THE CLASSIFICATION DEFICIT

Article 50 of the EU AI Act, the Epistemic Gap It Cannot Close, and the Formal Standard That Does

Vincent Gonzalez*

Abstract

The evidentiary problem of AI-generated content is not a disclosure problem. It is an epistemological one. Every major governance framework — the EU AI Act's Article 50, proposed FRE Rule 707, the Take It Down Act, platform watermarking policies — addresses synthetic content governance as a labelling problem: if generated content is properly disclosed, tagged, or marked, downstream harms can be managed. This Article argues that framing is structurally mistaken. The relevant distinction between observed and generated content is not a property that can be attached to content after production; it is a classification of epistemic status — specifically, the presence or absence of the constitutive causal connection to observable reality that makes content capable of anchoring knowledge claims. No label, watermark, or disclosure requirement can restore what is structurally absent.

This Article traces the history of evidentiary authentication through three transitions — physical artefact, photograph, digital file — demonstrating that each produced an authentication gap that law addressed procedurally rather than formally. The current transition to AI-generated content represents a structural rupture, not merely a further step in the sequence: synthetic content lacks not a causal chain that can be documented but the causal chain that makes evidence epistemically significant. Five categories of evidentiary failure produced by this rupture are identified and analysed: identity attribution without observational grounding (the Taylor Swift deepfakes), authority attribution without testimonial grounding (the Biden robocall), harm attribution without causal grounding (the AI-CSAM crisis), authentication without classification infrastructure (the

deepfake defense), and authorship attribution without creative grounding (the AI authorship problem).

The Modulign Standard v3.0 — a published open specification for Dimensional Address Grammar for Observable Reality (DAG-OR) — provides the formal infrastructure that closes the classification deficit. Its VR/·:SYN classification primitive is the first formal epistemic primitive for synthetic content, encoding the causal deficit at the classification level rather than as metadata. Its Address-Theoretic Non-Accidentality Principle (ATNA) provides the first structural dissolution of the Gettier problem as applied to observational evidence, guaranteeing that a properly classified observation cannot be accidentally true. Its confidence threshold architecture maps emergently onto legal standards of proof. Its ^EVID append-only ledger constitutes the chain of custody from the moment of classification. A clause-by-clause analysis of Article 50 identifies five structural gaps that DAG-OR closes without displacing the regulation. Proposed FRE Rule 901(b)(11) amendment language is provided. A unified four-step implementation framework concludes the Article.

PART I: THE HISTORICAL ARC OF EVIDENTIARY AUTHENTICATION

The law's relationship to observational evidence has always been mediated by technology. The forms of evidence that courts receive, authenticate, and weigh have changed dramatically — from physical artefacts and eyewitness accounts, through photographic and audio recordings, through digital files and metadata, to the AI-generated content that now poses the most serious challenge in the history of evidentiary practice. With each transition, the evidentiary problem has been structurally transformed. And with each transformation, law has responded through a characteristic pattern: procedural adaptation without formal epistemic infrastructure.

I.A The Evidentiary Problem in Its Original Form

The foundational problem of observational evidence is not technological; it is epistemological. Evidence is offered in legal proceedings to anchor factual findings — to establish that a particular event occurred, that a particular person did a particular thing, that a particular state of affairs obtained at a particular time. For evidence to perform this function, it must not merely assert; it must ground. The distinction between assertion and grounding is the distinction between testimony and evidence, between claim and proof.

What makes observational evidence capable of grounding a factual finding is the constitutive causal connection between the evidence and the state of affairs it represents. A footprint in mud is evidence of a person's presence because the person's foot causally produced it. A medical examination is evidence of injury because the injury causally produced the observable physiological state. In each case, the evidence carries epistemic weight because its causal history includes the fact it is offered to establish.

This is why authentication is the threshold requirement for admissibility across every major legal system. Authentication asks: is this evidence what its proponent claims? The answer is, at bottom, a question about causal history. Is this footprint the kind of artefact causally produced by the presence of a human foot? Is this document the kind of artefact causally produced by the institution it purports to represent? Authentication is the law's procedural mechanism for checking the causal pedigree of evidence.

For most of the common law's history, this check was performed through familiar, well-understood procedures. The procedures were imperfect, but the epistemology was legible: everyone involved understood, at least implicitly, what it meant for a piece of evidence to be causally connected to the fact it proved. This legibility has been progressively eroded by technological change.

I.B The First Transition: Photograph as Evidence

The introduction of photography into legal proceedings in the second half of the nineteenth century posed the first major challenge to common law evidentiary epistemology. Courts developed what became known as the 'silent witness' theory — the doctrine that a photograph can speak for itself, offering direct evidence of what it depicts, provided it is authenticated. The photograph was treated as a kind of mechanical eyewitness: a device that recorded what was present at the moment of exposure with a fidelity human memory could not match.

Authentication required that a witness with personal knowledge of the scene testify that the photograph was an accurate representation of it. This procedural resolution worked tolerably in an era when photographs were expensive and primarily produced by professionals. It became problematic as photography democratised, as darkroom manipulation became technically accessible, and as the gap between the photograph's apparent fidelity and its actual causal pedigree widened. The residue of the photographic transition was a set of ad hoc authentication procedures that depended on the assumption that photographs were difficult to fabricate convincingly. When digital image manipulation made this assumption untenable, the procedures became inadequate — and a new authentication gap opened.

I.C The Second Transition: Digital Evidence and the Metadata Era

The digitalisation of evidentiary media through the 1990s and 2000s produced a second structural transformation. Digital files could be copied without degradation, modified without visible traces, and transmitted across jurisdictions without physical custody transfers. The evidential assumptions built into the physical-evidence framework — that originals were distinguishable from copies, that tampering left visible traces, that custody was constituted by physical possession — failed systematically.

Law's response was procedural and metadata-based: chain-of-custody documentation requirements, hash verification, access logs. The NIST guidelines for digital forensics, the ACPO Good Practice Guide, and ISO/IEC 27037 collectively represent this approach. Hash verification is a genuine advance: it provides tamper detection far more reliable than visual inspection. But the response remained procedural rather than architectural. The integrity of digital evidence depended on the conduct of the collecting officer, the completeness of the documentation, the diligence of the reviewing court — not on structural properties of the evidence that made integrity violations impossible rather than merely prohibited. The gap between procedural compliance and genuine epistemic security remained.

I.D The Third Transition: AI-Generated Content and the Authentication Collapse

The emergence of AI-generated content as a significant evidentiary problem represents not merely a further step in the sequence of technological transitions but a structural rupture. The photograph-to-digital transition changed the medium and vulnerability profile of observational evidence; it did not change the fundamental character of the authentication problem. Both photographic and digital evidence, when genuine, stands in a constitutive causal relationship to the reality it depicts.

AI-generated content changes the authentication problem at the level of its foundation. The causal history of an AI-generated image does not include the scene it depicts. The causal history of a synthetic voice recording does not include the person whose voice it mimics. The content may be perceptually indistinguishable from genuine observational evidence; it may be technically sophisticated; it may be embedded in a plausible narrative. But it lacks the one thing that makes observational evidence epistemically significant: the constitutive causal connection to the reality it depicts.

Law's initial response has followed the same pattern as prior transitions: procedural adaptation through disclosure requirements, authentication guidelines, and evidentiary rules applicable to acknowledged synthetic content. These are genuine regulatory advances. They are not adequate responses to the authentication collapse they address. Disclosure requirements address acknowledged synthetic content; they do not address synthetic content whose origin is disputed — which is precisely the scenario where the authentication problem is most acute. An

adversary who can create synthetic evidence indistinguishable from genuine evidence, and deny its synthetic origin, has defeated every disclosure-based framework regardless of how stringent the disclosure requirements are.

What is needed is formal epistemic infrastructure that makes the synthetic/observed distinction a structural feature of how evidence is classified — established at the point of production, registered in a canonical and tamper-evident ledger, and verifiable by any party regardless of the producer's cooperation. This is the classification deficit: the absence, across the entire arc of evidentiary history, of formal infrastructure adequate to the authentication problem in its current form.

I.E The Pattern and Its Consequence

Each evidentiary transition has produced an authentication gap, and each gap has been addressed by the same characteristic response: new procedural requirements adapted to the specific vulnerability profile of the new evidentiary form. The pattern is not a failure of legal reasoning. The problem is structural: procedural responses are adequate only when the relevant parties cooperate with the authentication procedure.

Procedural responses fail — consistently, predictably — when the relevant party has the technical capacity to produce evidence that satisfies the surface requirements of the authentication procedure without the underlying epistemic substance those requirements are designed to certify. The classification deficit is this gap named precisely. Part II documents its consequences across five categories of evidentiary failure.

PART II: A TYPOLOGY OF EVIDENTIARY FAILURE

The five cases examined in this Part constitute a typology: each instantiates a structurally distinct form of evidentiary failure, located at a different stage in the content lifecycle, and producing a different category of harm. Together they demonstrate that the classification deficit is not a future risk. It is a present reality, manifesting across every context in which observational evidence matters.

Locus in Content Lifecycle	Case	Epistemic Failure Type
Production	Taylor Swift Deepfakes (Jan 2024)	Identity attribution without observational grounding
Transmission	Biden Robocall (Jan 2024)	Authority attribution without testimonial grounding
Aggregation at Scale	AI-CSAM Crisis (2024–2025)	Harm attribution without causal grounding
Adjudication	The Deepfake Defense (2024–2026)	Authentication without classification infrastructure
Economic Attribution	AI Authorship Problem (ongoing)	Authorship attribution without creative grounding

II.A Case 1: Dignity — The Taylor Swift Deepfakes (January 2024)

Epistemic failure type: *Identity attribution without observational grounding.*

In late January 2024, AI-generated sexually explicit images of Taylor Swift proliferated across social media platforms, with a single post reportedly viewed over forty-seven million times before removal. The images were produced using a text-to-image generative AI system with no causal connection to the depicted subject's person, conduct, or volition. Congressional response included the No AI FRAUD Act, the DEFIANCE Act, and ultimately the Take It Down Act (signed May 2025), requiring platforms to remove nonconsensual intimate imagery within forty-eight hours of notice.

The legal response correctly identified the harm. It did not correctly identify its epistemic structure. The harm is not that harmful content exists — it is that AI-generated content falsely

occupies the same epistemic space as observational content, circulating as if it had the causal connection to its depicted subject that genuine photographic evidence has. A photograph of a person performing an act carries epistemic weight because the person causally produced the photograph by performing the act. The AI-generated image carries no such weight — its causal history traces to a generative model's parameters and training data, not to any act by the depicted person.

In the Modulign address space, content depicting a real person in observed circumstances bears the SR/ realm classification. AI-generated content bearing VR/·:SYN cannot be addressed as SR/. The addresses are incompatible: not a rule that can be violated, but a structural property of the grammar. A classification infrastructure requiring point-of-production VR/·:SYN registration would make the identity attribution failure structurally impossible — because the epistemic distinction would be encoded in the content's classification record from the moment of production, verifiable by any party regardless of the producer's cooperation.

II.B Case 2: Democracy — The Biden Robocall (January 2024)

Epistemic failure type: *Authority attribution without testimonial grounding.*

Two days before the 2024 New Hampshire presidential primary, thousands of registered Democratic voters received a robocall in a voice that sounded unmistakably like President Biden's, instructing them not to vote. The audio was subsequently established to have been produced using ElevenLabs, a commercial AI voice synthesis tool, at a cost of approximately one hundred and fifty dollars. The FCC proposed a six-million-dollar fine, and criminal charges followed in New Hampshire.

The epistemic failure is a gap in the framework for authenticating claims of attributed speech. The robocall exploited the epistemic infrastructure that makes recorded speech trustworthy: the listener's assumption that a voice recording is causally produced by the person whose voice it bears. An audio recording of a person making a statement is evidence of that statement because the statement causally produced the recording. The Biden robocall substituted

a generative model's parameters for the causal act of speaking, while preserving the acoustic signature that triggers the listener's testimonial inference.

No classification requirement was imposed at the point of production that would have embedded the audio's synthetic origin in a permanent, verifiable record independent of the producer's disclosure. The VR/·:SYN primitive applied at the point of synthesis would have encoded the audio's causal history from the moment of production. Authentication of the audio as genuine speech — as content causally produced by Biden speaking — would require a SR/ classification, which the audio's causal history cannot support.

II.C Case 3: Child Safety — The AI-CSAM Crisis (2024–2025)

Epistemic failure type: *Harm attribution without causal grounding.*

Operation Cumberland — coordinated by Danish law enforcement with Europol support, concluding in February 2025 with arrests across nineteen countries — represented the first major international law enforcement operation targeting AI-generated CSAM at scale. The National Center for Missing & Exploited Children received 440,419 reports of AI-generated CSAM in the first half of 2025 alone, compared to 6,835 for the entirety of 2024 — a sixty-four-fold year-over-year increase.

The legal framework for CSAM prosecution treats content differently depending on its causal relationship to a real victim. Content depicting a real child is prosecuted under federal CSAM statutes with severe mandatory minimum sentences. Wholly AI-generated content is prosecuted under federal obscenity laws, where the penalty structure differs and the First Amendment analysis is more complex. Making this legally material distinction requires the formal classification infrastructure that does not currently exist. Forensic analysis to determine whether a real child was depicted costs between three thousand and eight thousand dollars per project. At the volume of H1 2025 reports, aggregate forensic costs at even one percent of reports would exceed one billion dollars — a structural consequence of the absence of production-stage classification.

A critical clarification on scope: the VR/·:SYN classification addresses causal grounding, not harm. Whether AI-generated CSAM causes harm, and how severely it should be penalised

relative to content that causally involves a real victim, is a question for legislators and courts. What the architecture provides is formal infrastructure for distinguishing content that stands in a constitutive causal relationship to a real victim from content that does not — the distinction the law requires in order to allocate criminal penalties appropriately. Production-stage classification would convert the forensic question into a registry lookup.

II.D Case 4: Justice — The Deepfake Defense and the Evidentiary Vacuum

Epistemic failure type: *Authentication without classification infrastructure.*

Courts are confronting AI-generated content as an evidentiary problem from two structurally distinct directions that must not be conflated. The first is the offensive use of synthetic content: the introduction of AI-generated evidence as genuine observational evidence. The second is the defensive invocation of synthetic content uncertainty: the argument, by a party seeking to exclude genuine observational evidence, that the evidence could be AI-generated.

In United States v. Reffitt, defense counsel argued that prosecution video evidence of Capitol riot participation could be a deepfake. In Huang v. Tesla, defense counsel argued that video evidence of a party's statements could be AI-generated. These arguments exploit the absence of production-stage classification to create reasonable doubt about genuine evidence. A genuine recording bearing SR/ classification — with an append-only ledger entry and cryptographic chain linking the classification to the recording instrument and event — is authenticated by the classification record itself. The deepfake defense fails not because a court adjudicates and rejects it, but because the classification record structurally forecloses it.

A separate category, which must be distinguished, involves AI-enhanced real observational content. In State v. Puloka, an AI-enhanced version of a bystander's cellphone video was processed using Topaz Video Enhance AI to clarify footage of a shooting; the court denied admission. AI-enhancement cases involve SR/ content with documented processing — a structurally different problem from AI-generation cases, requiring a different classification response. Conflating these categories produces inconsistent evidentiary rules. Proposed FRE Rule 707 — applying only to

acknowledged AI-generated evidence — addresses neither problem adequately in the absence of production-stage classification infrastructure.

II.E Case 5: Attribution — AI Authorship and the Epistemic Status of Creative Work

Epistemic failure type: *Authorship attribution without creative grounding.*

The fifth failure mode differs from the preceding four: the harm is not the misrepresentation of synthetic content as observed content, but the misrepresentation of AI-generated content as human-authored content. Copyright law conditions protection on human authorship; academic and professional standards condition attribution on human authorship; employment contracts include representations about the authorship of delivered work. In each context, the distinction between human-authored and AI-generated content is legally material.

AI detection tools address the problem at the wrong level: they attempt to detect AI origin through output analysis, rather than verify human origin through causal process documentation. A VR/·:SYN classification record at the point of AI generation, and the absence of any such record for human-authored work, would provide formal epistemic infrastructure for authorship attribution. The inversion of the default — from 'presume authentic unless proven synthetic' to 'presume epistemically unverified unless classified' — is the structural change that classification infrastructure makes possible and that no disclosure-based framework can achieve.

II.F The Structural Source: A Single Epistemic Deficit

The five failure modes are diverse in their factual contexts, legal frameworks, and the populations they harm. They share a single structural source: the classification deficit. In each case, harm arises from or is exacerbated by the absence of a formal epistemic distinction between observed and generated content at the classification level — encoded as a constitutive property of the content's identity, rather than as metadata attached after the fact or disclosed by the producer on request.

The five cases constitute a typology of loci at which the classification deficit produces harm: production, transmission, scale aggregation, adjudication, and economic attribution. Together they demonstrate that the classification deficit is pervasive — a structural feature of the current

information environment, manifest wherever the distinction between observed and generated content is legally or economically significant.

PART III: THE MODULIGN ARCHITECTURE AS FORMAL PROPOSAL

This Part introduces the formal infrastructure that closes the classification deficit at its source. The Modulign Standard v3.0 (Gonzalez 2026a) provides a complete architectural response to each failure mode identified in Part II. This Part proceeds in five sections: the three-registry DAG-OR architecture; the VR/·:SYN classification primitive; the Address-Theoretic Non-Accidentality Principle (ATNA); the confidence threshold mapping to legal standards of proof; and the ^EVID chain-of-custody architecture.

III.A The DAG-OR Architecture: Three Registries, One Address Space

The Modulign Standard v3.0 specifies a Dimensional Address Grammar for Observable Reality. The grammar assigns a structured address to every classifiable observation, encoding its domain, subdomain, spatial locus, scale, medium, activity state, observer identity, and jurisdictional context in a single machine-readable code string. The address is not a description of the observation; it is the observation's identity.

The architecture operates through three registries. The Observation Registry assigns a Global Stream Identifier (GSI-ID) — a UUID-v4 permanent unique identifier — to each classified observation. The GSI-ID is immutable: it never changes, regardless of any subsequent reclassification. It is the identifier of the classified observation as an epistemological event: the structured record of what was observed, where, when, by whom, with what confidence, and under what epistemic conditions.

The Entity Registry assigns a persistent Entity Identifier (EID) to entities — people, institutions, structures, legal instruments, AI systems — that have identity across multiple independent observations. An EID is issued only when referenced by at least two independent Modulign classifications. This two-observation requirement is an epistemic requirement: an entity independently observed and classified twice, with consistent classification under the CDP, has been identified under a procedure that is transparent, documented, and reproducible.

The Relation Registry records formally documented relationships between observations and entities — causal (→CAU→), temporal (→SEQ→), correlational (→COR→), contradictory (→CNT→), parallel (→PAR→). Each relation record carries its own timestamp, observer

annotation, and confidence score. Causal claims in the Relation Registry are not assertions; they are graded epistemic findings produced under a reproducible procedure, satisfying the foundational requirements for causal evidence in criminal, civil, and regulatory proceedings.

III.B VR/·:SYN: The First Formal Epistemic Primitive for Synthetic Content

III.B.1 What VR/·:SYN Is Not

A watermark is a signal embedded in content; it is a content property, and it can be stripped. C2PA provides cryptographically signed provenance manifests — the most sophisticated existing tool — but a C2PA manifest records that content was generated by a particular AI system; it does not encode what that means for epistemic weight in proceedings where proof standards apply. Article 50's machine-readable marking requirement and visible disclosure requirements are disclosure labels. A policy prohibition creates legal accountability for non-disclosure; it does not alter the epistemic status of the content.

VR/·:SYN is none of these things. It is a classification: the assignment of the content to its correct position in the DAG-OR address space. The assignment is not contingent on perceptual properties, intended audience, or disclosability. It follows necessarily from the content's causal history.

III.B.2 The Formal Structure and the Ontological Argument

The DAG-OR Realm axis encodes the ontological domain of the observed phenomenon. SR/ designates Surface Realm — the physical observable world. VR/ designates Virtual Realm — computational space and generated content. The Medium axis encodes how the phenomenon was captured. :SYN designates synthetic content — generated, not observed.

The VR/·:SYN combination is a biconditional in the address grammar:

REALM = VR AND MED = SYN if and only if SyntheticContent(ω) = TRUE

The biconditional is grounded in an ontological argument established in the Formal Logic of Modulign (Gonzalez 2026b, Step 5): synthetic content has no physical locus. It is not constituted at any spatio-temporal coordinate (x, y, z, t). It therefore cannot be assigned a physical realm without asserting a physical existence it does not have. VR/ is the only valid realm for synthetic content not because the Standard says so, but because synthetic content by definition has no physical locus.

Content bearing VR/·:SYN cannot function as an ATK-grade observational fact about the reality it depicts. It can function as an ATK-grade observational fact about its own existence as a generated artefact — about the fact that a particular generative process produced a particular output at a particular time. This consequence applies regardless of perceptual realism. This is the precise correction to Article 50(4)'s perceptual deepfake definition.

III.C The Address-Theoretic Non-Accidentality Principle (ATNA)

The deepest formal contribution of the DAG-OR architecture is the Address-Theoretic Non-Accidentality Principle, which addresses the foundational problem of observational evidence generally: ensuring that what a classification says is true, is true for the right reasons — not accidentally, but constitutively.

The Gettier problem (Gettier 1963) demonstrates that justified true belief is insufficient for knowledge. Justification and truth are independent properties: they can coincide accidentally. Zagzebski (1994) establishes that no analysis of the form 'knowledge = true belief + X' can escape Gettier cases under fallibilism, because any external bridge between justification and truth can always be satisfied accidentally. The legal significance is underappreciated: evidentiary standards ask whether a procedure was reliable, whether it was properly conducted, whether the expert was qualified. But they do not ask whether the procedure guarantees a non-accidental connection between the classification and the fact it purports to establish.

ATNA provides this guarantee structurally, as proven in the Formal Logic of Modulign (Gonzalez 2026b, Theorem 4.1):

> Address-Theoretic Non-Accidentality Principle (ATNA): A Modulign classification C of observation O is non-accidentally true if and only if: (a) C was produced by the full Classification Decision Protocol applied to the observable features of O; (b) each step's justificatory basis traces to the observable features causally produced by T(C), the state of affairs making C true; and (c) the Protocol contains no inference step epistemically isolated from O's observable features.

Under ATNA, there is no residual independence of the justificatory basis from the truth-making state of affairs. The Protocol constitutively engages the world in producing the classification. As stated in the Legal Treatise: 'A ^EVID-grade Modulign classification cannot be accidentally true. This is a stronger claim than any existing evidentiary standard makes about the observations it admits.' Existing evidentiary standards guarantee reliability across a population of cases. ATNA guarantees something stronger: that a particular classification of a particular observation is non-accidentally true.

The Address-Theoretic Knowledge (ATK) formulation grounds the CDP's legal analysis. The following table maps each ATK condition to its legal analogue:

ATK Condition	Requirement	Legal Standard	Citation
(iii)(a)	Constitutively world-engaged: justificatory basis traces to observable features causally produced by T(P)	Daubert: methodology must be reliably applied to the facts of the case; expert testimony based on sufficient facts	Daubert v. Merrell Dow Pharmaceuticals, 509 U.S. 579 (1993)
(iii)(b)	Transparent: steps explicit and auditable	FRE 705: expert must disclose basis for opinion; Daubert: methodology must be testable	FRE 705; Kumho Tire, 526 U.S. 137 (1999)
(iii)(c)	Intersubjective: reproducible by any competent observer with access to the same observable features	Daubert: methodology must be peer-reviewed and generally accepted in the relevant scientific community	General Electric Co. v. Joiner, 522 U.S. 136 (1997)

ATK Condition	Requirement	Legal Standard	Citation
(iii)(d)	Exhaustive: covering all segments of the reality-space relevant to P	Completeness requirements for forensic analysis; duty to consider contrary evidence	FRE 702(b)
(iii)(e)	Scale-sensitive: operating at the scale at which T(P) is constituted	Goldman's fake barns: a classification at the wrong scale is false regardless of other correctness	Goldman (1976); Formal Logic Theorem 4.3

III.D The Confidence Threshold Architecture and Legal Standards of Proof

The Modulign confidence scoring architecture produces a mapping to legal standards of proof that is, as the Legal Treatise observes, 'not engineered — it emerges from the epistemological structure of the Modulign confidence scoring system and the legal system's independent calibration of standards of proof. Both are attempts to calibrate certainty thresholds to the severity of the consequential decision being made.'

Every Modulign classification carries a confidence score $\kappa \in [0,1]$ computed as a weighted average of segment-level scores across the CDP's eight steps. Domain and Locus each carry weight 0.25 — the two failure modes most likely to produce a classification false in a way no downstream correction can repair. Subdomain carries 0.15. Observer, Scale, and Medium carry 0.10 each. Time carries 0.05.

Confidence Range	Modulign Status	Legal Standard	Evidentiary Consequence
$\kappa \geq 0.95$	^EVID — High confidence. All segments confirmed.	Beyond reasonable doubt (criminal conviction; ≥90–95%)	^EVID eligible. ATNA non-accidentality guaranteed. Maximum evidentiary weight.
$0.80 \leq \kappa < 0.95$	^EVID — Good confidence. Minor segment ambiguity.	Clear and convincing evidence (civil fraud, parental rights; 75–90%)	^EVID eligible at this tier. Sufficient for heightened civil standards.
$0.60 \leq \kappa < 0.80$	^CONT — Moderate confidence.	Preponderance of the evidence (civil liability; >50%)	Human review required before evidentiary use. Adequate for civil liability.

Confidence Range	Modulign Status	Legal Standard	Evidentiary Consequence
	Contested. Human review queued within 24 hours.		
$\kappa < 0.60$	^PROV — Provisional. Excluded from evidentiary use pending human review.	Probable cause (Fourth Amendment)	Excluded by default. Brady implications if undisclosed by prosecution.

The ^PROV exclusionary rule is self-enforcing: any classification that does not meet epistemic requirements is automatically marked ^PROV and excluded from evidentiary use. A proponent who presents a ^PROV classification in legal proceedings without disclosing its provisional status implicates the expert disclosure obligations of FRCP 26(a)(2) and 26(e) in civil proceedings, and may implicate Brady v. Maryland, 373 U.S. 83 (1963), in criminal proceedings.

III.E The ^EVID Architecture as Chain-of-Custody Instrument

The ^EVID flag is a classification status that requires a specific production procedure and produces a specific formal record. Three components together constitute the ^EVID architecture.

The append-only ledger: ^EVID-grade classifications use an append-only ledger — no classification can be deleted or modified after emission, only superseded. The Formal Logic establishes the constitutive identity: 'ChainOfCustody(ω) ≡ L(GSI(ω)). This is a constitutive identity, not a representation. The ledger does not document the chain of custody. It IS the chain of custody.' Evidence cannot be lost because the architecture does not permit deletion. This satisfies litigation hold obligations (Zubulake v. UBS Warburg, S.D.N.Y. 2003) and FRCP 37(e) — not by imposing obligations but by making violation impossible.

Cryptographic non-repudiation: each ^EVID classification is signed with a threshold signature scheme (k-of-n), covering the classification content, timestamp, and prior signature. The Formal Logic Theorem 3.3 establishes: verification proves that the holder of the signing key produced the entry at the stated time; any tampering is detectable at the specific position of the

tampered entry. The classifier cannot deny having made the classification; the timestamp cannot be retroactively altered.

The GSI-ID as evidential original: the GSI-ID and its attached Observational Ledger are the original. Not the media file; not the sensor reading. The classified observation — the structured record of what was observed, where, when, by whom — is the evidence. The best evidence rule (FRE 1002) is resolved structurally: any modification to the Ledger changes its SHA-256 hash, producing automatic detection. A certified copy of the Ledger is admissible under FRE 1003 as a duplicate.

PART IV: IMPLEMENTATION — WHAT REGULATORS, COURTS, AND LEGISLATURES SHOULD DO

Parts I through III established the problem and the formal proposal. This Part turns to implementation. Section IV.A performs a clause-by-clause gap analysis of Article 50 of the EU AI Act. Section IV.B addresses United States federal evidentiary practice, including proposed FRE amendment language and the Daubert analysis of automated classifier certification. Section IV.C addresses international proceedings. Section IV.D proposes a unified four-step implementation framework.

IV.A Article 50 of the EU AI Act: A Clause-by-Clause Gap Analysis

Article 50 of Regulation (EU) 2024/1689 constitutes the EU's primary legislative response to the epistemic problem of synthetic content, with transparency obligations entering into force in August 2026. Understanding what it achieves — and what it does not — requires examining each operative provision.

IV.A.1 Article 50(1): The Chatbot Disclosure Obligation

Article 50(1) requires that providers of AI systems intended to interact directly with natural persons ensure that those systems inform users they are interacting with an AI system at the latest at the time of first interaction. This is primarily an interaction-transparency rule rather than a content-classification rule. It does not create a formal distinction between the epistemic status of AI-generated statements and human statements.

The DAG-OR architecture addresses a complementary and deeper problem. The %AUT classification tag identifies observer type — automated classifier, human, or hybrid — as a permanent, auditable feature of every classification record. Where Article 50(1) requires a runtime disclosure, DAG-OR encodes observer identity structurally in every classification the system produces. The two requirements are compatible and mutually reinforcing. Gap identified: Article 50(1) creates no evidentiary consequence flowing from AI identity disclosure.

IV.A.2 Article 50(2): The Machine-Readable Marking Obligation

Article 50(2) requires providers of AI systems generating synthetic audio, image, video, or text to ensure outputs are marked in a machine-readable format and detectable as artificially

generated or manipulated. Providers must ensure technical solutions are effective, interoperable, robust, and reliable as far as technically feasible. This is a genuine advance, establishing a legal baseline for machine-readable content classification.

But Article 50(2) leaves three critical questions unanswered. First, it does not define what 'artificially generated' means at the level of epistemic architecture: the relevant distinction is not between content produced by a generative model and content that was not, but between content that stands in a constitutive causal relationship to observable reality and content that does not. Second, it does not specify a canonical registry architecture: markings can be stripped or disputed without a canonical record against which to verify the original classification. The Modulign Observation Registry with append-only ^EVID ledger and cryptographic integrity chain (Formal Logic Theorem 5.1) provides this architecture. Third, it does not specify who may produce valid classifications or what competence requirements they must satisfy. The %AUT Certification Framework (Gonzalez 2026c) addresses this directly with three competence levels, accuracy thresholds, explainability requirements, and continuous monitoring obligations.

IV.A.3 Article 50(3): The Emotion Recognition Obligation

Article 50(3) requires deployers of emotion recognition systems to inform natural persons of the system's operation. The DAG-OR architecture encodes a structural constraint directly relevant to this concern: Section 30 of the Modulign Standard v3.0 permanently excludes from classification first-person conscious experience, subjective mental states, and inferences about intent, emotion, or belief drawn from observed behaviour. This boundary condition cannot be altered by version updates; it is an architectural constraint. Article 50(3)'s concern with emotion recognition is addressed by DAG-OR not through a disclosure requirement but through a permanent scope exclusion.

IV.A.4 Article 50(4): The Deepfake Disclosure Obligation

Article 50(4) requires deployers of AI systems generating deepfakes — defined as AI-generated or manipulated image, audio, or video that 'would falsely appear to a person to be authentic or truthful' — to disclose that the content has been artificially generated. The deepfake definition has two structural weaknesses. First, it is perceptual rather than causal: content that happens to be perceptually distinguishable from authentic content is not a deepfake under this

definition, even though its epistemic status is identical to content that would be mistaken for authentic. Second, the text provision applies only to text 'published with the purpose of informing the public on matters of public interest,' leaving AI-generated legal documents, scientific data, and witness accounts outside the provision.

VR/∴SYN corrects both defects simultaneously: it applies to all content types without distinction, regardless of perceptual realism, and regardless of intended purpose or audience. Gap identified: Article 50(4)'s deepfake definition is perceptually rather than causally grounded, creating under-inclusion for non-realistic synthetic content and systematic gaps for synthetic text outside the public-interest category.

IV.A.5 The Structural Gap: What No Clause of Article 50 Addresses

The deepest gap is not specific to any clause; it runs through the entire Article. Article 50 treats synthetic content governance as a disclosure problem. None of its obligations addresses the epistemic consequence of disclosure — what it means for content to lack the constitutive causal connection to observable reality, and what evidentiary weight it can bear as a result. Article 50 creates legal accountability for non-disclosure. It does not create the formal epistemic infrastructure that courts need to reason about what synthetic content can and cannot prove.

IV.A.6 The Code of Practice: Convergence and Residual Gaps

The European Commission published the first draft Code of Practice on Transparency of AI-Generated Content in December 2025. The Code's distinction between 'fully AI-generated' and 'AI-assisted' content begins to approach the epistemic distinction DAG-OR encodes formally. Its requirement that 'automation alone is not sufficient' for labelling decisions reflects the same concern motivating the %AUT framework's prohibition on black-box classifications. Two residual gaps remain: the Code does not establish a canonical registry, and it remains within Article 50's disclosure paradigm rather than developing a classification standard.

IV.B United States Federal Evidentiary Practice

The Federal Rules of Evidence contain no rule specifically governing AI-generated evidence. Proposed Rule 707, released for public comment in August 2025 with the comment period closing

February 2026, applies only to evidence that the proponent acknowledges was created by AI. As the Advisory Committee's own documentation acknowledges, this scope limitation means the rule does little to help courts avoid deepfakes whose authenticity is disputed — a direct consequence of the classification deficit.

IV.B.1 Proposed FRE Amendment Language

The following amendment language is proposed for consideration by the Advisory Committee on Evidence Rules. It is intended to complement proposed Rule 707 and is drafted to be consistent with the existing structure of Article IX.

> Rule 901(b)(11) — Authenticity of Observed Content. For content offered as observational evidence of events or conditions, evidence that the content bears a valid classification record in a formally certified classification standard that: (A) encodes the content's causal provenance through a constitutive connection to observable reality; (B) was produced by a qualified observer through a documented, reproducible classification procedure; (C) is registered in an append-only, publicly queryable ledger with cryptographic integrity guarantees; and (D) achieves a confidence threshold corresponding to the applicable standard of proof.

> Rule 901(c) — Presumption for Unclassified AI-Generated Content. Content produced by an AI system that has not been classified in accordance with Rule 901(b)(11) shall be presumed to lack the observational grounding necessary to authenticate under Rule 901(a). A proponent may rebut this presumption by a preponderance of the evidence.

These amendments do not prohibit AI-generated content from being offered in evidence. They create a clear procedural framework for what proponents must show to authenticate observed content, and what courts should presume about unclassified AI-generated content. They are technologically neutral — specifying requirements for a formally certified classification standard rather than any particular standard — while being architecturally concrete.

IV.B.2 The Daubert Framework and Automated Classifier Certification

The Daubert trilogy established the framework for admissibility of expert testimony based on scientific methodology. The %AUT Certification Framework satisfies all four Daubert criteria as applied to automated classifiers. The CDP is a documented, reproducible methodology — testable by any observer with access to the same observable features. The Modulign Standard is published as an open specification and has been subjected to academic peer review. The confidence score architecture provides a known error rate per classification. And the Standard's Zenodo publication with permanent DOIs, its submission to peer review, and its formal logic documentation establish the basis for general acceptance in the relevant community. The %AUT framework's mapping to Daubert is not post-hoc rationalisation; it is a design requirement.

IV.C International Proceedings

ICC proceedings involve evidence collected across multiple jurisdictions, often under contested chain-of-custody conditions. Three challenges are particularly acute. First, chain-of-custody documentation may be unavailable in conflict zones. Second, jurisdictional fragmentation means evidence produced using AI in one jurisdiction may be subject to disclosure requirements that do not apply where it is offered. Third, the perceptual quality of AI-generated content has reached a level where courtroom examination cannot reliably distinguish observed from synthetic content.

The DAG-OR architecture addresses all three. The GSI-ID's immutability and the append-only registry mean that a classification record established at the point of production survives chain-of-custody disruptions. The jurisdiction-agnostic architecture means a classification produced under Article 50's EU framework carries the same epistemic weight as one produced under a US FRE framework, because the classification standard is the same. And VR/∴SYN provides a formal vocabulary for the synthetic/observed distinction independent of perceptual examination.

IV.D A Unified Implementation Framework

The gap analyses of Sections IV.A through IV.C converge on a common conclusion: the formal infrastructure required for adequate synthetic content governance exists and is published. What does not yet exist is a coordinated framework for deploying it across the relevant institutional actors.

Step 1 — Regulatory Adoption: EU AI Act Implementation

The European Commission's AI Office should incorporate the DAG-OR architecture into its Code of Practice. Specifically: (a) adopt VR/·:SYN and SR/ as the formal classification vocabulary for synthetic and observed content respectively; (b) require that Article 50(2) machine-readable marks be embedded in DAG-OR classification records rather than freestanding metadata; (c) adopt the %AUT Certification Framework's competence levels as the compliance standard for automated classifiers; and (d) require that classification records be registered in a canonical, append-only ledger. These adoptions do not require legislative amendment to the AI Act. They are implementation decisions within the scope of Article 50(7)'s implementing acts authority.

Step 2 — Evidentiary Adoption: US Federal Rules of Evidence

The Advisory Committee on Evidence Rules should adopt the proposed Rule 901(b)(11) and Rule 901(c) amendments set out in Section IV.B.1. These amendments provide a technologically neutral framework for authenticating observed content and a default presumption for unclassified AI-generated content. They are consistent with existing FRE architecture and do not require legislative action — amendment to the Federal Rules proceeds through the Rules Enabling Act process. The Committee should simultaneously note that Rule 707's limitation to acknowledged synthetic content creates a structural gap that Rule 901 amendments can address. The two rules are complementary.

Step 3 — International Coordination: ICC and Arbitral Tribunals

The ICC Assembly of States Parties should consider a resolution adopting DAG-OR classification records as a recognised authentication mechanism for digital evidence in ICC proceedings. This requires no amendment to the Rome Statute; it is a procedural adoption within the Court's authority to interpret its Rules of Procedure and Evidence. International arbitral institutions — ICC, ICSID, LCIA — should develop guidance on the evidentiary weight of DAG-OR classification records.

Step 4 — Technical Infrastructure: Registry and Certification

None of the institutional adoptions in Steps 1–3 is effective without underlying technical infrastructure. Three components are required: the Modulign Observation Registry must be operational and publicly queryable; a recognised Certification Authority structure must exist to

conduct %AUT certification; and the Standard must be published with adequate documentation to withstand scrutiny in legal proceedings. This documentation exists: the Modulign Standard v3.0, the Formal Logic of Modulign v1.1.0, and the Gettier dissolution paper (under review at Analysis) provide the technical and philosophical foundations required.

The implementation sequence is self-reinforcing. Regulatory adoption of the Code of Practice creates institutional demand for registry infrastructure. Evidentiary adoption of Rule 901 amendments creates litigation incentive for content producers to maintain classification infrastructure. International coordination creates multi-jurisdictional consistency, removing the jurisdictional arbitrage that has made synthetic content governance difficult in cross-border proceedings.

IV.E The Non-Legislative Claim

Throughout this Article, a recurring theme has been that the architecture proposed here satisfies existing legal requirements rather than requiring new legal obligations. FRE 901 already requires authentication. FRE 702 and the Daubert framework already require documented methodology, known error rates, and peer review. EU AI Act Article 50 already requires machine-readable marking and effective, interoperable, robust technical solutions. ICC evidentiary practice already weighs probative value against authentication concerns.

None of these requirements was written with AI-generated content in mind. None specifies what the formal infrastructure adequate to meet their requirements looks like. But all of them contain the structural space into which DAG-OR's architecture fits. The implementation steps proposed in Part IV.D do not require new legislation. They require courts, regulatory agencies, and standards bodies to recognise that the formal infrastructure adequate to satisfy the requirements they already impose has been developed, and to adopt it accordingly.

CONCLUSION

The history of evidentiary authentication is a history of authentication gaps. Each technological transition — from physical artefact to photograph, from photograph to digital file, from digital file to AI-generated content — has produced a new category of evidence whose epistemic status could not be established through the procedures developed for the prior regime. Each gap has been addressed through procedural adaptation rather than formal epistemic infrastructure. And each adaptation has been adequate only when the relevant parties cooperated with the authentication procedure.

The current transition is different in kind. AI-generated content does not merely create a new vulnerability in existing authentication procedures; it renders the foundational assumption of those procedures — that evidence, when genuine, stands in a constitutive causal relationship to the reality it depicts — contingent rather than structural. The classification deficit is the absence of formal infrastructure adequate to make this assumption verifiable: to encode the synthetic/observed distinction at the classification level, as a constitutive property of the content's identity, rather than as metadata attached after the fact.

Five categories of evidentiary failure document the deficit's present consequences: identity attribution without observational grounding, authority attribution without testimonial grounding, harm attribution without causal grounding, authentication without classification infrastructure, and authorship attribution without creative grounding. The structure of each failure is the same: in the absence of a formal epistemic primitive at the classification level, synthetic content occupies the same evidential space as observed content, with consequences that range from individual harm to democratic integrity to child safety to the integrity of the justice system itself.

The Modulign Standard v3.0 provides the formal infrastructure that closes the deficit. Its VR/·:SYN classification primitive encodes the causal deficit of synthetic content at the classification level — not as a label, not as a disclosure, but as a constitutive property of the content's address. Its ATNA principle provides the first structural dissolution of the Gettier problem as applied to observational evidence, guaranteeing that a properly classified observation cannot be accidentally true. Its confidence threshold architecture maps emergently onto legal standards of proof. Its ^EVID append-only ledger with cryptographic integrity constitutes the

chain of custody from the moment of classification. A clause-by-clause analysis of Article 50 identifies five structural gaps that DAG-OR closes without displacing the regulation. Proposed Rule 901 amendments provide the FRE vehicle for evidentiary adoption.

The governance gap can be closed. The standard is ready. The question before regulators, courts, and legislatures is not whether formal epistemic infrastructure for synthetic content is needed — every framework discussed in this Article acknowledges that it is. The question is whether the formal infrastructure that exists will be recognised and deployed before the next wave of evidentiary failures makes the cost of inaction impossible to ignore.

Notes

1. Gonzalez, V. (2026a). Modulign Standard v3.0: A Dimensional Address Grammar for Observable Reality. Open technical specification. Zenodo. https://doi.org/10.5281/zenodo.19348703.

2. Gonzalez, V. (2026b). The Formal Logic of Modulign, v1.1.0. Zenodo. https://doi.org/10.5281/zenodo.19350847. Theorem 4.1 establishes the ATNA proof. Theorem 3.2 establishes ChainOfCustody(ω) ≡ L(GSI(ω)). Theorem 3.3 establishes cryptographic non-repudiation. Theorem 5.1 establishes the cryptographic integrity chain for the ^EVID registry.

3. Gonzalez, V. (2026c). Modulign Standard: Automated Classifier Certification Framework (%AUT). Proposed Appendix to v3.1 Specification. Zenodo.

4. Gonzalez, V. (2026d). 'Structural Dissolution of the Gettier Problem through Address-Theoretic Epistemology.' Under review at Analysis (Oxford University Press).

5. Gonzalez, V. (2026e). Modulign as Evidence: A Legal Epistemology Treatise. v1.0.1. modulign.org.

6. Gonzalez, V. (2026f). Synthetic Content as Epistemic Category: The Governance Gap No Label Can Close — and the Formal Primitive That Does. v1.1.0. modulign.org.

7. Gettier, E. L. (1963). 'Is Justified True Belief Knowledge?' Analysis, 23(6): 121–123. Zagzebski, L. (1994). 'The Inescapability of Gettier Problems.' Philosophical Quarterly, 44(174): 65–73.

8. Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993); General Electric Co. v. Joiner, 522 U.S. 136 (1997); Kumho Tire Co. v. Carmichael, 526 U.S. 137 (1999).

9. Brady v. Maryland, 373 U.S. 83 (1963). Zubulake v. UBS Warburg LLC, 220 F.R.D. 212 (S.D.N.Y. 2003).

10. Goldman, A. I. (1976). 'Discrimination and Perceptual Knowledge.' Journal of Philosophy, 73(20): 771–791. The fake barns case establishes that a reliably formed belief, produced in an environment where the procedure would typically fail, does not constitute knowledge. Formal Logic Theorem 4.3 establishes the analogous result: a classification at the wrong scale is false regardless of correctness at all other levels.

11. Take It Down Act, Pub. L. 119-__ (signed May 19, 2025). No AI FRAUD Act, H.R. 6943 (118th Cong.). DEFIANCE Act, S. 4569 (118th Cong.).

12. FCC Notice of Apparent Liability for Forfeiture, In the Matter of Steve Kramer, May 23, 2024. League of Women Voters of New Hampshire v. Kramer, No. 24-cv-73 (D.N.H. 2025).

13. Europol, 'Operation Cumberland,' press release, February 28, 2025. NCMEC CyberTipline: 6,835 AI-generated CSAM reports (CY2024); 440,419 reports (H1 2025). Forensic cost range from Grimm, P.W., Grossman, M.R., Linna, D.W. et al. (2024). 'Deepfakes in Court.' University of Chicago Legal Forum.

14. United States v. Reffitt, No. 21-cr-32 (D.D.C. 2022). State v. Puloka, King County Superior Court, Washington, 2024. Huang v. Tesla (N.D. Cal., ongoing). United States v. Alvarez, 567 U.S. 709 (2012) (government's power to prohibit false statements of fact substantially limited by First Amendment — the evidentiary claim is distinct from the prohibitory question).

15. Advisory Committee on Evidence Rules, Agenda Book, May 2025 Meeting; comment period for proposed Rule 707 closed February 16, 2026. Federal Rules of Evidence 901, 901(b)(9), 901(b)(11) (proposed), 901(c) (proposed), 1002, 1003. FRCP 26(a)(2), 26(e), 37(e).

16. Regulation (EU) 2024/1689 (EU AI Act), Article 50 (transparency obligations). Articles 9–15 (high-risk AI system requirements). Article 50(7) (implementing acts authority). Entry into force of Article 50 obligations: August 2026.

17. European Commission, AI Office. 'Code of Practice on Transparency of AI-Generated Content,' first draft December 17, 2025. Final code anticipated June 2026.

18. Rome Statute of the International Criminal Court, Article 69(4). ISO/IEC 27037:2012. Electronic Signatures in Global and National Commerce Act (E-SIGN), 15 U.S.C. § 7001 et seq. Uniform Electronic Transactions Act (UETA).

19. The silent witness theory of photographic evidence: State v. Tatum, 360 P.2d 754 (Wash. 1961). NIST SP 800-86, Guide to Integrating Forensic Techniques into Incident Response (2006). ACPO Good Practice Guide for Digital Evidence, 5th ed. (2012). ISO/IEC 27037:2012.

20. U.S. Copyright Office. Copyright and Artificial Intelligence (2024). The Copyright Office's position: works produced without human creative input are not copyrightable; human authorship is a constitutional and statutory requirement.

** Independent Scholar, Modulign Standard Author. modulign.org. VinceGonzalez@me.com.*

The Classification Deficit — Complete Manuscript — Version 1.0.0