SYNTHETIC CONTENT AS EPISTEMIC CATEGORY

The Governance Gap No Label Can Close — and the Formal Primitive That Does

A White Paper on AI-Generated Content, Evidentiary Epistemology, and the Modulign Standard

Vincent Gonzalez · Independent Scholar | Founder, Modulign Standard
VinceGonzalez@me.com · 2026 · v1.1.0

This white paper is a companion document to "Structural Dissolution of the Gettier Problem through Address-Theoretic Epistemology" (Gonzalez, under review at Analysis). That paper establishes the formal epistemological foundation. This paper applies it to the most urgent practical problem in AI governance: the absence of a formal epistemic distinction between observed and generated content.

Executive Summary

Every major AI governance framework — the EU AI Act, the NIST AI Risk Management Framework, proposed US federal legislation, and the emerging body of evidentiary case law — treats synthetic content as a labeling problem. If generated content is properly disclosed, tagged, or watermarked, the reasoning goes, downstream harms can be managed.

This paper argues that framing is wrong, and that its wrongness is not a matter of degree but of kind. Synthetic content is not observed content with a label attached. It is a categorically different epistemic object — one that lacks the constitutive causal connection to observable reality that makes observed content capable of anchoring knowledge claims. No label, watermark, or disclosure requirement can restore what is structurally absent.

The Modulign Standard's `VR/·:SYN` classification signature is the first formal grammar to encode this distinction as a **primitive** — not a tag applied after classification, but a constitutive element of the address that determines what epistemic weight the classified content can bear. Content bearing `VR/·:SYN` structurally fails the authentication predicate of FRE 901(b)(9) — not because a rule prohibits its admission, but because its address structure encodes the absence of the causal chain that observational evidence requires.

FIX #3 — Core Terminology Definitions (blocking fix)

Four Terms Used Throughout This Paper

OBSERVATIONAL GROUNDING: The constitutive causal relationship between a piece of content and the state of affairs it represents. Content is observationally grounded when the state of affairs causally produced the content through a documentable physical process.

CAUSAL CHAIN: The traceable sequence of physical causation from a real-world event or state through a recording or transmission process to the resulting content artifact. Synthetic content has a causal chain that terminates in a generative model's parameters, not in a physical event.

EPISTEMIC WEIGHT: The degree to which a piece of content can anchor knowledge claims about the reality it depicts. Observationally grounded content carries epistemic weight proportional to the fidelity and integrity of its causal chain. Synthetic content carries zero epistemic weight about the reality it depicts, regardless of its perceptual realism.

CONSTITUTIVE CAUSAL CONNECTION: A causal relationship that is not merely correlative but definitional — the connection between a crime scene photograph and the crime scene is constitutive of what makes the photograph evidence. Without this connection, the content is not degraded evidence; it is a categorically different kind of object.

This paper proceeds in four parts:

Part I: The Governance Gap — what existing frameworks treat as a labeling problem and why that framing fails

Part II: Five Cases — a typology of synthetic content harms across five loci of epistemic failure, each illustrating a distinct stage at which governance frameworks break down

Part III: The Formal Primitive — what `VR/·:SYN` is, why it is categorically different from a content label, and what it formally entails

Part IV: The Prospective Standard — what governance bodies, courts, and legislatures should adopt, and why the Modulign architecture provides the necessary foundation

Part I: The Governance Gap

1.1 The Labeling Consensus and Its Limits

The dominant approach to AI-generated content governance rests on a disclosure model: generated content must be labeled, watermarked, or disclosed as such, and recipients can then adjust their epistemic posture accordingly. This model underlies **EU AI Act Article 50**'s requirements for transparency in AI-generated outputs — requirements that mandate disclosure of synthetic origin but do not create a formal epistemic classification primitive distinguishing synthetic from observed content. The same model underlies the FTC's guidance on AI disclosure, proposed US federal legislation including the No AI FRAUD Act and the DEFIANCE Act, and platform policies from X, Meta, YouTube, and others.

FIX #6 — EU AI Act Article 50 — explicit citation and distinction

It is important to be precise about what EU AI Act Article 50 achieves and what it does not. Article 50 imposes a **disclosure requirement**: systems that generate synthetic audio, visual, or audiovisual content must ensure that output is marked in a machine-readable format and is detectable as artificially generated or manipulated. This is a genuine regulatory advance. It creates legal accountability for non-disclosure. It shifts liability to producers.

What Article 50 does not create is a **formal epistemic classification primitive**. A disclosure requirement tells downstream recipients that content is synthetic. It does not encode *what that means* for the epistemic weight the content can bear. It does not create a grammar that makes the synthetic/observed distinction machine-readable at the classification level — as a structural feature of the content's address — rather than as a separately applied metadata tag. The `VR/·:SYN` primitive operates at the classification level. Article 50 operates at the disclosure level. Both are necessary. Neither substitutes for the other.

The disclosure model is not without value. Labeling requirements reduce the information asymmetry between content producers and content consumers. Watermarks create audit trails. Disclosure requirements shift legal liability. These are genuine goods.

But the model contains a structural error that no amount of labeling can correct. The error is this: it assumes that the difference between observed and generated content is a *property of the content* — something that can be attached to or stripped from a given piece of media. Under this assumption, the right intervention is to ensure the property is always attached. Label the deepfake. Watermark the synthetic audio. Disclose the AI-generated text.

The structural error emerges when we ask: what is the epistemic status of labeled synthetic content? Does a disclosure requirement transform a deepfake into something that can anchor a knowledge claim? Does a watermark restore the causal connection to observable reality that the content lacks? The answer is no — and understanding why requires engaging the epistemology of observational evidence.

1.2 What Observational Evidence Is

Observational evidence is not merely content that happens to depict something real. It is content that stands in a specific causal-constitutive relationship to the state of affairs it represents. A photograph of a crime scene is observational evidence not because it looks like the scene, but because the scene causally produced it — the photons from the scene struck the sensor, the sensor recorded their pattern, the recording preserves a causal trace of the scene's physical properties at the moment of observation.

This causal chain is what gives observational evidence its epistemic weight. Courts have long recognized this, which is why authentication requirements under FRE 901 ask not merely whether evidence looks real, but whether it was produced by a process that accurately captures reality. The process matters because the process is what connects the evidence to the fact.

Synthetic content breaks this chain constitutively. A deepfake of a crime scene is not a photograph with incorrect content. It is a categorically different kind of object: one whose causal history traces to a generative model's parameters, not to the physical properties of any real scene. No amount of labeling changes this. A disclosed deepfake is still a deepfake. A watermarked synthetic voice recording is still a synthetic voice recording. The label describes the content's provenance; it does not alter the content's epistemic status.

1.3 The Gap in Current Frameworks

FIX #11 — 'First Formal Grammar' Claim — verified against C2PA, W3C, ISO/IEC 42001

Before establishing what the `VR/·:SYN` primitive provides, it is necessary to verify the claim that it is the **first formal epistemic primitive** for synthetic content — as distinct from labeling frameworks. The three principal existing standards require examination:

Standard	What it provides	What it does not provide
C2PA (Coalition for Content Provenance and Authenticity)	Cryptographically signed provenance manifests attached to content files. Records who produced content, what tools were used, and what modifications were made.	A formal epistemic classification primitive. C2PA encodes provenance as metadata; it does not encode the epistemic consequence of that provenance. A C2PA manifest can record that content was generated by AI; it cannot encode that this means the content lacks observational grounding and therefore cannot anchor knowledge claims about the reality it depicts.

W3C Verifiable Credentials / PROV Ontology	Vocabulary for expressing provenance graphs and credential chains for digital content.	An epistemic primitive. W3C PROV records causal chains as graph structures; it does not make structural claims about what epistemic weight content can bear based on those chains. It is a representation framework, not an epistemological architecture.
ISO/IEC 42001 (AI Management Systems)	Governance and risk management framework for AI systems, including requirements for transparency and documentation of AI-generated outputs.	Formal epistemic classification. ISO/IEC 42001 is an organizational governance standard; it does not define a formal grammar for classifying the epistemic status of individual content artifacts.

The `VR/·:SYN` primitive is therefore, to the authors' knowledge, the first formal grammar that encodes the synthetic/observed distinction as a **structural epistemic primitive** — one that makes structural claims about epistemic weight derivable from the classification itself, not from separately applied policy or metadata. C2PA, W3C PROV, and ISO/IEC 42001 are labeling and provenance frameworks. They record facts about content provenance. `VR/·:SYN` encodes the epistemic consequence of those facts. Both layers are necessary; only the Modulign architecture provides the second.

The governance gap is the space between what disclosure-based frameworks can achieve and what the epistemic difference between observed and generated content actually requires. Disclosure frameworks can tell you that content is synthetic. They cannot tell you what that means for the epistemic weight the content can bear — because they have no formal account of what epistemic weight is or how it is grounded. They treat the disclosure as sufficient; they have no framework for reasoning about sufficiency.

What is needed is not a better label. What is needed is a formal primitive that encodes the epistemic difference as a structural feature of how content is classified — one that propagates through any downstream use of that content and makes its epistemic limitations legible at the architectural level, not the policy level.

Part II: Five Cases

FIX #10 — Lifecycle Locus Paragraph — Part II Introduction

The five cases below are not chosen for their notoriety. They are chosen because together they constitute a **typology of loci of epistemic failure** — each case represents a *different stage in the content lifecycle* at which the absence of a formal epistemic classification primitive causes harm:

LOCUS 1 — PRODUCTION (Case 1: Dignity)

Synthetic content is produced without formal epistemic classification.
The harm is identity attribution — content is produced as if it
were observationally grounded when it structurally cannot be.
LOCUS 2 — TRANSMISSION (Case 2: Democracy)
Synthetic content is transmitted through channels that presuppose
observational grounding (robocalls conveying political instruction).
The harm is authority attribution without testimonial grounding.
LOCUS 3 — CONSUMPTION (Case 3: Child Safety)
Synthetic content is consumed and adjudicated under frameworks
designed for observationally grounded content. The harm is
causal attribution without physical victim grounding.
LOCUS 4 — ADJUDICATION (Case 4: Justice)
Courts must authenticate content without a formal classification
standard. The harm is authentication without classification.
LOCUS 5 — ATTRIBUTION (Case 5: Labor and Authorship)
Content is attributed to human creative acts when its causal
chain terminates in a generative model. The harm is authorship
attribution without creative grounding.
VR/·:SYN is the ONLY intervention that operates at LOCUS 1 —
the production stage — the only stage where causal history is
knowable with certainty by the party producing the content.

All other interventions are downstream and therefore partial.

2.1 Case 1: Dignity — The Taylor Swift Deepfakes (January 2024)

In late January 2024, AI-generated sexually explicit images of Taylor Swift proliferated on social media platforms including 4chan and X. One post was reportedly viewed over 47 million times before removal. The images were created using a text-to-image AI system and depicted Swift in explicitly nonconsensual sexual scenarios.

Congressional response included the No AI FRAUD Act and the DEFIANCE Act, both of which aim to provide civil remedies for victims. The **Take It Down Act** (Pub. L. 119-__, signed May 19, 2025) subsequently created a federal requirement for platforms to remove nonconsensual intimate imagery, including AI-generated content, within 48 hours of notice.

The epistemic failure is one of *identity attribution*. The images purported to depict a real person engaging in real acts. They did not. In the absence of a formal epistemic category that encodes this distinction structurally — not as a label but as a classification primitive — the content circulated as if it occupied the same epistemic space as observed imagery. Content bearing `VR/·:SYN` in the Modulign address space cannot be addressed as `SR/` (observed, physical space) content. The addresses are incompatible. The incompatibility is not a rule that can be violated — it is a property of the grammar.

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

2.2 Case 2: Democracy — The Biden Robocall (January 2024)

Two days before New Hampshire's 2024 presidential primary, thousands of voters received a robocall featuring what appeared to be President Biden's voice telling them not to vote in the primary. The fake audio was created using ElevenLabs, an AI voice creation tool. Steve Kramer, a Democratic political consultant, later acknowledged commissioning the robocall. A New Orleans magician named Paul Carpenter was paid $150 to create the audio using AI software.

The FCC proposed a $6 million fine for apparent violations and ruled in February 2024 that AI-generated voices in robocalls are illegal. Criminal charges followed in New Hampshire.

The epistemic failure is one of *authority attribution*. The robocall claimed to convey the expressed intention of a specific public official. It did not. The official's voice was used as a generative input — a pattern to be replicated — not as a channel through which the official communicated. The distinction between 'this audio was produced by Biden speaking' and 'this audio was produced by a model trained on Biden's voice' is precisely the distinction the Modulign `VR/·:SYN` classification encodes.

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

2.3 Case 3: Child Safety — Operation Cumberland and the AI-CSAM Crisis (2024–2025)

FIX #7 — Operation Cumberland date correction

Operation Cumberland (corrected timeline): Suspects were initially arrested in November 2024 as part of the operation's active enforcement phase. In February 2025, law enforcement agencies from 19 countries announced the conclusion of the operation, with 25 arrests total, coordinated by Danish law enforcement and supported by Europol. The operation seized 173 electronic devices and identified 273 suspected members of the network.

FIX #12 — Cost quantification paragraph

AI-CSAM Scale and Cost Data
NCMEC reports of AI-generated CSAM, H1 2025: 440,419
NCMEC reports of AI-generated CSAM, full year 2024: 6,835
Year-over-year increase: ~64x
Forensic analysis cost per project (industry range): $3,000–$8,000
At 440,419 reports × $3,000 minimum: $1.32 billion in forensic costs
for H1 2025 alone, if each report required independent analysis.
Modulign registry query cost: trivial by comparison.
A formal classification at point of production converts a $3,000–$8,000
forensic question into a registry lookup.
The economic case for production-stage classification is not marginal.
It is categorical.

The central legal challenge is precisely the epistemic one: if an image depicts a real child, modified with AI or not, it is prosecuted under federal CSAM laws. If an image is wholly generated using AI without depicting a real child, it is prosecuted under federal Obscenity laws. Depending on which statute an offender is charged under, the resulting penalties can differ significantly.

This sentencing disparity reflects the absence of a formal framework for reasoning about the epistemic status of synthetic content. The legal system distinguishes between content that causally involves a real victim and content that does not — a distinction that maps directly onto

the Modulign `SR/` versus `VR/ ·:SYN` distinction. But because no formal grammar encodes this distinction at the classification level, courts must reconstruct it ad hoc in each case, with inconsistent results.

FIX #8 — CSAM harm paragraph — classification vs. harm (Equity Sage)

A clarification essential to the integrity of this analysis: the `VR/ ·:SYN` classification addresses **causal grounding**, not **harm**. Whether AI-generated CSAM causes harm, and how severely it should be penalized relative to content that causally involves a real victim, is an empirical and normative question for legislators and courts — not a question the Modulign architecture resolves. The architecture determines whether content has observational grounding. Harm is constituted by the reality the content depicts or causally involves, and the degree to which those depicted or involved suffer real consequences. These are separate inquiries. Conflating them would be a category error that serves neither the victims of real abuse nor the integrity of the classification system.

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

2.4 Case 4: Justice — The Deepfake Defense and the Evidentiary Vacuum

FIX #4 — Separate AI-enhancement cases (Puloka, Rittenhouse) from AI-generation (Reffitt)

Courts are now confronting synthetic content from two *structurally distinct* directions, which must not be conflated:

CATEGORY A: AI-GENERATION CASES
Content alleged to be wholly synthetic — no real-world observation
in its causal history.
United States v. Reffitt: defense counsel argued that prosecution
evidence of Capitol riot participation could be a deepfake.
This is the deployment of VR/·:SYN uncertainty as a defense against
content that may be SR/ (observationally grounded).
CATEGORY B: AI-ENHANCEMENT CASES
Real observational content (SR/) that has been processed, upscaled,
or modified by AI after capture — preserving partial observational
grounding but introducing computational modification.

State v. Puloka (Wash. Super. Ct. 2024): an AI-enhanced version of
a bystander's cellphone video was processed using Topaz Video Enhance AI
to clarify footage of a nightclub shooting. The court analyzed the
evidence under Frye, FRE 702, FRE 401, and FRE 403, and denied admission.
Sz Huang v. Tesla: defense counsel argued that video evidence of
a party's statements could be a deepfake. Alleged AI-generation,
not AI-enhancement.
WHY THE DISTINCTION MATTERS:
AI-generation cases involve VR/·:SYN content with no SR/ component.
AI-enhancement cases involve SR/ content with documented processing.
Modulign handles these differently:
— `AI-generated content: REALM = VR, MED = SYN → ^PROV`
— `AI-enhanced content: REALM = SR, MED = VID, META ∋ CONT`
`(CONT = content modification flag), processing documented in ledger.`
Conflating these categories produces inconsistent evidentiary rules.

The Advisory Committee on Evidence Rules released proposed Rule 707 for public comment in August 2025. Rule 707 applies only to evidence that the proponent acknowledges was created by AI — doing little to help courts avoid deepfakes or other falsified evidence whose authenticity is in dispute. This limitation reflects the absence of a formal classification infrastructure. A formal standard requiring classification at point of production would transform this problem: unclassified content would be epistemically unverified by default, not epistemically presumed authentic.

> **Epistemic failure type:** *Authentication without classification.*

2.5 Case 5: Labor and Authorship — The AI Authorship Problem

The 2023 Hollywood writers' strike foregrounded a question that governance frameworks have not formally answered: what is the epistemic relationship between a human author and a text generated by a system trained on human-authored texts?

The governance frameworks emerging from the strike and its aftermath — SAG-AFTRA AI agreements, WGA contract provisions, proposed legislation — all treat this as a property rights question. Who owns the output? How must AI assistance be disclosed? These are important questions. But they presuppose a framework for determining what kind of thing AI-generated content is — and that framework does not yet exist at the formal level.

The `VR/·:SYN` classification encodes the relevant distinction: content whose causal chain traces to a generative model's parameters rather than to a human creative act occupies a different epistemic space, one that should be structurally distinguished in any governance framework that aims to reason about authorship, attribution, or creative rights with precision.

FIX #13 — Second-order effect — default 'unverified' presumption and distributional fairness

A governance framework that treats unclassified content as epistemically unverified by default creates a **second-order distributional effect** that must be acknowledged. Large content producers — studios, platforms, media organizations — have the infrastructure to produce and maintain formal classification records. Individual creators, small publishers, and marginal voices may not. A default 'unverified' presumption that is not accompanied by accessible, low-cost classification infrastructure will systematically disadvantage individuals in favor of institutional producers.

The Modulign architecture is designed to be low-cost at the classification level — registry queries are computationally trivial and the specification is an open standard with no licensing fee. But the governance framework built on this architecture must explicitly distinguish **contested evidentiary contexts** (legal proceedings, regulatory enforcement) from **ordinary social contexts** (social media, personal communication). The 'epistemically unverified by default' presumption should apply in evidentiary contexts where proof standards are high and consequences of error are severe. It should not be mechanically extended to social contexts where it would function as a censorship mechanism against unclassified but genuine content.

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

Part III: The Formal Primitive

3.1 What VR/·:SYN Is

In the Modulign Standard's address grammar, every observation is classified along two axes relevant to the synthetic content question:

The Realm axis encodes the spatial/ontological status of the observed phenomenon:

`SR/` — Surface Realm: the physical observable world

`AT/` — Atmospheric/aerial space

`AQ/` — Aquatic/subaqueous space

`OR/` — Orbital/exoatmospheric space

`VR/` — Virtual Realm: knowledge space, computational space, and generated content

The Medium axis encodes how the phenomenon was captured or transmitted:

`:VID` — video `:AUD` — audio `:IMG` — still image `:TXT` — text `:SYN` — synthetic content (generated, not observed)

The `VR/·:SYN` combination is not a tag. It is a **biconditional** in the address grammar:

REALM = VR AND MED = SYN
if and only if
SyntheticContent(ω) = TRUE
This means two things simultaneously:
— Content that is synthetic MUST receive VR/·:SYN (not optional)
— Content bearing VR/·:SYN IS synthetic (address is the claim)
The biconditional is a structural feature of the grammar, not a policy
requirement. An address bearing :SYN in a non-VR/ realm is invalid —
parseable but semantically incoherent, like a street address specifying
a building on a road that does not exist in the specified city.
Why VR/ is the only valid realm for synthetic content:
Synthetic content has no physical locus. It is not constituted at any
spatio-temporal coordinate (x, y, z, t). It cannot be assigned a
physical realm (SR/, AT/, AQ/, OR/) without asserting a physical
existence it does not have — a contradiction with Modulign's
foundational definition of observable phenomena.

3.2 ATK-Grade vs. ^PROV: The Certification Distinction

FIX #9 — ATK-grade contradiction resolved — certification locus

KEY DISTINCTION: WHO CLASSIFIES DETERMINES EPISTEMIC GRADE

VR/·:SYN can be ATK-grade (^EVID-eligible) OR ^PROV depending on WHO produces the classification and WHEN.

ATK-GRADE VR/·:SYN classification requires ONE OF:

```
(A) A certified classifier applying the full 8-step CDP to the
    synthetic content artifact at point of production, OR
(B) A post-generation classification step performed by an observer
    satisfying eligible_EVID(O, DOM) who applies the full CDP.
```

^PROV VR/·:SYN classification results from:

```
— Self-report by the generating system without separate certification
— Classification that does not complete all 8 CDP steps
— Classification by an observer not satisfying eligible_EVID(O, DOM)
```

PRACTICAL CONSEQUENCE:

An AI system that tags its own output as 'synthetic' produces a ^PROV classification — useful, better than nothing, but not ^EVID.

A certified classifier that applies the CDP to that same output produces an ATK-grade ^EVID classification.

This distinction matters for legal proceedings. ^PROV is excluded from evidentiary use per Invariant I3 of the Modulign Standard.

^EVID satisfies FRE 901(b)(9), chain of custody, and Daubert.

The governance implication: requiring AI systems to self-classify
their output creates a ^PROV record. Requiring certified classifiers
to verify that output creates an ^EVID record. Both are valuable.
Both serve different purposes. Policy must specify which it requires
and in what context.

3.3 Why This Is Categorically Different From a Label

A label is applied to content after classification. It is separable from the content — it can be removed, altered, or ignored. Its relationship to the content is contingent: the label says the content is synthetic, but the content itself does not structurally encode this fact.

The `VR/·:SYN` primitive is not separable from the content's address. The Modulign address *is* the classification — it is not a description of the classification. An observation classified as `VR/·:SYN` has that classification as a permanent, append-only entry in its Observational Ledger. The classification cannot be removed; a subsequent classification that reclassifies it as `SR/` would create a contradiction record in the Relation Registry, not replace the original classification.

3.4 What VR/·:SYN Formally Entails

The epistemic consequences of the `VR/·:SYN` classification follow from the Address-Theoretic Knowledge formulation (ATK). Under ATK, knowledge of observable reality requires that the justificatory procedure constitutively engage the observable features causally produced by the truth-making state of affairs (ATK condition iii(a)).

Synthetic content, by definition, was not produced by a causal process that engaged the observable features of any real-world state of affairs. A deepfake of a crime scene was not produced by the crime scene. A synthetic voice recording was not produced by the person whose voice it mimics.

FORMAL CONSEQUENCE:
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 synthetic artifact — its address, its provenance,

its generative parameters, its timestamp.

But it CANNOT ground knowledge claims about the state of affairs

it represents, because no classification procedure that constitutively

engaged that state of affairs produced it.

This is not a prohibition. It is a structural consequence of the

epistemic architecture. It holds regardless of how convincing the

synthetic content is, regardless of whether it is labeled, and

regardless of whether any governance body has adopted the standard.

3.5 The Detection Problem and Why the Primitive Solves It

A persistent objection to synthetic content governance is the detection problem: if we cannot reliably detect whether content is synthetic, how can we require its classification?

This objection misapplies the problem. Detection is a question of verification: given content of unknown provenance, can we determine whether it is synthetic? Verification is hard and getting harder as generative models improve. Classification is a different question: given content we produced, can we classify it correctly? For content producers — AI companies, individuals using AI tools — classification is trivial. They know whether the content is synthetic because they produced it.

FIX #14 — Third-order risk — defense attorneys and causal chain argument

ANTICIPATED DEFENSE ARGUMENT (Operator 7 — Impact Measurer)

Defense attorneys will argue: 'All digital processing breaks the causal

chain. If digital modification introduces causal discontinuity, then all

digital evidence is epistemically suspect, not just AI-generated content.'

This argument has surface plausibility. The answer is in the Modulign

architecture:

Computational processing of a physical signal does NOT break the causal chain if three conditions are met:

```
(1) DOCUMENTED: The processing steps are fully recorded in the
    Observational Ledger.
(2) REPRODUCIBLE: The processing is deterministic and can be
    independently replicated from the documented inputs.
(3) ORIGINAL PRESERVED: The original signal (pre-processing) is
    retained and content-addressable via its GSI-ID.
```

Under these conditions, the causal chain from physical event to processed artifact is not broken — it is extended and documented. The processing becomes part of the chain of custody (L(GSI(ω))), not a rupture in it.

Synthetic content fails condition (3) constitutively: there is no original physical signal to preserve because the content was not produced by a physical signal. This is the structural difference that distinguishes AI-generated content from AI-enhanced content.

The defense argument, if accepted without this distinction, would make all digital evidence epistemically suspect — an outcome that would be catastrophically disruptive to modern evidentiary practice. The Modulign architecture is the answer to this argument, not its casualty.

The VR/ ·:SYN primitive addresses the classification side of the problem, not the detection side. It provides the formal grammar for classification at the point of production. Content that lacks a verifiable VR/ ·:SYN or SR/ classification is unclassified, and unclassified content should be treated as epistemically unverified in evidentiary contexts.

Part IV: The Prospective Standard

4.1 What Governance Bodies Should Adopt

The governance gap cannot be closed by disclosure requirements alone. What is needed, across all three audiences this paper addresses, is a formal epistemic classification standard for content provenance. The four recommendations below are jurisdiction-agnostic.

Recommendation 1 — Formal classification at point of production

AI systems that generate content should be required to produce, alongside that content, a formal classification record that encodes: (a) that the content is synthetic, (b) the generative parameters or system that produced it, (c) a timestamp, and (d) a cryptographic hash of the content at time of generation. This is not a watermark. It is a classification record in a formally defined grammar — one that is permanent, auditable, and machine-readable.

Recommendation 2 — Epistemic status as a structural default

FIX #5 — Rec. 2 scoped — evidentiary not prohibitory; US v. Alvarez flagged

Content bearing a formal VR/ ·:SYN classification structurally fails the authentication predicate of FRE 901(b)(9) — not because a rule prohibits it, but because its address encodes the absence of the causal chain that observational evidence requires. Courts should be empowered to treat unclassified audiovisual content as epistemically unverified, shifting the burden of proof to the proponent to demonstrate observational grounding.

This is an **evidentiary claim**, not a **prohibitory claim**. The inadmissibility of VR/ ·:SYN content as observational evidence in legal proceedings does not mean such content is legally prohibited from existing, from circulating, or from being used for purposes that do not require observational grounding (art, entertainment, satire, research, education). As the Supreme Court held in *United States v. Alvarez*, 567 U.S. 709 (2012), the government's power to prohibit false statements of fact is substantially limited by the First Amendment. The Modulign architecture does not address the prohibitory question — that is outside its scope. It addresses the **evidentiary** question: what epistemic weight can a given piece of content bear in proceedings where proof standards apply? The answer for VR/ ·:SYN content is: none, as observational evidence about the reality it depicts.

Recommendation 3 — Canonical registry for classification records

Formal classification records should be registered in a canonical, append-only, publicly queryable registry. The purpose is not surveillance but epistemic infrastructure: any party — court, journalist, researcher, platform — should be able to query the registry for a given content's classification record and determine its epistemic status. The Modulign Observation Registry provides this infrastructure.

Recommendation 4 — Jurisdictional scope acknowledgment

Each jurisdiction should map the formal classification standard to its existing evidentiary requirements. The structural claim — synthetic content lacks the causal grounding that observational evidence requires — is jurisdiction-agnostic. The procedural implementation — how that structural fact is encoded in evidentiary rules, how burden of proof is allocated, what remedies are available — is jurisdiction-specific. The standard provides the former; jurisdictions supply the latter.

4.2 The Advisory Committee and Proposed Rule 707

Proposed Rule 707, released for public comment in August 2025, applies only to evidence that the proponent acknowledges was created by AI — doing little to help courts avoid deepfakes or other falsified evidence whose authenticity is in dispute. The Rule 707 public comment period closed February 16, 2026.

This limitation reflects the absence of a formal classification infrastructure. Rule 707 can only address acknowledged synthetic content because there is no standard against which unacknowledged content can be measured. A formal classification standard — one that requires classification at the point of production and makes that classification auditable — would transform this problem: unclassified content would be epistemically unverified by default.

The Advisory Committee is working toward the right answer. The Modulign architecture provides the formal foundation that would make that answer structurally sound rather than procedurally provisional.

4.3 Why This Is Not a Technology Problem

The final resistance to formal epistemic classification typically takes this form: AI generation technology is advancing faster than any classification standard can keep pace with. What is classifiable today may be undetectable tomorrow.

This objection misunderstands what formal classification achieves. Detection technology may or may not keep pace with generation technology. The formal epistemic distinction between observed and generated content is not a technological claim — it is a structural one. A photograph causally produced by a real scene is epistemically different from a synthetic image that looks identical to it, regardless of whether any detection system can tell them apart. The `VR/·:SYN` primitive encodes that difference. It does not solve the detection problem. It makes clear why the detection problem matters.

Conclusion

The five cases examined in this paper share a common structure: in each, synthetic content was treated as if it occupied the same epistemic 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 governance frameworks emerging in response — disclosure requirements, watermarking mandates, proposed evidentiary rules, criminal penalties — are all necessary. None of them is sufficient, because none of them addresses the structural epistemic difference between observed and generated content. They treat it as a labeling problem. It is an epistemological problem.

The Modulign Standard's `VR/·:SYN` primitive is the first formal grammar to encode this difference as a classification primitive rather than a content tag. Its consequences follow from the architecture, not from policy: content that lacks the causal connection to observable reality cannot anchor knowledge claims about that reality, and its address structure encodes this absence permanently, auditably, and machine-readably.

The governance gap can be closed. But closing it requires acknowledging what it is: not a problem of insufficient labels, but a problem of absent formal epistemology. The primitive exists. The architecture is built. The cases are here.

The standard is ready.

References

Advisory Committee on Evidence Rules, Agenda Book, November 2024 Meeting. United States Courts.

Advisory Committee on Evidence Rules, Agenda Book, May 2025 Meeting. United States Courts.

Delfino, R. A. (2025). "Deepfakes on Trial 2.0: A Revised Proposal." Submission to the Advisory Committee on Evidence Rules, April 2025.

EU Artificial Intelligence Act, Regulation (EU) 2024/1689, Article 50 (Transparency obligations for providers and deployers of certain AI systems).

Europol (2025). "Operation Cumberland." Press release, February 28, 2025.

FCC (2024). Notice of Apparent Liability for Forfeiture, Steve Kramer / Biden AI Robocall. May 23, 2024.

Goldman, A. I. (1976). "Discrimination and perceptual knowledge." Journal of Philosophy, 73(20), 771–791. [Fake barns case]

Gonzalez, V. (2026). The Formal Logic of Modulign, v1.1.0. Open specification. modulign.org

Gonzalez, V. (2026). Modulign Standard v3.0: A Dimensional Address Grammar for Observable Reality. Open technical specification. modulign.org

Gonzalez, V. (under review). "Structural Dissolution of the Gettier Problem through Address-Theoretic Epistemology." Submitted to Analysis.

Grimm, P. W., Grossman, M. R., Linna, D. W., et al. (2024). "Deepfakes in Court." University of Chicago Legal Forum.

League of Women Voters of New Hampshire v. Kramer, No. 24-cv-73 (D.N.H. 2025).

Linna, D. W., et al. (2024). "Deepfakes in Court." Northwestern Law & Economics Research Paper No. 24-18.

NCMEC (2024). "Generative AI CSAM Is CSAM." National Center for Missing & Exploited Children.

NIST AI Risk Management Framework (AI RMF 1.0), January 2023.

State v. Puloka, King County Superior Court, Washington, 2024.

Take It Down Act, Pub. L. 119-__ (signed May 19, 2025).

Thorn (2025). "The ENFORCE Act: Critical Updates to Federal Law for Addressing AI-Generated CSAM Offenses."

United States v. Alvarez, 567 U.S. 709 (2012).

MODULIGN™ · Synthetic Content White Paper · v1.1.0 · 2026

VinceGonzalez@me.com