ROOT ZERO VAULT

Provenance Collapse Is a Governance Problem:

How Constitutional Infrastructure Enables Deterministic Media Verification in the Generative AI Era

Hosameldeen (Deen) Saleh
Founder & CEO, Root Zero Vault, Inc.
Designer, Recursive Stage-Based Identifier System (RSBIS)

Published: January 20, 2026
Correspondence: deen.saleh@rootzerovault.com

Abstract

Generative AI has collapsed the practical boundary between authentic and synthetic media. The resulting governance crisis is not merely that forgeries exist, but that courts cannot deterministically recompute provenance when evidence is contested—enabling both the liar's dividend (authentic media becomes deniable) and fraud proliferation (synthetic media becomes plausibly authentic).

We distinguish three provenance types: integrity (these bytes were committed at time T), attribution (the commitment is bound to a specific device/holder), and reality authenticity (the media depicts actual events). Current systems often conflate these, producing disputes that cannot be resolved by recomputation.

RSBIS provides deterministic integrity and cryptographic attribution via media Deeds bound to content commitments (CVIDs), device/holder attestations, tamper-evident editing lineage, and continuity bundles enabling offline verification independent of platforms or vendors. RSBIS makes integrity + attribution deterministically verifiable; it makes reality authenticity adjudicable by giving courts a cryptographic floor, not a full epistemology. Reality authenticity remains corroborative, but courts gain a deterministic foundation for admissibility.

A photojournalism defamation scenario shows how offline recomputation can authenticate integrity and attribution without relying on expert testimony alone, while preserving a clear boundary between verifiable provenance and factual claims about reality. Contesting party claims "AI deepfake"; court verifies camera signature + hash chain + editing transparency through deterministic recomputation. We include normative governance specimens demonstrating deterministic acceptance and the formal doctrine of "inadmissible ≠ fake" (Rule P-ADMIT): content without recomputable provenance is unverifiable and inadmissible for truth claims, without asserting falsity.

The contribution establishes constitutional provenance as structural verification replacing operational attestation, enabling courts to prove tampering mathematically while acknowledging the boundaries of what cryptography can determine about physical reality.

---

1. Introduction: When Courts Cannot Prove Provenance

1.1 The Dual Crisis

Crisis 1 - Liar's Dividend: Authentic evidence becomes deniable
"That video of me accepting bribes is AI-generated"

Crisis 2 - Fraud Proliferation: Synthetic content becomes plausibly authentic
"This CEO deepfake authorizing $25M wire transfer looks legitimate"

Both stem from same cause: provenance is operational attestation (trust platforms, trust experts) not structural verification (recompute offline, prove mathematically).

1.2 Documented Impact (Examples, Not Comprehensive Totals)

Deepfake proliferation trends: Industry trackers report sharp year-over-year increases in detected deepfake content (Sensity AI and others report order-of-magnitude growth 2022-2024, though methodologies vary). Cost collapse: sophisticated deepfake creation required $10K+ in compute and technical expertise (2019) versus mobile applications generating convincing results in seconds (2024).

Verified fraud cases (documented with primary reporting):

- **Hong Kong multinational firm:** Employee deceived via deepfake video conference call into transferring HK$200 million (~US$25.6M) to fraudulent accounts (February 2024; widely reported by Reuters, CNN, South China Morning Post)
- **UK energy company:** CEO voice clone used to authorize fraudulent €220,000 (~£243,000) wire transfer (2019; reported by Wall Street Journal, BBC)
- **FBI Internet Crime Report:** $3.5 billion in cryptocurrency fraud losses (2023), with increasing use of AI-generated personas and deepfake endorsements as attack vector (FBI IC3 Annual Report 2023)
- **Romance scams:** FTC reports $1.3 billion annual losses to romance fraud (2023), with growing deployment of AI-generated profiles and synthetic personas

Election interference (documented instances):

- Pakistan: Fake audio attributed to Imran Khan circulated during 2024 elections (verified by multiple fact-checking organizations)
- Indonesia: Candidate impersonation videos detected during presidential campaigns (Tempo, Jakarta Post reporting)
- United States: Robocalls using AI-generated voice impersonating President Biden during New Hampshire primary (January 2024; FCC investigation)

Journalism credibility concerns:

- Pew Research Center studies show declining trust in social media news accuracy, with concerns about misinformation rising among news consumers
- Provenance verification challenges cited by major news organizations (AP, Reuters, BBC) as operational concern requiring new authentication infrastructure

Note on aggregate figures: Various estimates of total economic impact ($50B-$100B+ annually) appear in media and industry reports but lack rigorous methodological documentation. Constitutional governance claim does not depend on contested aggregate figures—it depends on provenance being legally insufficient under current operational approaches, which documented cases demonstrate.

1.3 Why Current Approaches Fail

Metadata (EXIF/IPTC): Stripped via `exiftool -all=`, platforms remove on upload, trivially forgeable.

Watermarks: AI removal 80%+ success (Saberi 2023), screenshot bypass, transcoding loss.

Forensic detection: Non-deterministic, expensive ($200-500/hr), latest AI models fool tools, experts disagree.

C2PA (Adobe/Microsoft/BBC): Strong progress with industry consortium backing, major platform adoption, tamper-evident signatures. **What C2PA provides:** Manifest structure (JSON-LD), embedded signatures (JUMBF), ecosystem tooling, ingredient/claims model, existing adoption by Adobe Creative Cloud, newsroom workflows.

C2PA is necessary but not sufficient for court-grade determinism because adversaries can route around preserved manifests (screenshot laundering, platform migration to non-preserving systems, metadata stripping via conversion tools). RSBIS treats that circumvention as expected adversarial behavior and makes the provenance loss itself provable.

RSBIS + C2PA integration: C2PA = payload format (manifests, tooling, ecosystem); RSBIS = constitutional layer (identity lineage, immutable journals, offline recomputation, court-grade continuity surviving platform collapse).

What RSBIS inherits from C2PA: Manifest format, signing infrastructure, ingredient tracking, existing tooling ecosystem.

What RSBIS adds to C2PA:

- Offline recomputation independent of any single vendor (continuity bundles self-contained)
- Court continuity surviving platform/vendor failure (vendor bankruptcy doesn't break verification)
- Long-horizon verification logic (signature policy declarations enable verification across cryptographic transitions)
- "Inadmissible ≠ fake" governance outcomes (formal admissibility doctrine separate from truth claims)
- Cross-standard anchoring (IPTC, proprietary formats, blockchain timestamps all reference same RSBIS identity substrate)

1.4 The Governance Insight

Don't detect deepfakes perfectly. Make integrity provenance deterministically verifiable and attribution provenance cryptographically provable. RSBIS makes integrity + attribution deterministically verifiable; it makes reality authenticity adjudicable by giving courts a cryptographic floor, not a full epistemology.

Three provenance types (critical distinction):

Integrity: These bytes committed at time T; tampering detectable → **RSBIS provides deterministically** (pure mathematics: hash verification)

Attribution: Linked to device/holder → **RSBIS provides cryptographically** (signature verification under declared policy)

Reality authenticity: Depicted events actually occurred in physical world → **Requires corroboration** (witnesses, sensors, temporal/spatial consistency, motive analysis)

Boundary principle: RSBIS makes integrity + attribution recomputable without operational trust. Reality authenticity remains a factual question requiring defense-in-depth: provenance provides cryptographic floor for admissibility; courts evaluate reality claims through traditional evidentiary standards (multiple witnesses, corroborating evidence, expert analysis).

What courts gain: Mathematical certainty about tampering (integrity), cryptographic proof of source (attribution). **What courts still require:** Human judgment about whether depicted events actually occurred (reality authenticity).

1.5 Adversary Model

Attack 1 - Metadata stripping + re-encoding: Platform removes all provenance → Defense: Continuity bundle travels separately

Attack 2 - Screenshot laundering: Screenshot removes signatures → Defense: Creates new CVID; cannot claim original identity

Attack 3 - Honest camera, dishonest scene: Stage fake event, legitimate capture → Limitation: Cannot detect; requires witness corroboration

Attack 4 - Compromised device keys: Adversary extracts camera signing key through supply chain attack, side-channel analysis, or insider access → **RSBIS defense:** Device key compromise handled as first-class governance event:

yaml

```
key_compromise_handling:
  event_type: DEVICE_KEY_COMPROMISE_DECLARED
```

affected_device: Canon_EOS_R5_Serial_012345678901

compromise_timestamp: 2024-09-15T10:00:00Z

governance_response:

- DEVICE_ATTESTATION_QUARANTINED_FROM(2024-09-15T10:00:00Z)
- All signatures after timestamp require additional corroboration
- Pre-compromise signatures remain valid (temporal anchoring)
- Compromised device signatures flagged: E-SIG (revoked key)

admissibility_impact:

- Content created before compromise: Valid with primary attestation
- Content created after compromise: Inadmissible without secondary attestations (witness corroboration, multiple devices, forensic redundancy)

Governance handling stronger than economic arguments: Rather than claiming compromise "too expensive," RSBIS treats compromise as expected adversarial event with defined response. When detected, Journal records revocation; post-compromise content requires defense-in-depth. This moves from "trust our cost estimates" to "we have governance procedures when trust breaks."

Attack 5 - Synthetic claiming authentic: AI image forged camera metadata → Defense: Cannot forge hardware signature without factory key

Attack 6 - Liar's dividend: Authentic content falsely accused → Defense: Provenance shifts burden to claimant

---

2. Constitutional Provenance Architecture

2.1 Media Deeds with Cryptographic Binding

Canon EOS R5 captures image; hardware secure element creates attestation:

yaml

```
media_deed:
  identity: RootZero0634_Syria_Hospital_2024
  content_cvid: cvid:blake3:syria_8f3a9d2e...

  device_attestation:
    model: Canon_EOS_R5
    serial: 012345678901
    signature: sig:ed25519:Canon_Factory:4f7a...
    timestamp: 2024-08-15T14:23:47Z
    gps: [36.2021, 37.1343]
    settings: {ISO_400, f5.6, 1/500s}

  provenance_claim: Original_camera_capture
```

Properties: Any pixel change = different CVID (tampering detectable). Device signature binds content + timestamp + GPS (cannot forge without Canon factory key).

Multi-attestation architecture: Device attestation is a sufficient but not necessary attribution source. RSBIS supports multi-attestation provenance where policy thresholds determine admissibility:

- **Device + institutional signer:** Camera signature + newsroom editorial signature (defense-in-depth)
- **Device + witness:** Hardware attestation + human witness affidavit (corroborative)
- **Institutional only:** Newsroom signature without device attestation (legacy workflows, whistleblower submissions)
- **Policy-determined thresholds:** High-stakes cases may require multiple attestations; routine cases accept single device signature

This enables: (i) legacy device support (content without hardware signing gains institutional attestation), (ii) whistleblower protection (anonymous submissions verified through institutional vouching), (iii) gradual ecosystem adoption (partial attestation better than none).

Key security: Hardware secure element / Trusted Execution Environment (TEE) / secure enclave class attestation. Attack cost substantial (nation-state capability for secure element extraction). **Governance handling:** Compromise treated as first-class event with defined

response (revocation, temporal anchoring, secondary attestation requirements). Not claimed perfect; claimed detectable when occurs + procedural handling when detected.

2.2 Editing Transparency

NYT photo desk edits for publication:

yaml

```
derivative_deed:
  identity: RootZero06340 (child of 0634)
  parent_cvid: cvid:blake3:...8f3a...

  modifications:
    - CROP: [336,224,4800,3200]
    - EXPOSURE: +0.7_stops
    - COLOR: +200K_temperature

  derivative_cvid: cvid:blake3:...2c9f...
  editor_signature: sig:ed25519:NYT:7e2d...
```

Verification: Apply disclosed edits to parent → compute hash → matches derivative CVID? If yes: editing transparent. If no: E-CHAIN (provenance broken).

2.3 Offline Verification via Continuity Bundle

Bundle contains: Original Deed, derivative Deed, Journal entries, signatures, public keys. **Platform-independent:** Verifiable without NYT cooperation, platform access, or vendor servers. **Court verification:** Recompute hashes, verify signatures, check chain—pure mathematics.

2.4 C2PA Integration

RSBIS Constitutional Layer (identity, journals, offline recomputation)

↓ references via CVID

C2PA Payload Layer (manifests, tooling, ecosystem adoption)

What RSBIS adds: Platform-independent verification when C2PA lost, vendor-collapse resilience, offline court verification, cross-standard anchoring, temporal policy (algorithm transitions).

---

3. Photojournalist Defamation Case Walkthrough

Scenario: Ahmed captures Syria hospital bombing (Canon R5). NYT publishes with disclosed edits. Military commander sues: "AI deepfake to defame me." Court must prove: authentic or synthetic?

Phase 1: Capture (Aug 15, 2024)

Camera creates signed RAW: content CVID + device signature + timestamp + GPS. Media Deed issued: RootZero0634.

Phase 2: Editorial (Aug 16)

NYT verifies: Canon signature valid ✓, GPS matches location ✓, timestamp plausible ✓, forensics consistent ✓. Creates edited version: RootZero06340 with disclosed modifications.

Phase 3: Legal Challenge (Aug 17-20)

Commander: "AI-generated deepfake, modern models photorealistic, cannot prove authenticity." NYT: "Deterministic provenance via constitutional governance."

Phase 4: Court Verification (Sep 2024)

Independent expert applies formal verification procedure:

Court-Grade Verification Checklist (Deterministic + Corroborative Layers):

LAYER 1: DETERMINISTIC VERIFICATION (Integrity + Attribution)

Step 1: Recompute CVID from submitted bytes

→ Does computed hash match Media Deed commitment?

→ Result: MATCH ✓ or MISMATCH ✗

Step 2: Verify device/holder signature under declared policy

→ Extract signature + public key from continuity bundle

→ Cryptographic verification: signature covers (content + timestamp + metadata)

→ Result: VALID ✓ or INVALID ✗

Step 3: Verify chain continuity (parent → derivative)

→ Apply disclosed modifications to parent content

→ Recompute derivative hash

→ Compare to claimed derivative CVID

→ Result: CHAIN_VALID ✓ or CHAIN_BROKEN ✗

Step 4: Verify timestamps anchored (registry receipts / journal hash-chain)

→ Check journal hash chain unbroken from capture to present

→ Verify registry economic finality receipts

→ Result: ANCHORED ✓ or DISPUTED ✗

LAYER 2: CORROBORATIVE VERIFICATION (Reality Authenticity)

Step 5: Multi-source consistency (GPS/time vs known event timeline)

→ Cross-reference embedded GPS with event location

→ Verify timestamp plausibility relative to reported incident

→ Result: CONSISTENT or ANOMALOUS

Step 6: Independent witnesses/sensors (other cameras, satellites, logs)

→ Multiple photographers at scene? Satellite imagery available?

→ Traffic cameras, security footage, third-party documentation?

→ Result: CORROBORATED or UNCORROBORATED

Step 7: Motive/opportunity analysis for staging (human factors)

→ Could scene be artificially staged?

→ Does staging require implausible coordination?

→ Result: PLAUSIBLE_AUTHENTIC or PLAUSIBLE_STAGED

ADMISSIBILITY DETERMINATION

IF Layer 1 (Steps 1-4) ALL PASS:

→ Integrity + attribution deterministically verified

→ Content ADMISSIBLE for truth-claim purposes

→ Layer 2 evaluation informs weight of evidence

IF Layer 1 ANY FAIL:

→ Integrity or attribution cannot be recomputed

→ Content INADMISSIBLE under Rule P-ADMIT

→ Does not assert falsity; asserts unverifiability

Application to Syria hospital photograph:

Step 1: Compute CVID of original file → Matches claimed? **YES ✓** (unaltered)

Step 2: Verify hash-chain continuity (capture → edit → publish) → **VALID ✓**

Step 3: Verify Canon device signature → **VALID ✓** (cannot forge factory key)

Step 4: Check AI synthesis disclosure → **ABSENT** (claimed camera capture, not AI)

Step 5: Verify editing transparency → Apply disclosed edits → Hash matches derivative? **YES** ✓

Step 6: Forensic corroboration (supplementary) → No AI artifacts detected

Step 7: Temporal/spatial consistency → Photo 23min post-incident, GPS 50m from hospital ✓

Expert report: "Integrity provenance deterministic (unaltered since capture ✓). Attribution provenance cryptographic (Canon R5 signed ✓). Reality authenticity corroborated (timeline/location plausible ✓). Plaintiff 'deepfake' claim **unsupported by evidence**."

Court ruling: "Photograph authenticated via cryptographic provenance providing mathematical certainty unavailable through forensics alone. Burden-shifting: Defendant proved integrity deterministically; plaintiff provided no countervailing forgery evidence. Defamation claim **DISMISSED**."

Phase 5: Counterfactuals

If actually AI-generated:
Cannot forge Canon key (economically implausible). Options: (a) No signature = inadmissible, (b) Forge = detectable, (c) Honest disclosure = legal (art/satire).

If camera captured staged scene:
Device signature authentic, but scene dishonest. RSBIS limitation: Cannot detect staging. Requires: multiple witnesses, angles, temporal consistency, third-party verification (satellites, UN observers).

---

4. What Constitutional Provenance Does NOT Do

Provides deterministically:

- ✓ Integrity (content unaltered since commitment)
- ✓ Attribution (linked to device/holder cryptographically)
- ✓ Editing transparency (modifications disclosed, verifiable)
- ✓ Cross-platform portability

- ✓ Offline court verification

Does NOT provide:

- ✗ Reality authenticity determination (cannot prove events occurred; requires corroboration)
- ✗ Perfect key extraction prevention (nation-state attacks theoretically possible; economically implausible + detectable)
- ✗ Staged scene detection (legitimate camera photographs fake events)
- ✗ Automatic platform enforcement (platforms choose preservation)

Critical: Integrity provenance (RSBIS) + reality authenticity (corroboration) = defense-in-depth.

---

5. Canonical Specimens

RSBIS Reason Code Glossary (governance outcomes):

- **E-FORMAT:** Missing required provenance structure (no device signature, no CVID commitment, missing synthesis disclosure)
- **E-MODEL:** Content hash mismatch (tampering detected: submitted bytes ≠ committed CVID)
- **E-CHAIN:** Provenance chain broken (derivative recomputation doesn't match parent + disclosed edits)
- **E-SIG:** Signature verification failed (invalid signature, wrong key, revoked key, expired certificate)
- **E-SCOPE:** Access/modification outside declared policy scope
- **E-IMMUTABILITY:** Attempt to alter immutable governance record (Journal tampering, Deed modification)

Formal Admissibility Doctrine (Rule P-ADMIT):

Rule P-ADMIT (Provenance-Based Admissibility): If integrity + attribution provenance cannot be deterministically recomputed from continuity artifacts, the content is unverifiable and treated as inadmissible for truth-claim purposes under constitutional governance standards, without asserting the content is false.

Rationale: This doctrine directly counters the "liar's dividend" problem. Courts do not argue "this content is fake" (which may be unprovable). Courts argue "this content does not meet the verifiability threshold for deterministic authentication." Burden shifts to claimant: provide recomputable provenance or content inadmissible.

Distinction: "Inadmissible" ≠ "fake." Authentic content that loses provenance (through screenshot laundering, platform migration, metadata stripping) cannot be verified deterministically. Alternative verification paths remain (forensic analysis, witness testimony, circumstantial evidence), but constitutional governance requires structural verification as admissibility foundation.

Acceptance (deterministically verifiable provenance):

A1: RootZero0240020700_Camera_Capture_Verified

- Original RAW file with hardware device signature
- Content CVID matches file hash (integrity ✓)
- Device signature cryptographically valid (attribution ✓)
- No editing (original unaltered)
- **Outcome:** ADMISSIBLE - Authentic camera capture deterministically verified

A2: RootZero0240020701_Transparent_Edit_Disclosed

- Derivative of camera original
- Parent CVID referenced correctly
- Modifications explicitly listed (CROP, EXPOSURE, COLOR)
- Recomputation: apply disclosed edits to parent → matches derivative CVID (chain ✓)
- Editor signature valid
- **Outcome:** ADMISSIBLE - Authentic edited derivative with transparent modification history

A3: RootZero0240020702_AI_Synthesis_Disclosed

- AI-generated content
- Model + prompt + parameters disclosed
- Authenticity claim: AI_Synthetic (not camera capture)
- No camera signature present (correctly absent for AI)
- **Outcome:** ADMISSIBLE - Honest AI art disclosure (clearly identified as synthesis)

Rejection (inadequate or fraudulent provenance):

R1: RootZero0240020710_Screenshot_Laundered

- Content CVID different from claimed original (pixel-level changes from recompression)
- No device signature present (metadata stripped)
- Claimed to be "original photograph" but lacks attestation
- **Outcome:** INADMISSIBLE under Rule P-ADMIT (not provable fake, but unverifiable) → E-FORMAT
- **Note:** This is the critical "inadmissible ≠ fake" case. Content might be authentic but provenance lost through adversarial laundering.

R2: RootZero0240020711_Tampered_Content_Detected

- Claimed original CVID: cvid:blake3:...8f3a...
- Actual file hash: cvid:blake3:...DIFFERENT...
- Hash mismatch = content altered post-commitment
- **Outcome:** TAMPERED (integrity provenance broken) → E-MODEL

R3: RootZero0240020712_Broken_Provenance_Chain

- Derivative claims parent: cvid:blake3:...8f3a...
- Apply disclosed edits to parent → cvid:blake3:...MISMATCH...
- Derivative CVID doesn't match recomputation
- **Outcome:** PROVENANCE_BROKEN (editing not transparent; undisclosed modifications) → E-CHAIN

R4: RootZero0240020713_Undisclosed_Synthesis

- Content claims "camera capture" (authenticity_claim: Original_photographic)
- No device signature present (required for camera claim)
- Forensic analysis suggests AI generation artifacts
- Missing synthesis_disclosure field (required for AI content)
- **Outcome:** DISCLOSURE_VIOLATION (fraud attempt: synthetic claiming authentic) → E-FORMAT

R5: RootZero0240020714_Forged_Signature

- Device signature present in file

- Signature cryptographic verification fails (doesn't match content + public key)
- Possible causes: wrong public key, signature doesn't match content hash, revoked key, expired certificate
- **Outcome:** FORGERY_DETECTED → E-SIG

R6: RootZero0240020715_Post_Compromise_Attestation

- Device signature present and cryptographically valid
- BUT: Device key revoked via Journal (DEVICE_KEY_COMPROMISE_DECLARED)
- Content timestamp: after compromise timestamp
- **Outcome:** INADMISSIBLE (compromised key; requires secondary attestations per governance policy) → E-SIG (revoked)

Key governance insight: Specimens R1 and R6 both result in inadmissibility but for different reasons. R1 = provenance lost (might be authentic, cannot verify). R6 = compromised source (might be authentic, requires additional corroboration). Both demonstrate "inadmissible ≠ fake" principle.

6.5 Limitations and Open Questions

Acknowledged limitations (not failures, but boundaries):

Legacy media adoption: Billions of images/videos exist without device attestation. RSBIS provides governance for new content; legacy content requires institutional vouching or remains unverifiable under deterministic standards.

Mixed-provenance ecosystems: Transition period where some platforms preserve provenance, others don't. Screenshot laundering remains possible; RSBIS makes provenance loss detectable, not preventable.

Human factors in disclosure enforcement: Honest AI disclosure requires cultural shift (artists/creators voluntarily labeling synthetic content). Legal mandates may be required; voluntary compliance uncertain.

Jurisdictional variance in admissibility: Different courts may set different thresholds for Rule P-ADMIT application. Constitutional governance provides framework; jurisdictional adoption varies.

Device manufacturer coordination: Camera signing requires hardware manufacturers (Canon, Nikon, Sony, Apple, Samsung) to implement secure elements. Adoption timeline depends on industry coordination, not technical feasibility.

Whistleblower scenarios: Anonymous sources cannot provide device attestation without revealing identity. Institutional attestation (newsroom vouching) becomes sole attribution source—acceptable under multi-attestation framework but weaker than device + institutional.

Deepfake detection evolution: As generative models improve, forensic detection becomes less reliable. RSBIS provides governance foundation regardless of detection capabilities, but reality authenticity evaluation remains challenging when staging sophisticated.

Economic incentives for adoption: News organizations, camera manufacturers, platforms must invest in provenance infrastructure. Business case depends on liability reduction, trust restoration, regulatory mandates—not guaranteed without policy intervention.

Open questions for future work:

- **Optimal attestation thresholds:** What combination of device/institutional/witness attestations provides sufficient admissibility confidence for different case types?
- **Temporal degradation:** How long do device signatures remain trustworthy before key extraction risk requires additional corroboration?
- **Cross-jurisdictional recognition:** Will courts across different legal systems recognize RSBIS provenance as admissible evidence?
- **Platform liability:** Should platforms be required to preserve provenance metadata, or is screenshot laundering accepted adversarial behavior?
- **AI synthesis disclosure enforcement:** Should undisclosed AI generation be legally prohibited, or handled through civil liability only?

---

7. Impact and Deployment

Deployment ladder:

Phase 1 (2025-27): High-liability first - courts require provenance for high-stakes evidence, insurance mandates for $100K+ claims, news organizations (AP, Reuters, NYT, BBC) adopt.

Phase 2 (2026-28): Camera manufacturers embed signing - Canon/Nikon/Sony flagship models, Apple/Samsung smartphones add secure elements.

Phase 3 (2027-29): Platform ecosystem - Meta/X preserve provenance where available, YouTube/Vimeo support verification, Adobe/Final Cut maintain through workflow.

Phase 4 (2028-30): Legal standards - Federal Rules of Evidence require provenance for authentication, state deepfake laws reference RSBIS, UN/OSCE adopt frameworks.

---

8. Conclusion

The crisis: courts cannot deterministically prove provenance. Result: authentic becomes contestable (liar's dividend), synthetic becomes plausible (fraud proliferation).

Constitutional infrastructure distinguishes three provenance types: integrity (deterministic via RSBIS), attribution (cryptographic via RSBIS), reality (corroborative, requires defense-in-depth).

Courts verify integrity + attribution mathematically through offline recomputation. Reality authenticity through traditional evidence (witnesses, sensors, consistency). Honest AI disclosure transparent; fraudulent claims detectable.

Provenance is not detection problem (forensics arms race) but governance problem (structural verification). With constitutional infrastructure, tampering becomes mathematically detectable, attribution cryptographically provable, and honest disclosure transparently distinguishable from fraud.

Constitutional infrastructure applicability: This provenance layer shares structural foundations with other governance domains requiring deterministic validation under explicit policy with permanent, recomputable evidence.*

*See Root Zero Deed specification (UnknownMe_RootZeroDeed_V39) for complete problem taxonomy addressing sixteen governance challenges including digital inheritance, supply chain fraud, refugee identity, research integrity, environmental accountability, healthcare interoperability, and election verification—all utilizing the same validation infrastructure, reason code taxonomy, and offline recomputation principles demonstrated in this paper.

Correspondence: deen.saleh@rootzerovault.com