The Authenticity Protocol

Inverting the Burden of Proof in the Age of Synthetic Media

Aniket Dash
Machine Learning Engineer Researcher
aniket.addash@gmail.com

January 2026

Executive Summary

The rise of generative AI has fundamentally compromised the traditional mechanism of digital trust. As synthetic media achieves parity with human perception, the industry's reliance on "AI Detection" has led to a technically asymmetric arms race. This whitepaper introduces the **Creator Passport**, a platform-agnostic framework designed to shift the burden of proof from detection to verification. By implementing the C2PA standard through a modular "Stripe-for-Trust" architecture, the protocol enables creators to cryptographically sign their work and platforms to prioritize verified human content. The goal is to establish a new social contract for the internet: one where authenticity is a premium, provable currency.

Contents

The Crisis: The Collapse of Default Trust

For decades, "seeing was believing." The digital transformation of information relied on the assumption that media captured by a lens or recorded by a microphone was, by and large, an objective reflection of reality. Generative AI has permanently shattered this assumption.

The AI Detection Trap

Current responses to synthetic media focus on "Detection", heuristic-based algorithms that search for artifacts or statistical patterns indicating AI generation. This approach is inherently flawed:

- **The Asymmetry of Progress:** Generative models (GANs, Diffusion) improve exponentially through adversarial training. Detectors, which rely on identifying known flaws, will always lag behind.
- **The False Positive Penalty:** As detectors become more aggressive, they inevitably flag human-created art and journalism, effectively "censoring" authenticity.
- **Computation vs. Cryptography:** Detection is a probabilistic guess, verification is a mathematical certainty.

Context Collapse

The problem is not just the content (the "What") but the lack of provenance (the "Who" and "How"). A "real" photo distributed by a malicious actor to spread disinformation is as damaging as a deepfake. Digital trust requires knowing the source, the intent, and the integrity of the asset.

The Lifecycle of Authenticity

The Authenticity Protocol defines a four-stage operational lifecycle for media, ensuring provenance is preserved from capture to consumption.

1. Capture: Hardware-Level Identity

Trust begins at the sensor. We advocate for and implement the C2PA (Coalition for Content Provenance and Authenticity) standard, where hardware manufacturers (Nikon, Sony, Leica) cryptographically sign raw assets at the moment of capture.

- **Action:** Binding a unique hardware ID to the initial image hash.
- **Result:** A tamper-proof "Genesis Manifest" for the media.

2. Soft-Binding: The Creator Passport

In a transition period where hardware signing is not yet ubiquitous, the Creator Passport acts as a software bridge. It allows creators to "soft-bind" their verified identity (x.509 certificate) to existing assets.

- **Action:** Issuing digital certificates to verified human users.
- **Utility:** Creating a point-of-ingestion signature for assets before they enter distribution networks.

3. Edits and Iteration: The Provenance Log

Authentic media is rarely raw. The protocol ensures that edits (cropping, color correction, but not generative addition) are appended as signed manifest ingredients.

- **Integrity:** If a manifest is stripped, the asset is flagged as "Incomplete Provenance."

4. Consumption: The Truth Lens

The final stage is the transparent display of this data. The Truth Lens SDK allows social platforms and news aggregators to display a "Nutrition Label" for media, surfacing credibility signals directly to the end-user.

System Architecture: "Stripe for Trust"

The Authenticity Protocol is designed as a ubiquitous infrastructure layer that any platform can integrate.

The Trust Engine (Backend)

A containerized microservice that orchestrates the cryptographic heavy lifting:

- **Registration API:** Handles identity verification and certificate issuance.
- **Signing Service:** Ingests media and appends C2PA-compliant manifests.
- **Verification API:** Analyzes incoming assets and returns structured "Trust Signals."

The Integration Layer (SDK)

A lightweight JavaScript SDK that enables platforms to verify content with a single line of code.

```
<script src="auth-protocol.js"></script>
<img src="photo.jpg" data-c2pa-verify>
```

Credential Identity Checklist

- [ ] Digital Identity verified against government or social ID.
- [ ] x.509 Certificate issued with RSA-2048 or EC signatures.
- [ ] Private keys stored in secure enclaves (HSM/TPM).
- [ ] Manifests signed using C2PA JWS (JSON Web Signatures).

Policy and Industry Impact

The technical implementation is only half the battle. The Authenticity Protocol proposes a paradigm shift in regulatory and platform policy.

Credibility as a Ranking Signal

Platforms like Instagram and Facebook currently optimize for engagement. We propose optimizing for **Originality**.

- **The Human Tier:** A dedicated feed or "Verified" badge for content with a complete, signed provenance chain.
- **Ranking Boost:** Signed media receives higher distribution weight than unverified synthetic content.

Informing the Social Contract

By surfacing "Nutrition Labels," we empower the user to make informed decisions. We move from a "Default Trust" environment (prone to exploitation) to a "Default Skepticism" environment (resilient to deepfakes).

Conclusion

The era of "AI Detection" is a defensive, losing battle. The Authenticity Protocol represents an offensive stance: rebuilding the foundations of digital reality through cryptographic proof. By providing a platform-agnostic "Stripe-for-Trust" layer, we ensure that human creativity, journalism, and truth remain the premium currency of the internet.