of Defense Instruction 8320.04, marking that identifies individual items rather than item types.

The same principle was written into acquisition regulation. The Defense Federal Acquisition Regulation Supplement now requires defense contractors, wherever possible, to obtain electronic parts from the original manufacturer or its authorized suppliers (48 C.F.R. § 252.246-7008). In software services the same principle takes a concrete and daily form. A government customer does not accept software simply because it runs correctly; it asks where the code came from, who built it, and whether it has been altered on its way to deployment. Two copies of a program can be byte-for-byte functional equivalents; the one with a documented origin is acceptable and the one without it is not. The record, not the object, carries the trust. That is the intuition this article carries from the procurement office into the gallery.

The mapping to images is direct. A style, a genre, or a description such as a photograph of a harbor at dusk is class identity, the visual equivalent of a stock number. Generative models are exceptionally good at producing items that satisfy a class specification; that is what they are trained to do. What an artist needs to establish is instance pedigree: that this particular file descends from this person's hand, camera, or working session. In supply-chain terms the artist is the original component manufacturer, and galleries, journals, competitions, and platforms are distributors.

Three lessons follow. First, trust must be anchored at the origin; no amount of downstream inspection can reliably reconstruct a pedigree that was never recorded. Second, every handoff is an attack surface. The Senate committee found counterfeits entering through unvetted intermediaries, and the image economy has its own equivalent in the platforms and aggregators that re-encode and repost content. Third, silence is itself a failure mode. The committee found that suspect counterfeits routinely went unreported. In the image

economy, the analogous silence is the quiet removal of provenance records in distribution, which leaves viewers unable to distinguish an image that never had a record from one whose record was lost.

4. Threat Model: Four Harms

A design standard needs to know what it is defending against. I organize the analysis around four harms, each associated with a characteristic adversary or failure.

H1, loss of credit. The adversary is the launderer, who detaches attribution from a work or reproduces a living artist's style without credit. Protective tools aimed at style mimicry, such as Glaze, add imperceptible perturbations to published artwork so that models fine-tuned on it learn a distorted style (Shan et al., 2023). Subsequent work has shown that such protections can be undermined by simple robust mimicry techniques, such as image upscaling (Hönig et al., 2025). Protection against copying, in short, is fragile; proof of origin is a complementary and more durable defense.

H2, deception. The adversary is the forger, who presents synthetic media as a record of real events. Chesney and Citron (2019) identified a second-order effect that is equally important: the liar's dividend, in which the mere plausibility of fakes allows authentic evidence to be dismissed as fabricated. A provenance system therefore has two jobs under H2: to expose fakes and to protect the credibility of the real.

H3, loss in distribution. Here there is often no adversary at all. Social platforms and publishing systems routinely strip metadata from uploaded files, frequently for legitimate reasons of privacy and file size, and a screenshot removes it entirely. The World Privacy Forum (2025) notes that such stripping breaks the provenance chain and with it the promise of tracing content through the media ecosystem. H3 is the harm of the indifferent pipeline.

H4, provenance as surveillance. The adversary is the observer. A record that binds an image to a name, a device, a location, and a time can expose a creator, a journalist, or a witness to retaliation. The C2PA's own harms modeling acknowledges that use of the technology may result in human rights violations and that malicious actors, including state actors, could misuse it (World Privacy Forum, 2025).

These harms pull in opposite directions. H1 and H2 push toward stronger, more persistent binding between an image and its maker. H4 pushes toward weaker, more selective binding. H3 reveals that strong binding is worthless if it does not survive transit. No single parameter setting satisfies all four. Resolving that tension is not an optimization problem with a closed-form answer; it is a design problem, and it calls for design principles.

5. The Framework: Ten Principles, Ten Criteria

Dieter Rams, the German industrial designer who served as chief designer at Braun for several decades, articulated ten principles of good design that remain among the most cited statements of design ethics (Lovell, 2011; Vitsœ, n.d.). Good design, in his formulation, is innovative; makes a product useful; is aesthetic; makes a product understandable; is unobtrusive; is honest; is long-lasting; is thorough down to the last detail; is environmentally friendly; and is as little design as possible.

Rams is an apt reference for security engineering for a reason that may not be obvious. His principles treat restraint as a virtue, and security practitioners learn the same lesson from the other direction: complexity is the enemy of security. Every additional field, interface, and dependency is a new place for a system to fail or to leak. A provenance standard that tries to record everything about an image will be harder to understand, easier to break, and more dangerous to its users than one that records only what a viewer needs to know. Rams supplies a vocabulary for saying this precisely.

Table 1 translates each principle into an evaluation criterion for provenance systems, with a test question that an evaluator can apply. Two translations require comment. I read environmentally friendly broadly, to include both computational and energy cost and the health of the wider ecosystem, meaning openness and interoperability. And I read as little design as possible as a principle of minimal disclosure: a provenance system should bind only what is necessary to support the claim being made, which is the design expression

of privacy by default.

Table 1. *Rams-Derived Evaluation Criteria for Provenance Systems*

Principle (Rams)	Criterion for provenance systems	Test question
1. Innovative	Advances trust	Does it address a threat that existing mechanisms could not?
2. Useful	Answers the viewer's question	Does it tell the viewer who made the work and how, when they need to know?
3. Aesthetic	Respects the work	Does it leave the image visually unaltered?
4. Understandable	Is read correctly	Will a non-expert interpret the signal accurately?
5. Unobtrusive	Fits existing practice	Does it work without forcing artists to change how they create?
6. Honest	Claims only what it proves	Does it avoid implying that a signed image is a true image?
7. Long-lasting	Survives time and transit	Does the record survive re-encoding, reposting, and vendor exit?
8. Thorough	Covers the whole chain	Does it record edits and handoffs, and fail visibly?
9. Environmentally friendly	Is proportionate and open	Is its compute and energy cost proportionate, and is it interoperable?
10. As little design as possible	Discloses minimally	Does it bind only what the claim requires?

6. Applying the Framework

I now evaluate four mechanism families against the ten criteria, rating each as Strong, Partial, or Weak. The ratings in Table 2 summarize the discussion that follows, and Table 3 summarizes performance against the four harms. They are analytical judgments based on the published evidence cited, intended to make trade-offs visible rather than to rank products.

6.1 C2PA Content Credentials

Content Credentials score well on the principles that concern the integrity of the claim. They are honest in a precise sense: a valid manifest establishes who signed a set of assertions and that neither the assertions nor the asset have been altered since, and it does not purport to establish that the depicted scene is true. They are thorough, because the manifest can record a chain of edits and ingredients rather than a single label. And as an open specification governed by a broad coalition, they score well on the ecosystem reading of the ninth principle (C2PA, 2026).

Their weaknesses are concentrated in durability and comprehension. Metadata-based credentials do not survive the indifferent pipeline of H3; stripping on upload, or a simple screenshot, removes them (World Privacy Forum, 2025). And the distinction between a valid signature and a true image is subtle. A viewer who sees a provenance badge may reasonably infer that the content has been verified as accurate, which the system never claims. The joint NSA guidance draws exactly this line between verifiability and truth (NSA et al., 2025), but a distinction drawn in a technical guidance document does not automatically reach a viewer scrolling past a badge. Heuer's (1999) observation that vivid evidence is over-weighted suggests that a prominent badge will tend to be read as a verdict unless the interface works deliberately against that reading.

On the principle of minimal disclosure the picture is mixed: the specification permits creators to omit identifying fields and includes harms analysis, but the same architecture can bind identity, device, and location tightly when an implementation chooses to. My overall judgment is that Content Credentials are the best available backbone for provenance and the weakest link in transit.

6.2 Invisible Watermarking

Watermarks are the natural complement to metadata. Because the signal lives in the pixels, it can survive the stripping of file metadata, and well-designed watermarks are imperceptible, satisfying the aesthetic criterion. They are also unobtrusive for the user, since embedding can happen automatically inside a generation or editing tool.

Their honesty is asymmetric. The presence of a watermark can indicate that a model produced an image, but its absence proves nothing, since most images carry no watermark and an adversary can remove one. On that last point the evidence is now strong: Zhao et al. (2024) demonstrated, with formal proofs and empirical evaluation, that pixel-level invisible watermarks are vulnerable to regeneration attacks that pass an image through a generative model and reconstruct it without the mark. Watermarks therefore cannot serve as a root of trust. They are best understood as a recovery channel, a durable pointer that allows a stripped image to be re-associated with its manifest, which is the role the joint NSA guidance assigns them under the heading of durable credentials (NSA et al., 2025).

6.3 Synthetic-Image Detection

Detection is the only mechanism that requires nothing of the creator. It is maximally unobtrusive and, because it binds no identity, poses little surveillance risk. That makes it attractive as a last line of defense when no provenance record exists.

Its weaknesses mirror its strengths. Detection produces probabilistic judgments that must be continually retrained against new generators, and its outputs are easily misread as verdicts. The intelligence community has studied this communication problem for decades. Kent (1964) found that readers of national estimates assigned very different probabilities to terms such as probable and likely, and a detector that labels an image likely synthetic inherits the same ambiguity, now delivered to a mass audience without training in how to read it. Under H2 it is especially exposed to the liar's dividend: a false positive on an authentic photograph does exactly the damage that Chesney and Citron (2019) warned about. On honesty and understandability it rates Weak, not because detectors are useless but because the signal they produce is harder to communicate truthfully than a signed statement. NIST (2024) treats detection as a distinct family from provenance tracking for this reason; it answers a different and less certain question.

6.4 Blockchain Registries

Blockchain registries are useful, with limits. Their genuine strength is long-lastingness. A hash anchored in a widely replicated public ledger provides durable, independently checkable evidence that a particular record existed at a particular time, without dependence on any single company's continued operation. That is the timestamping property Haber and Stornetta (1991) formalized, and it is valuable to an artist who may one day need to show that a work predates a contested copy.

Four limits constrain that value. First, a ledger records what was registered, not who truly made it. If a launderer registers a stolen image first, the ledger faithfully preserves the lie. The trust problem is moved to the moment of registration, not solved. Second, many registries store only a pointer or a cryptographic hash, while the image itself lives elsewhere; when the hosted file disappears or is re-encoded, the link between the record and the work weakens. A cryptographic hash changes completely when a single pixel or compression setting changes, so a reposted copy will not match the registered hash at all, which makes ledgers poorly suited to H3 unless paired with perceptual matching. Third, immutability is in direct tension with the principle of minimal disclosure: an identity written permanently to a public ledger cannot be withdrawn if it later endangers the person it names. Fourth, environmental cost depends heavily on the ledger's consensus mechanism, with proof-of-work systems far more energy-intensive than proof-of-stake alternatives.

The appropriate role for a ledger is therefore narrow and specific: an optional timestamp anchor for a hash of an artist-signed manifest, not an identity registry and not a substitute for a signature at the source.

6.5 Synthesis

Tables 2 and 3 make the central finding visible: no column is Strong across the board, and the mechanisms fail in complementary places. Content Credentials are honest and thorough but fragile in transit. Watermarks survive transit but cannot carry trust. Detection asks nothing of creators but promises less than it appears to. Ledgers last but record only what they are given. The natural conclusion is a layered architecture in which each mechanism does the job it is good at, and in which the weakest property of one layer is covered by the strongest property of another. That is defense in depth, the organizing principle of serious security engineering, expressed through Rams's discipline of doing only what is necessary.

Table 2. *Mechanism Families Rated Against the Ten Criteria*

Criterion	C2PA Credentials	Invisible watermark	Detection	Blockchain registry
1. Advances trust	Strong	Partial	Partial	Partial
2. Useful	Strong	Partial	Partial	Partial
3. Respects the work	Strong	Strong	Strong	Strong
4. Read correctly	Partial	Weak	Weak	Weak
5. Fits practice	Partial	Strong	Strong	Weak
6. Honest	Strong	Partial	Weak	Partial
7. Survives transit	Weak	Partial	Partial	Strong
8. Covers chain	Strong	Weak	Weak	Weak
9. Proportionate, open	Strong	Partial	Partial	Partial
10. Discloses minimally	Partial	Strong	Strong	Weak

Table 3. *Mechanism Families Rated Against the Four Harms*

Harm	C2PA Credentials	Invisible watermark	Detection	Blockchain registry
H1 Loss of credit	Partial	Partial	Weak	Partial
H2 Deception	Partial	Partial	Partial	Weak
H3 Loss in distribution	Weak	Partial	Partial	Weak
H4 Surveillance risk (higher is safer)	Partial	Partial	Strong	Weak

7. The Artist as Root of Trust

In public key infrastructure, a root of trust is the entity whose key is accepted without further proof and from which all other trust is derived. Every provenance architecture has one, whether or not its designers name it, and the choice is consequential.

Three alternatives to the artist are currently prominent, and each has a defect. A platform-rooted design makes the distributor the arbiter of authorship, reproducing the gatekeeping that digital culture was supposed

to loosen and making an artist's credibility contingent on a company's policies. A device-rooted design, in which cameras sign captures, proves that a sensor recorded light at a moment, which is valuable for photojournalism but says nothing about authorship of a painting, an illustration, or a composite, and excludes most visual art entirely. A model-rooted design, in which generators mark their outputs, tells a viewer that a model was used but not who directed it or how much of the result reflects human choices, which is exactly the question the Copyright Office and the courts now ask.

An artist-rooted design places the signature where the claim originates. The artist signs at the moment of creation, inside the tools of their practice, and stakes their reputation on the statement that the work is theirs and was made in the way the manifest describes. This is honest in Rams's sense. The system does not certify that the artist is truthful; it makes the artist's claim attributable and accountable, as a supplier certification makes a supplier accountable without guaranteeing that no supplier will ever lie. Responsibility rests with the person best positioned to know the truth and with the most to lose from misrepresenting it.

Rooting trust in the artist also resolves much of the tension between H1 and H4. Because the artist controls the key, the artist controls disclosure. A well-known painter may bind a legal name; a documentary photographer working under a hostile government may sign with a pseudonymous credential whose consistency across works establishes a reputation without revealing an identity. Identity assertions developed alongside the C2PA standard are moving in this direction, and the design commitment proposed here is that pseudonymous, selectively disclosed identity should be a first-class option rather than an exception.

The architecture I propose, summarized in Table 4, has five layers. Each addresses a specific harm, and each is chosen because it is the least design that will do the job.

Table 4. *A Five-Layer, Artist-Rooted Provenance Architecture*

Layer	Mechanism	Primary harm addressed	Governing principle
L1 Origin	Artist key; signing at creation inside the creative tool	H1, H2	Honest; unobtrusive
L2 Claim	C2PA manifest with minimal assertions and optional pseudonymous identity	H1, H4	As little design as possible
L3 Recovery	Watermark or perceptual fingerprint pointing to the manifest	H3	Long-lasting
L4 Anchor	Optional ledger timestamp of the manifest hash	H1	Long-lasting
L5 Presentation	Viewer display naming the signer and recorded actions, without a truth verdict	H2	Understandable; honest

Two features of this architecture deserve emphasis. The first is that the layers are ordered by necessity, not by sophistication. Layer 1 and Layer 2 carry the claim; without them nothing else has meaning. Layers 3 and 4 exist only to keep that claim alive in a hostile or indifferent environment, and Layer 4 is explicitly optional, because a timestamp anchor adds value for some artists and none for others. Layer 5 is where most of the trust is actually won or lost, because it is the only layer a viewer ever sees. The second feature is that the architecture deliberately declines to do several things it could do. It does not record location by default, it does not require a legal name, and it does not route signing through a platform. Each omission is a choice in favor of the tenth principle, and each removes a category of failure along with a category of capability.

7.1 An Illustrative Walk-Through

Consider a hypothetical digital painter who works in a raster editor and publishes primarily on social platforms. Under the proposed architecture, her editing tool holds a signing key bound to a professional pseudonym she has used for a decade. When she exports a finished work, the tool generates a manifest recording that the work was created in that application, that it incorporates two of her own earlier files as ingredients, and that no generative model was used. It signs the manifest with her key (Layers 1 and 2). The export also embeds an imperceptible watermark carrying a short identifier that resolves to the manifest in a public lookup service (Layer 3). Because she has had work copied before, she elects to anchor a hash of the manifest in a public ledger (Layer 4).

She uploads the work to a platform that strips metadata. The manifest is gone from the posted file, but the watermark survives ordinary compression. Weeks later, a stranger reposts the image under a different name. A viewer who suspects the attribution runs the image through a verifier. The verifier reads the watermark, retrieves the manifest, checks the signature against the painter's public key, and confirms that the manifest hash appeared in the ledger on a date that precedes the repost. The display (Layer 5) states plainly that the work was signed by the pseudonymous creator, records the declared editing history, and makes no claim about anything the manifest does not assert.

Now consider the adversary's options. A determined launderer can attack the watermark with a regeneration technique of the kind Zhao et al. (2024) describe. If the attack succeeds, the repost loses its link to the manifest, and the verifier reports only that no provenance was found. That is a loss, but an honest one: the system does not falsely attribute the work to the launderer, and the painter retains a signed, timestamped original that predates the copy and can be produced in a dispute. The launderer cannot forge her signature without her key, and cannot backdate a competing manifest past the ledger anchor. The architecture does not make theft impossible. It makes theft provable, which is the realistic goal of any supply-chain control.

The same walk-through clarifies what the design does for the painter's safety. At no point did the system require her legal name, her location, or her device serial number. A reputation accrued to a stable pseudonym, and the viewer learned exactly what the painter chose to assert and nothing more. This is the practical meaning of less, but verified.

7.2 Failure Modes

The design has failure modes, and honesty requires naming them. Keys can be lost or stolen, so revocation and key rotation must be part of any deployment. Artists can lie, signing a generated image as handmade; the architecture does not prevent this, but it converts an anonymous deception into an attributable one, which is the precondition for reputational and legal consequences. And adoption depends on tooling: an artist-rooted system succeeds only if signing is as unobtrusive as saving a file.

8. Recommendations

The framework yields concrete recommendations for five groups of actors.

8.1 For Artists

Sign at the source, using tools that create Content Credentials at the moment of creation rather than after publication. Preserve process records, such as layered files, sketches, drafts, and session histories, because these are the evidence that copyright examiners and juries now ask for. Choose a level of identity disclosure deliberately, using a pseudonymous credential where personal safety requires it. Treat protective perturbation tools as one layer of defense, not as a guarantee, given the evidence of their fragility (Hönig et al., 2025).

8.2 For Platforms

Stop treating provenance records as disposable metadata. Where stripping is necessary to protect users, strip identifying fields while preserving or re-signing the minimal assertions needed to show that a credential

existed and what it claimed. Display provenance in language that matches what the cryptography proves, naming the signer and the recorded actions, and avoid labels that imply a verdict about truth. Support watermark-based recovery so that credentials stripped elsewhere can be re-associated with the work. The joint guidance from the NSA and its Five Eyes partners already asks organizations to commit to preserving unaltered metadata across the media lifecycle (NSA et al., 2025); platforms that host most of the world's images should hold themselves to at least that standard.

8.3 For Standards Bodies

Make privacy-minimal profiles the default rather than an option, consistent with the principle of as little design as possible. Specify interoperable recovery, so that a watermark or perceptual fingerprint embedded by one vendor can retrieve a manifest held by another. Continue and expand conformance testing, and publish user-comprehension research alongside technical specifications, because a provenance signal that viewers misunderstand fails the fourth principle regardless of its cryptographic strength.

8.4 For Exhibiting and Publishing Institutions

Galleries, competitions, and journals sit at exactly the point in the supply chain where pedigree can be checked. They should request provenance information at submission, publish clear policies on the use of generative tools, and treat credentials as evidence to be weighed rather than as automatic verdicts. An institution that asks for provenance creates demand for it, and demand is what moves artists and toolmakers to adopt it.

8.5 For Regulators

Labeling obligations such as those in Article 50 of the AI Act should recognize artist-signed provenance as a valid route to transparency and should avoid mandating disclosure of a creator's legal identity where a pseudonymous, accountable credential would serve the public interest equally well. Regulation that requires identity in order to prove authenticity will, under H4, deter precisely the witnesses whose images matter most.

9. Limitations and Future Work

This study is conceptual. The ratings in Tables 2 and 3 are argued judgments based on published evidence, not measurements, and other evaluators applying the same criteria might reasonably assign different ratings in particular cells. The translation of Rams's principles into evaluation criteria is itself interpretive; the principles were written for physical products, and their extension to cryptographic infrastructure involves choices that I have tried to make explicit. The field is also moving quickly. The C2PA specification has passed through several versions in two years, and European implementing guidance was finalized only in 2026, so specific technical judgments may date faster than the framework itself.

Three lines of future work follow. The first is user research on how viewers interpret provenance displays, which would test the fourth and sixth criteria directly. The second is measurement of how often Content Credentials survive distribution across major platforms, which would quantify H3. The third is a prototype of the artist-rooted, five-layer architecture inside a common creative tool, which would test whether the design can meet the unobtrusiveness criterion in daily practice.

10. Conclusion

Benjamin feared that reproduction would dissolve the authority of the original. Generative AI has completed that dissolution and added a new problem: an image can now carry the appearance of origin without having one. The defense supply chain learned, at considerable cost, that knowing what an item is does not tell you where a particular item came from, and that authenticity must be recorded at the source and protected at every handoff. Image provenance faces the same structural problem and should learn the same lessons.

It should also learn from design. A provenance system that records everything will break, confuse, and endanger. A system that records only what is necessary, signed by the person who made the work, carried through the pipeline by complementary mechanisms, and presented to viewers in honest language, can do

something more modest and more valuable: let an artist say, credibly and safely, I made this. Rams asked for weniger, aber besser, less, but better. For the age of generative images, the standard should be less, but verified.

Editorial & Review Credits

Editor-in-Chief: Elizaveta Akimova. Author: Laxmi Narayan Behera.

Peer Review Board:

- Josephine Florens (Contemporary figurative oil painter, Odesa and Germany; CAA Annual Conference Committee member; Apexart International Open Call juror): “This article offers a thoughtful interdisciplinary perspective on the challenges of artistic authorship and authenticity in the age of generative artificial intelligence. By connecting Dieter Rams’s design principles with digital provenance technologies, the author encourages an important discussion about protecting creative integrity, artistic identity, and trust in contemporary visual culture.”
- Maria Raskin (Mixed-media artist, Turkey; MA Fashion and Costume Design, Moscow State Textile University; Diploma in Traditional Iconography and Ancient Fine Art Techniques; Apexart New York juror): “The article’s most valuable contribution is its insistence that image provenance is a matter of design and artistic agency, not merely technical verification. By placing the artist at the root of trust and distinguishing a verifiable claim from proof of truth, it offers a thoughtful framework for protecting creative credit without making privacy the price of credibility.”
- Yiting Chen (Visual artist and illustrator, Beijing; Gold Winner, French Design Awards and MUSE Design Awards; exhibited at ICFF New York and Maison&Objet Paris): “What I found most interesting is the way this paper treats image provenance as a question of design, not just technology. The connection to Dieter Rams is useful because it keeps the discussion focused on what artists and viewers actually need to know. I also appreciate the attention given to privacy, especially the idea that proving where an image came from should not require an artist to reveal everything about themselves.”

This article was reviewed by members of the Curatone.art Peer Review Board. How to cite: Behera, L. N. (2026). Less, but Verified: A Rams-Derived Framework for Evaluating Image Provenance in the Age of Generative AI. Curatone Art & Research Journal, 1(2).

References

Benjamin, W. (2008). The work of art in the age of its technological reproducibility, and other writings on media (M. W. Jennings, B. Doherty, & T. Y. Levin, Eds.). Belknap Press of Harvard University Press. (Original work published 1936)

Chesney, R., & Citron, D. K. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. California Law Review, 107(6), 1753–1820. https://doi.org/10.15779/Z38RV0D15J

Coalition for Content Provenance and Authenticity. (2026). Content Credentials: C2PA technical specification (Version 2.4). https://spec.c2pa.org/specifications/specifications/2.4/specs/C2PA_Specification.html

Defense Federal Acquisition Regulation Supplement: Sources of Electronic Parts, 48 C.F.R. § 252.246-7008. https://www.ecfr.gov/current/title-48/chapter-2/subchapter-H/part-252/subpart-252.2/section-252.246-7008

Defense Logistics Agency. (2022). National Stock Number (NSN) assignment [Fact sheet]. https://www.dla.mil/Portals/104/Documents/J3LogisticOperations/Brochures/J3_NSNSPECSHEET_220722.pdf

European Commission. (2026). Quick facts: Transparency rules for AI systems. https://digital-strategy.ec.europa.eu/en/factpages/quick-facts-transparency-rules-ai-systems

Haber, S., & Stornetta, W. S. (1991). How to time-stamp a digital document. Journal of Cryptology, 3(2), 99–111. https://doi.org/10.1007/BF00196791

Heuer, R. J., Jr. (1999). Psychology of intelligence analysis. Center for the Study of Intelligence, Central Intelligence Agency. https://www.cia.gov/resources/csi/static/9a5f1162fd0932c29bfed1c030edf4ae/Pyschology-of-Intelligence-Analysis.pdf

Hönig, R., Rando, J., Carlini, N., & Tramèr, F. (2025). Adversarial perturbations cannot reliably protect artists from generative AI. In Proceedings of the Thirteenth International Conference on Learning Representations.

Kent, S. (1964). Words of estimative probability. Studies in Intelligence, 8(4), 49–65.

Lovell, S. (2011). Dieter Rams: As little design as possible. Phaidon.

Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. https://bitcoin.org/bitcoin.pdf

National Institute of Standards and Technology. (2024). Reducing risks posed by synthetic content: An overview of technical approaches to digital content transparency (NIST AI 100-4). https://doi.org/10.6028/NIST.AI.100-4

National Security Agency, Australian Signals Directorate's Australian Cyber Security Centre, Canadian Centre for Cyber Security, & UK National Cyber Security Centre. (2025). Content credentials: Strengthening multimedia integrity in the generative AI era [Cybersecurity information sheet]. https://www.cyber.gov.au/sites/default/files/2025-01/content-credentials-strengthening-multimedia-integrity-in-the-generative-ai-era.pdf

National Security Agency, Federal Bureau of Investigation, & Cybersecurity and Infrastructure Security Agency. (2023). Contextualizing deepfake threats to organizations [Cybersecurity information sheet]. https://www.nsa.gov/Press-Room/Digital-Media-Center/Document-Gallery/igphoto/2003298925/

Rams, D. (1995). Weniger, aber besser [Less but better]. Jo Klatt Design+Design Verlag.

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). (2024). Official Journal of the European Union, L 2024/1689. http://data.europa.eu/eli/reg/2024/1689/oj

Shan, S., Cryan, J., Wenger, E., Zheng, H., Hanocka, R., & Zhao, B. Y. (2023). Glaze: Protecting artists from style mimicry by text-to-image models. In Proceedings of the 32nd USENIX Security Symposium (pp. 2187–2204). USENIX Association.

Steyerl, H. (2009). In defense of the poor image. e-flux journal, (10). https://www.e-flux.com/journal/10/61362/in-defense-of-the-poor-image/

Thaler v. Perlmutter, 130 F.4th 1039 (D.C. Cir. 2025).

U.S. Copyright Office Review Board. (2023). Re: Second request for reconsideration for refusal to register Théâtre D'opéra Spatial. https://www.copyright.gov/rulings-filings/review-board/docs/Theatre-Dopera-Spatial.pdf

U.S. Senate Committee on Armed Services. (2012, May 21). Senate Armed Services Committee releases report on counterfeit electronic parts [Press release]. https://www.armed-services.senate.gov/press-releases/senate-armed-services-committee-releases-report-on-counterfeit-electronic-parts

Vitsœ. (n.d.). Dieter Rams: Ten principles for good design. https://www.vitsoe.com/us/about/good-design

World Privacy Forum. (2025). Privacy, identity and trust in C2PA: A technical review and analysis of the C2PA digital media provenance framework. https://worldprivacyforum.org/posts/privacy-identity-and-trust-in-c2pa/

Zhao, X., Zhang, K., Su, Z., Vasan, S., Grishchenko, I., Kruegel, C., Vigna, G., Wang, Y.-X., & Li, L. (2024). Invisible image watermarks are provably removable using generative AI. In Advances in Neural Information Processing Systems 37. https://doi.org/10.52202/079017-0276

© 2026 Laxmi Narayan Behera. Published by Curatone Art & Research Journal (ISSN 3054-6621) under a Creative Commons CC BY-NC-SA 4.0 licence.

Available online at: https://curatone.art/journal/less-but-verified-image-provenance