Beyond the Visual Bias: Why Synthetic Media Provenance Must Be Multi-Modal[1]

Eddan Katz and Erik Svilich
Encypher Corporation
eddan.katz@encypher.com, erik.svilich@encypher.com

***Abstract*—Content provenance research and regulation have developed around a visual assumption: synthetic media means synthetic images and video, and detection is a signal-processing problem. The resulting regulatory landscape—from the EU AI Act to state-level US mandates—operationalizes provenance for visual media while leaving text, the dominant medium of AI-generated content, without equivalent infrastructure. This paper argues that the next phase of content provenance must be multi-modal, with text at its center, and that the C2PA open standard is the first to bring two historically distinct functions—authentication and declared history—together in a vendor-neutral form, creating an authenticated substrate over which a third function, metering, can be exercised by external observation. That layering, rather than any single implementation, is what creates conditions for business models prior provenance approaches could not sustain. We situate this convergence against statistical watermarking, capture-based credentials, and access-control gateways; analyze its adversarial robustness; and outline a research agenda for the synthetic media detection community.**

***Index Terms*—content provenance, text watermarking, C2PA, multi-modal authentication, synthetic media, AI governance, content credentials.**

I. Introduction

The first International Symposium on Synthetic Media Attribution and Detection (ISSMAD 2025) arrived at a moment when detection-focused approaches to synthetic media were reaching their practical limits. Generative adversarial networks, diffusion models, and large language models have driven the quality of synthetic content past the threshold of reliable human detection. A 2026 ACM study using over 2,000 human judgments found that participants could not reliably distinguish LLM-generated news articles from human-written ones across multiple models [1]. A parallel expert survey identified what the authors call a "verification crisis": the detection arms race is structurally unsustainable, and the research community must shift from reactive detection to proactive provenance infrastructure [2].

This paper accepts that premise and extends it. The provenance infrastructure now taking shape—anchored by the Coalition for Content Provenance and Authenticity (C2PA) specification—has been designed, tested, and deployed primarily for visual media. Text was added to the C2PA specification only in January 2026, as Section A.7 of C2PA 2.3, renumbered Section A.8 in C2PA 2.4 after HTML was added as A.7 [3]. The delay was not an oversight. It resulted from a specific set of technical, institutional, and disciplinary choices. This paper identifies those choices and traces their consequences through regulatory instruments, industry surveys, and academic research agendas.

The paper then argues that three functions historically served by separate cryptographic traditions—authentication, chain of custody, and metering—now stand in a new relation to one another. The first two converge inside a single open standard for the first time. The third, which no specification performs, becomes available to any observer willing to verify against that standard. It is the combination—an open authenticated substrate plus unrestricted external observation—that supports business models detection-only approaches and proprietary watermarking schemes cannot.

Between this paper's submission and the symposium, the argument acquired its first public test. A frontier provider disclosed a planned two-part marking architecture for supported models under the EU AI Act's Article 50(2) Code of Practice: an embedded model-level watermark for generated text, and C2PA-signed provenance metadata for supported files [14]. The technical details of the text mark, and of the mechanism by which it is detected, remain forthcoming. Text was assigned the detection layer; the governance-bearing layer was assigned where metadata containers already exist. Section V-E explains why that assignment is the rational one at the model layer, and why it does not displace publication-time provenance.

Positions. We advance four positions for ISSMAD 2026:

1) The visual bias in synthetic media research is structural, not incidental, and must be repaired before text-dominant generative content overwhelms existing detection infrastructure.
2) Provenance is not a single function but a convergence of three—authentication, chain of custody, and metering—that previously lived in separate, often proprietary, systems.
3) C2PA is the first open standard in which authentication and chain of custody are interoperable and not controlled by a single vendor, and the first over which metering can therefore be performed by any observer rather than by a gatekeeper; this architectural fact, not any specific implementation, is what enables sustainable business models for provenance infrastructure.
4) Detection research is most valuable when integrated into this provenance substrate, where forensic findings can be attributed, traced, and metered rather than merely classified.

[1] Position paper presented at the International Symposium on Synthetic Media Attribution and Detection (ISSMAD 2026). The authors are affiliated with the reference implementation discussed in Section IV-F and cited as reference [15].

II. The Visual Bias: How Text Was Excluded

A. Disciplinary Origins

The term "deepfake" entered public discourse in late 2017, naming a specific practice: using deep neural networks to swap faces in video. The research community that formed around this problem—and that produced the foundational survey literature—was rooted in computer vision and signal processing. DARPA's Media Forensics (MediFor) program (2016–2020) and its successor, Semantic Forensics (SemaFor), funded detection research almost exclusively for images and video [4]. The academic conferences that incubated this work—CVPR, ICCV, ACM Multimedia—are signal-processing venues. The problem definition imported the assumptions of the discipline: synthetic media is a signal that can be analyzed in the frequency domain, and detection means finding statistical anomalies in pixel distributions, compression artifacts, or temporal inconsistencies.

Text does not have pixels. It does not have a frequency domain in the signal-processing sense. It cannot be analyzed through the forensic methods that image and video detection research developed. When the synthetic media detection community surveyed the landscape, text was either absent from the frame or treated as a secondary concern addressable through stylometry or statistical watermarking.

B. File-Format Affordances

The C2PA specification, jointly developed by Adobe, Microsoft, Intel, Arm, and Truepic beginning in 2019, was built around file formats that support embedded metadata containers. JPEG, PNG, TIFF, MP4, and other binary media formats have defined structures—EXIF headers, XMP metadata, ISO BMFF boxes—into which a JUMBF (JPEG Universal Metadata Box Format) manifest can be placed. The C2PA manifest sits inside the file, signed with a cryptographic certificate, and travels with the asset through distribution.

Plain text has no equivalent structure. A UTF-8 text file is a sequence of bytes with no native metadata container, no header, and no box structure. The C2PA specification initially addressed this gap by proposing external manifest storage: the text asset's hash would serve as a lookup key in a remote manifest repository [3]. This approach works for controlled environments but breaks under copy-paste, syndication, scraping, and redistribution through AI training pipelines. Section A.7, published January 8, 2026 and renumbered Section A.8 in C2PA 2.4, addressed unstructured text with an in-band mechanism: a manifest wrapper encoded into the text itself through non-rendering Unicode Variation Selectors and bound to the content by a data-hash assertion, with sidecar manifest references as an alternative [3]. The operative distinction is therefore native container versus in-band embedding, not the presence or absence of a mechanism. This was a necessary extension, but it arrived six years after the specification's visual-media foundations were laid.

C. Regulatory Tracking

Regulatory instruments tracked the research community's assumptions. The EU AI Act's Article 50 transparency obligations, finalized in 2024, require that AI-generated content be marked in a "machine-readable" way. The implementing Code of Practice is modality-general in its commitments: the marking obligation reaches audio, image, video, and text alike. It distributes the governance-bearing measures by file-format affordance, however. Digitally signed metadata is required where content is "generated or exported in a data format that supports adding information as part of the metadata," the examples given being audio, image, video, and document files; for free text and other outputs that cannot host metadata, signatories are only encouraged to offer a downloadable signed manifest, and richer provenance information under Measure 1.3 is optional throughout. The commitment that reaches text is the imperceptible watermark, and it reaches text with qualifications of its own: the final Code requires watermarking for free-form text longer than 200 tokens, excepts very short text, allows that free-form text watermarks may have lower reliability, and permits access to the corresponding detection mechanisms to be restricted for a period on that basis [5].

The bias is therefore subtler than exclusion. Text is marked, but with the layer that carries neither authentication nor declared history, while the layer that carries both is conditioned on a format that supports metadata. The first signatory architecture to be disclosed follows exactly this contour: a frontier provider's announced plan assigns a model-level watermark to generated text and C2PA-signed metadata to supported files [14]. That provider also states that the marking is not limited to the Union, which would turn an EU instrument into a global default through a provider's implementation choice.

Washington State HB 1170, signed March 2026 and effective February 1, 2027, is the first US state law mandating provenance detection tools [6]; its definition of "synthetic media" is broad enough to include text, but the supporting technology landscape assumed visual content as the primary target. China's mandatory AI content labeling measures (effective September 2025) similarly operationalize marking requirements through visual indicators and metadata structures [7]. The pattern is consistent: regulation followed research, and research followed the disciplinary frame established by the visual deepfake problem. Text was not deliberately excluded—it was simply not present in the technical infrastructure that the instruments assumed, and where the instruments now reach it, they reach it with the weakest of the three functions.

III. Three Functions Coming Together

Content provenance, as now implemented through C2PA, is not a single cryptographic function. It is the convergence of three functions that developed independently in separate traditions over decades. Two of them, authentication and chain of custody, inhabit a single open standard for the first time. The third, metering, is performed by no specification at all; it is performed by observers, and what C2PA changes is the

substrate those observers can verify against. Fig. 1 places these functions in relation to one another and to prior systems.

Figure 1 — Venn diagram: Authentication, Chain of Custody, and Metering as distinct cryptographic functions
[artwork to be inserted]

Fig. 1. Three governance-bearing cryptographic functions. Prior systems occupied subsets; C2PA is the first vendor-neutral open standard in which authentication and chain of custody are interoperable, and the first over which metering can be performed without the standard's own permission.

A. Authentication

Authentication is the cryptographic assertion of who or what produced a given artifact. This function descends from public-key infrastructure (PKI), X.509 certificates, and digital signature schemes developed for secure communications beginning in the 1970s. In the C2PA context, authentication means that a claim bound to an asset is signed by a certificate traceable to a trust anchor, establishing that the signer—a camera, a newsroom CMS, an AI model, a publisher—asserts something about that asset and can be held to the assertion. A valid signature establishes who signed the claim; it does not by itself establish authorship, ownership, or the truth of what is claimed. Authentication alone answers the question *from where?* It is the precondition for any liability claim, any trust decision, and any downstream verification.

The distinction between an attribution and a bare signal is now stated in first-party industry documentation. The guidance accompanying the first disclosed Article 50 text-marking architecture warns that a detected mark shows only that the content may have been processed by the model, because users route human-authored material through it to proofread, translate, or summarize [14]. A mark without an authentication layer carries no assertion of who is claiming what, so it cannot distinguish authorship from processing, and it misattributes in both directions: stripped provenance underclaims machine involvement, and an authorship-agnostic mark overclaims it. Institutions that read a probabilistic mark as a verdict are not misreading a good signal; they are supplying by assumption the authentication layer that the signal lacks.

B. Chain of Custody: Declared History

Chain of custody is the ordered record of how content was created, modified, and distributed, as declared by the actors who performed those operations. It draws on Merkle tree constructions (1979), hash chains, and the forensic chain-of-evidence principles used in legal proceedings. The word declared is load-bearing: what a manifest chain contains is a set of signed assertions by claim generators, not an independently witnessed history, and the distinction matters whenever a manifest is offered as evidence. In C2PA, the provenance chain is a linked sequence of manifests, each signed by the actor who performed a given operation—capture, edit, resize, AI-assisted modification, publication—and answers the question *through what?* For text, custody takes on particular importance. An article published by a wire service may pass through a CMS, a B2B licensee, an aggregator, a content farm, and an AI training pipeline. Document-level C2PA manifests establish provenance for the article as a whole; granular text authentication, in which sub-document units are independently verifiable against a commitment recorded in the manifest, extends declared history to the granularity at which AI systems actually consume text [8]. Verification requires that the signing and verifying parties divide the normalized text identically. That requirement is a security property rather than an implementation convenience, and the division method is recorded in the manifest so that any conforming verifier can reproduce it.

C. Metering

Metering is the systematic observation and counting of how provenanced content is used. It descends from digital rights management (DRM), broadcast monitoring, and the telemetry systems used in advertising and content licensing. Metering answers *how often, by whom, and at what scale?* DRM systems of the early 2000s bundled metering with access control, producing consumer backlash and interoperability failures. It is necessary to be precise about where this function lives, because the convergence claim is easy to overstate at exactly this point. C2PA does not meter. The specification supplies signed claims, hard bindings to assets, declared ingredients and actions, provenance-related assertions, and validation machinery; it does not crawl the web, observe copying, count uses, aggregate detection events, or produce a usage ledger. What it supplies instead is an authenticated substrate that external observation can verify against. A signed article that appears on a content farm, in a RAG retrieval cache, or in an AI model's training corpus carries its provenance metadata with it; a monitor that encounters and verifies that metadata generates a verification event, and the aggregation of those events into a record of frequency, distribution path, and consumption context is the metering function. Metering is therefore built on the standard rather than contained in it. What the standard does contribute is the absence of a constraint: because content credentials neither prevent copying nor gate verification, observation is open to anyone, and no party can be excluded from metering by the party that signed. That separation from access control is architecturally significant—provenance can be deployed without restricting content flows, a prerequisite for adoption by publishers whose business model depends on wide distribution, and it is why metering over a C2PA substrate does not reproduce DRM's failure.

D. The Convergence

Each function is independently useful. Authentication without custody is a signature. Custody without metering is forensics. Metering without authentication is telemetry. When the first two inhabit a single open standard and the third can be exercised against it without permission, together they constitute a governance interface. Previous content-protection systems combined some of these functions but in proprietary, closed architectures. Broadcast watermarking systems (Teletrax, Digimarc) provided metering and authentication but not custody, and operated as vendor-specific services rather than open standards. DRM systems (FairPlay, Widevine) combined

authentication with access control but locked metering data inside platform silos. Statistical watermarking for AI-generated text—SynthID [9], and the model-level text marking announced under the Article 50 Code [14]—provides a detection signal but no authentication certificate and no declared history, and supports no metering infrastructure of its own; its robustness limits are examined in Section IV. C2PA is the first open specification in which authentication and chain of custody are available, interoperable across implementations, and not controlled by a single vendor, and the first over which the metering function can be exercised by any party able to verify a signature.

E. Cryptographic Primitives

C2PA mandates COSE signatures (RFC 9052) over the manifest claim, with permitted algorithms ES256 (ECDSA P-256), ES384 (ECDSA P-384), ES512 (ECDSA P-521), and EdDSA (Ed25519). Certificate chains follow X.509, with verifiers consulting a designated trust list of issuing authorities. Each claim binds the manifest to the asset through one of two mechanisms: a hard binding, which is a cryptographic hash (typically SHA-256) of the asset's content bytes, and a soft binding, which is a perceptual or content-derived identifier tolerant of permitted transformations such as re-encoding or resizing. For text under Section A.7 (Section A.8 in C2PA 2.4), granular text authentication commits sub-document units to a single value recorded in the manifest claim as a hard binding, with per-unit markers sufficient to verify any one unit independently against that commitment. Streaming generation from LLMs requires an incremental construction that finalizes on end-of-sequence. The specific construction, normalization profile, and marker encoding used by the reference implementation in [15] are outside the scope of this paper.

IV. Comparative Landscape

The three-function framework provides a basis for comparing the current generation of provenance, watermarking, and access-control systems. Table I summarizes nine representative systems against each function and against four operational dimensions—paraphrase survival, open-standard status, whether the system requires AI-vendor cooperation, and granularity.

TABLE I

System	Au	Cu	Me	Pa	Op	Co	Gr
C2PA text (A.7/A.8, granular)	Y	Y	Ext	P	Y	N	D/S
Adobe Content Credentials	Y	Y	N	—	Y	N	D
Truepic capture credentials	Y	Y	N	—	Y	N	Cap
Numbers Protocol	Y	Y	P	—	Y	N	Ast
Google SynthID Text	N	N	P	lim	N	Y	Tok
Art. 50 model-provider text marking	N	N	P	deg	N	Y	Tok
Originality.ai (post-hoc)	N	N	N	deg	N	N	D
Stylometry / forensic	N	N	N	deg	—	N	D
robots.txt + TollBit / AI Crawl	N	N	Y	—	P	Y	Acc

Comparative landscape of text-relevant provenance and detection systems. Columns: Au = authentication; Cu = chain of custody; Me = metering; Pa = behavior under paraphrase/translation; Op = open standard; Co = requires AI-vendor cooperation; Gr = granularity. Values: Y = yes, N = no, P = partial, Ext = available only through observation external to the system, — = n/a, lim = limited, deg = degrades; D = document, S = segment, Cap = capture event, Ast = asset, Tok = token, Acc = access. The Article 50 model-provider row characterizes the text layer of the first Code architecture to be disclosed [14]; the file layer inherits the C2PA row's properties for the formats it covers. Metering is never a property of the specification named in a row: where a row shows metering, the function is performed by observation and aggregation outside that system.

A. Capture-Based Credentials: Adobe and Truepic

Adobe's Content Credentials and Truepic's capture-pipeline credentials both implement C2PA for visual media. Both provide strong authentication (PKI-anchored signatures) and custody (linked manifest chains across editing operations). Neither implements metering at the platform level; verification is offered as a public good, but downstream usage telemetry is left to individual signers. As evidence for the three-function thesis, these systems demonstrate that the C2PA substrate supports authentication and custody at production scale—and that metering is a deliberate extension built on that substrate rather than a missing feature of it.

B. Blockchain-Anchored Provenance: Numbers Protocol

Numbers Protocol layers a blockchain commitment over C2PA manifests, providing immutable timestamping and a public usage ledger that approximates partial metering. The blockchain anchor strengthens custody at the cost of additional infrastructure and per-write cost. As evidence for the three-function thesis, Numbers Protocol shows that the open standard composes with external commitment systems without sacrificing interoperability—different vendors can layer different metering and audit substrates atop the same C2PA-signed asset.

C. Statistical Watermarking: Google SynthID Text

SynthID Text [9] modulates the token-sampling distribution of an LLM at generation time, producing an output whose statistical signature can later be detected. Detection is open-sourced, but the watermark itself is applied only by participating models. Generation-time text watermarking is no longer a single-vendor practice: a second frontier lab announced model-level text marking for supported models in August 2026 under the Article 50 Code, and further Code signatories are expected to follow [14]. The mechanism of that second mark has not been disclosed—the announcement does not state whether it uses C2PA unstructured-text embedding, coexists with a manifest, links to a remote one, interoperates with C2PA, or is entirely proprietary—but the governance profile it describes reproduces the SynthID profile: a detection

signal with no authentication certificate and no declared history, paired with C2PA metadata for supported files. Section V-E takes up that division. Research has shown the statistical signal degrades substantially under paraphrasing, translation, and substantive editing [10]. As evidence for the three-function thesis, generation-time watermarking illustrates that marking without authentication is a probabilistic detector rather than a governance interface—the inverse of cryptographic provenance without watermarking.

D. Post-Hoc Classifiers: Originality.ai and Stylometry

Originality.ai and similar commercial classifiers, together with academic stylometric methods, operate entirely post-hoc and require no cooperation from the generating model. They are detection-only: no authentication, no custody, no metering. Accuracy degrades on shorter texts, paraphrased content, and human-edited LLM outputs, and false-positive rates on human writing remain a documented concern. As evidence for the three-function thesis, post-hoc classifiers are the rear-guard regime the introduction described—still necessary in environments where provenance has been stripped, but unable on their own to sustain licensing, compliance, or attribution at scale.

E. Access-Control Gateways: robots.txt, TollBit, AI Crawl Control

Opt-in access gates and revenue-sharing marketplaces (robots.txt extensions, TollBit, Cloudflare AI Crawl Control, ProRata) provide metering and access logging but no cryptographic authentication or custody of the underlying content. They function only when AI companies voluntarily participate—a cooperation-dependent control that breaks when crawlers ignore directives or when content is redistributed off-origin. As evidence for the three-function thesis, access gateways show that metering alone, without authentication and custody embedded in the content itself, cannot enforce ownership claims downstream of the cooperation boundary.

F. A Reference Implementation for Text

One reference implementation [15] demonstrates that the open standard can deliver all three functions for text. Authentication operates through C2PA certificate chains. Declared history operates through granular text authentication and linked manifests. Metering is performed by neither the standard nor the implementation's signing path; it depends on downstream observation of marked content, with verification events aggregated into per-document transparency log entries recording where, when, and how often signed content was encountered—the quantitative usage record that content licensing and compliance require. The implementation covers 31 MIME types across six media categories—text, images, audio, video, documents, and fonts—combining C2PA manifests with modality-appropriate marking. The point is not that this is the only such implementation—Adobe, Truepic, and others are extending their stacks toward text—but that the open standard now carries authentication and declared history for text in a way no single vendor's pre-C2PA system did, and leaves metering open to any observer.

Figure 2 — Text Provenance Pipeline: Authentication, Custody, and Metering Across the Distribution Path (Sign → Distribute → Verify → Meter)

[artwork to be inserted]

Fig. 2. The same artifact carries authentication and declared history through every stage of distribution; external observation at the metering stage closes the loop.

V. Threat Model and Robustness

Provenance infrastructure must operate under adversarial conditions. We identify four classes of attack against text provenance and describe the system's behavior under each, then draw out the architectural division of labor that the threat model implies.

A. Metadata Stripping

Direct removal of the C2PA manifest—through systems that do not preserve sidecar manifests, or through deliberate excision—severs document-level provenance. The in-band granular layer survives this attack because its markers are embedded in the content stream itself. Stripping the manifest while preserving embedded markers creates a verifiable but unattributed signal: the granular commitment remains computable but cannot be checked against a signed manifest claim. The absence of an expected manifest is itself an observable event, recordable by downstream verifiers and useful as evidence of unauthorized redistribution.

B. Content Transformation

Paraphrasing, translation, and substantive editing destroy both statistical watermarks and exact-hash segment matches. Granular commitments over normalized units degrade gracefully under light editing (typographic corrections, formatting changes, Unicode normalization) but fail under semantic transformation. This is the regime in which forensic detection—stylometry, semantic similarity, training-data membership inference—retains its primary utility. The provenance system does not replace these tools; it shifts the burden from detection to attribution where provenance survives, and back to detection where it does not.

C. Adversarial Embedding and Forgery

A malicious actor could attempt to forge a manifest by signing arbitrary content with a valid certificate or by transplanting steganographic markers from one document into another. The first attack is bounded by the trust hierarchy: forgery requires compromising a trust-anchor certificate or operating with a self-signed certificate that no verifier accepts. The second is bounded by the granular commitment, which is computed over the specific text content; transplanted markers will not validate against the commitment recorded in the manifest unless the surrounding text is also transplanted, in which case the attack reduces to redistribution of the original signed content—the regime that metering is designed to detect.

D. Platform-Level Normalization

Several major platforms (Slack, X, Substack, various CMS sanitizers) strip zero-width Unicode characters during ingestion. This is the most significant unsolved problem for text steganography. Alternative in-band embedding channels remain under exploration, each trading payload capacity against detectability and linguistic quality; none is a complete solution. Platform-level normalization is fundamentally a coordination problem—provenance survives only when downstream systems agree to preserve it. Section A.7 (Section A.8 in C2PA 2.4) mitigates the risk by specifying sidecar manifest references that survive normalization at the cost of requiring an external repository lookup; the long-run answer is platform-side support for provenance preservation, which the EU AI Act Code of Practice begins to specify.

E. Division of Labor: Generation Time and Publication Time

The threat model explains an architectural division that would otherwise read as a refutation of this paper. The first architecture disclosed under the Article 50 Code assigns an embedded watermark to text and C2PA to supported files, and does not state any relationship between the text mark and C2PA text provenance [14]. If C2PA text provenance is the emerging substrate, how should a generation-time text mark connect to publication-time signed provenance?

The two instruments attach at different points in the pipeline and face different adversaries. A model provider marks at generation time. Its output is a token stream leaving for an unknown destination, and the first thing that stream encounters is the normalization regime of Section V-D: CMS sanitizers, messaging clients, and format converters that strip zero-width characters and discard sidecars. Embedded steganography is brittle in exactly that environment, and a detached manifest has nothing durable to bind to when the asset is a fragment of a chat response with no stable boundary. A distribution-level watermark, which survives copy-paste and degrades rather than vanishes, is the rational instrument at that layer, notwithstanding that it carries none of the three functions.

C2PA text provenance attaches at publication time. A newsroom CMS, a wire service, or an enterprise pipeline signs a finished document with an organizational certificate, in a controlled environment, with a stable asset boundary and a manifest repository it can host. That is where authentication and declared history become available and where external observation can meter against them, and the actor is not the model provider.

The Code anticipates both layers, and its own structure sorts them the same way: watermarking is required for free-form text above a length threshold, the signed manifest is conditioned on a format that supports metadata, and richer provenance information is optional [5]. On the current public record the two layers address different boundaries and are best read as complements assigned to different actors rather than competitors for one slot. Whether they interoperate is not something the disclosure answers, and that unanswered question is itself a research surface. This sharpens Position 4. The detection community's contribution at the generation layer is watermark robustness; at the publication layer it is verification and metering against a signed substrate; and the interface between them—where a model-level mark and a publisher-level credential describe the same passage—is the least-characterized surface in the stack, as Section VII-3 argues.

VI. Business Models Enabled by the Convergence

The detection research community has historically engaged with synthetic media as a security problem. Attribution and detection produce forensic evidence. The question of who pays for detection infrastructure, and how that infrastructure sustains itself commercially, has received less attention. The convergence of authentication, custody, and metering within C2PA creates business model possibilities that detection-only approaches do not support.

A. Content Licensing at Scale

Authentication establishes who signed a claim bound to the asset. Chain of custody records the relationships and actions those signers declared. Metering, performed by observation, records where and how often the asset was encountered. Together the three enable content licensing for AI training and retrieval at a granularity and evidentiary standard no existing mechanism provides. A publisher who signs content with C2PA credentials and later observes that content in AI outputs or training corpora holds an authenticated claim bound to the asset, a set of declared relationships and actions, and a record of verification events. That material may support licensing, notice, and evidentiary workflows [11]. It does not by itself establish copyright ownership, authorship, an exhaustive distribution path, copying, infringement, knowledge, or willfulness, each of which remains a fact-specific determination for the forum in which it is raised. Existing access controls (robots.txt, TollBit, AI Crawl Control, ProRata [12]) function only when AI companies voluntarily participate; embedded provenance functions unilaterally.

B. Verification Markets

C2PA credentials are free to verify. Any implementation of the open standard can validate a signed asset without registration, payment, or vendor relationship. This design choice—verification as a public good—creates a market structure analogous to the early web's certificate infrastructure: the verification layer is free, the signing and enforcement layers are commercial. Free verification removes friction from adoption; each verification strengthens the trust chain and increases the value of signing infrastructure. This is a network-effect business model that detection products, which typically charge per query, do not replicate.

The contrast with vendor-operated detection is no longer hypothetical. The Article 50 Code requires signatories to make detection available free of charge to deployers, end users, and other legitimate parties [5], and permits that to take the form of a public specification, locally runnable software, or a cloud API; it also permits access to be restricted for a period where free-form text watermark reliability is lower. Where a signatory chooses to operate the endpoint, the signatory determines which

parties are legitimate and sets the terms of scaled access. In the first disclosed architecture the detection mechanism is described as forthcoming [14]. Cost is not the distinguishing variable; permission is. Any conforming implementation can verify a C2PA asset without asking anyone, which keeps the metering function distributed. Where detection runs through a vendor endpoint, whoever controls the endpoint controls who may meter—a governance question the open-standard path does not raise, and one the research agenda should.

C. Ad-Tech Provenance Signals

Digital advertising already pays for content verification: IAS and DoubleVerify built a combined $1.2 billion business proving that ads were seen by humans in brand-safe environments [13]. Content provenance is the next verification signal, proving that the content adjacent to an ad is authentic, not synthetic, and attributable to a known publisher. CPM-based provenance verification maps directly onto existing ad-tech billing infrastructure and scales with advertising volume.

D. Compliance Infrastructure

The EU AI Act's Article 50 transparency obligations took effect on August 2, 2026. The Code of Practice sets out a multilayered approach: digitally signed metadata where the output format supports it, plus an imperceptible watermark [5]. Demand for infrastructure that delivers both layers is no longer a forecast. A frontier provider signed the Code in both roles the regulation names, as a provider of generative AI models and of generative AI systems, and has disclosed a dual-layer architecture for supported models that it states is not limited to the Union [14]. Infringement of Article 50 exposes providers to administrative fines of up to €15 million or 3% of worldwide annual turnover [16]. Adherence to the Code is a means of demonstrating compliance and expressly not conclusive evidence of it [5], which leaves every signatory needing a durable record of what was marked, by which layer, and when. That audit record is what a signed and metered substrate produces and a bare watermark does not. Compliance-as-a-service—providing enterprises with the signing, watermarking, and audit infrastructure to meet regulatory requirements—is a business model that detection research enables but cannot itself monetize.

E. Multi-Modal Portfolio Signing

The convergence matters most when it operates across media types. A news article contains text, photographs, embedded video, and audio; a multi-modal signing platform binds these into a single composite C2PA manifest, each component independently verifiable, letting publishers sign an entire content portfolio under one provenance infrastructure. The commercial implication: one vendor relationship, one integration, one provenance chain across all media types. The regulatory implication: a single platform satisfying both the provenance-manifest and invisible-watermarking requirements that the Code of Practice specifies. This is a platform business, not a point solution.

VII. Implications for the Detection Community

Three concrete research agendas follow from the analysis above.

1) Robust text steganography. Embedding channels that survive platform-level Unicode normalization, format conversion, and CMS sanitization remain open. Every candidate channel trades payload capacity against detectability and linguistic quality, and characterizing that trade across channels is open work.
2) Granular authentication over streaming output. Verifying content while it is still being produced requires incremental finalization and low-latency verification, and the security property depends on the signing and verifying parties dividing the text identically. Open questions include division policies for non-English text, handling of tool-use and citation outputs, and integration with token-level statistical watermarks.
3) Cross-modal binding and desynchronization. When signed text is rendered to a screenshot, transcribed from audio, or extracted from a signed PDF, the cryptographic binding to the original manifest breaks. The interaction between watermarks applied at generation and credentials applied at publication creates a desynchronized verification surface whose failure modes—including cases in which two valid signals appear to support different conclusions—are not yet well characterized. The architecture that produces this surface has now been announced rather than merely projected: a model-drafted passage would carry a generation-time watermark, the publishing newsroom would sign the finished article with its own C2PA credential, and the two signals would be produced by different actors, at different layers, with different survivability profiles [14]. Neither signal is what it is casually taken for. A watermark detection result is not a cryptographic attestation, and a valid C2PA signature does not make an authorship assertion factually true; both may validate under their respective systems while appearing to answer the same question differently. The failure mode has been formalized for images, where a valid C2PA manifest asserting human authorship coexists with an AI watermark on the same asset [17]; the text analogue is unstudied and, on the announced architecture, near-term.

Forensic detection becomes more valuable, not less, when it operates within a provenance ecosystem that can attribute findings, establish custody, and meter usage. The detection community's next contribution is not only better classifiers—it is the integration of detection capabilities into the provenance substrate.

VIII. Conclusion

The visual bias in synthetic media research was a rational response to the most visible and politically urgent manifestation of generative AI, not a failure of imagination. But text is now the dominant medium of AI-generated content, and the medium in which AI systems most directly intersect with journalism,

law, governance, and commerce. The first architecture disclosed under the Article 50 Code shows what the missing layers cost: a mark that its own vendor documents as evidence of processing rather than authorship [14], arriving in institutions that will read it as a verdict on who wrote the text. Authentication is what turns a signal into an attribution, declared history is what makes the attribution traceable, and accountable observation is what makes it countable. The relation of these three functions to C2PA creates conditions that no prior content-protection system achieved: an open standard in which the first two are interoperable and not controlled by a single vendor, and over which the third can be exercised by anyone able to verify a signature. That layering enables licensing, verification, ad-tech, compliance, and multi-modal portfolio businesses that detection-only approaches cannot sustain—and gives the detection community a substrate on which forensic capability can be attributed, traced, and metered rather than merely classified.

References

[1] A. Loth, M. Kappes, and M.-O. Pahl, "Can Humans Tell? A Dual-Axis Study of Human Perception of LLM-Generated News," in Proc. ACM, 2026.

[2] A. Loth et al., "The Verification Crisis: Expert Perceptions of GenAI Disinformation and the Case for Reproducible Provenance," in Proc. ACM, 2026.

[3] Coalition for Content Provenance and Authenticity, "C2PA Specification 2.3, Section A.7: Embedding Manifests into Unstructured Text," Jan. 2026; renumbered Section A.8 in C2PA Specification 2.4, HTML having been added as A.7. [Online]. Available: https://spec.c2pa.org/specifications/specifications/2.3/specs/C2PA_Specification.html

[4] Defense Advanced Research Projects Agency, "Media Forensics (MediFor) Program," 2016–2020; "Semantic Forensics (SemaFor) Program," 2020–present.

[5] European Commission, "Code of Practice on Transparency of AI-Generated Content under Article 50 of the AI Act," final Code, Jun. 10, 2026. Drafted by independent experts in a process facilitated by the AI Office and subsequently confirmed by the Commission and the AI Board as an adequate voluntary means of demonstrating compliance. [Online]. Available: https://digital-strategy.ec.europa.eu

[6] Washington State Legislature, "HB 1170: Concerning the Disclosure of Synthetic Media in Political Advertising and Other Contexts," signed Mar. 2026, effective Feb. 1, 2027.

[7] Cyberspace Administration of China, "Provisions on the Management of AI-Generated Synthetic Content," effective Sep. 2025.

[8] R. C. Merkle, "A Digital Signature Based on a Conventional Encryption Function," in Advances in Cryptology — CRYPTO '87, Springer, 1988, pp. 369–378.

[9] S. Dathathri et al., "Scalable Watermarking for Identifying Large Language Model Outputs," Nature, vol. 634, pp. 818–823, Oct. 2024 (Google DeepMind SynthID Text).

[10] S. Sadasivan et al., "Can AI-Generated Text be Reliably Detected?," arXiv:2303.11156, 2023; K. Krishna et al., "Paraphrasing Evades Detectors of AI-Generated Text," NeurIPS, 2023; J. Kirchenbauer et al., "On the Reliability of Watermarks for Large Language Models," ICLR, 2024.

[11] 17 U.S.C. § 504(c)(2) (enhanced statutory damages for willful infringement of registered works). Cited for the statutory framework only; nothing in this paper asserts that provenance records establish willfulness, which remains a fact-specific finding.

[12] TollBit Inc., https://tollbit.com; Cloudflare, "AI Crawl Control," https://developers.cloudflare.com/ai-crawl-control/; ProRata.ai, https://prorata.ai.

[13] Integral Ad Science, Annual Report 2025; DoubleVerify Holdings, Annual Report 2025.

[14] Anthropic, "How Claude marks AI-generated content," Claude Help Center, updated Aug. 2026. The page describes planned and in-progress marking for supported models; it does not document a completed end-to-end deployment, and it does not disclose the text watermark mechanism. [Online]. Available: https://support.claude.com/en/articles/16266773

[15] Encypher Corporation, "Enterprise Content Provenance API," 2026. [Online]. The authors are affiliated with this implementation.

[16] Regulation (EU) 2024/1689 (Artificial Intelligence Act), Art. 99(4) (administrative fines for infringement of obligations including those under Art. 50).

[17] Case Western Reserve University and University of California, Los Angeles, "Authenticated Contradictions from Desynchronized Provenance and Watermarking," Apr. 2026.