EVIDE GOVERNANCE LAB

From Research Protocols to Real-World Defensibility

What each Lab project means for your legal defensibility

emanuel
celano
10.07.2026
19:18:11
GMT+02:00

Every organization deploying AI eventually faces the same question:

Can we still defend this decision months after it happened?

DECISION

↓

Evidentiary Deposit - proves the event happened

↓

Authority Continuity - confirms authority stayed valid

↓

Epistemic Stabilization Buffer - confirms the decision held stable

↓

Governance Preservation Loop - keeps checking for the life of the process

↓

DEFENSIBLE IN A FUTURE DISPUTE

EVIDE documents. It does not decide. Every capability on this page produces evidence for a human or a court to evaluate - none of them determines an outcome.

Two further Lab studies - Observer Reconstructability and Authority Visibility Asymmetry - validate the assumptions this chain relies on. See Sections 2–3.

The EVIDE Governance Lab develops and validates the protocols that answer that question. Each project below answers a different piece of it - what happens when an AI-assisted decision is later challenged, in a dispute, an audit, or a courtroom.

Six validated Lab projects: four operational evidentiary capabilities, and two controlled calibration studies that confirm the theoretical foundations the other four rely on.

Together, these projects form a progressively layered evidentiary architecture, not six independent research efforts.

Every entry follows the same structure: the question it answers, a real-world scenario, what is missing without it, what changes with it, and what it concretely provides.

1. Evidentiary Deposit

CORE QUESTION

> ***Could we prove today that a decision made by an AI system happened this way, at this time, and was not altered afterward?***

REAL-WORLD SCENARIO

A company uses an AI system to approve or reject requests - loan applications, job candidates, published content. Months later, a customer disputes the decision. The company has only generic application logs: no hash, no independent certainty that the log was not modified afterward, no verifiable timestamp.

WITHOUT THIS

The decision exists only in the system's memory - repeatable in words, not provable.

WITH THIS

Every intake produces a canonical SHA-256 hash of the payload, a UTC timestamp anchored to that hash, and an integrity classification (stable / degraded / broken) captured at the moment of entry.

This classification - FCC - recurs across every project in this document; it is the one signal every Lab schema computes. A verified test in Section 6 (Governance Preservation Loop) confirms other engines evaluate independently of it, rather than simply relaying it.

CONCRETE VALUE

Creates a verifiable technical record that can be independently re-examined, even long after the event occurred. This record rests on:

- A canonical SHA-256 hash of the submitted content
- A UTC timestamp cryptographically anchored to the payload hash
- A structured, reproducible payload format (canonical JSON)

WHAT THIS PROVIDES *a stronger evidentiary basis for defending a position. It does not determine legal outcomes or replace judicial evaluation.*

---

2. Observer Reconstructability (v0.1)

CORE QUESTION

> ***Does the same AI-mediated event produce the same accountability read for everyone who examines it - or can the read change depending on who is looking, and do we know exactly why?***

REAL-WORLD SCENARIO

Two parties to the same event - for example, a platform operator with full internal visibility and an external auditor or counterparty with only partial access - each examine the identical AI-mediated event. If their formal read of its integrity and accountability differs, is that a flaw, a sign of manipulation, or simply an expected consequence of each party seeing a genuinely different slice of the same event? Without a documented answer, a divergence in findings looks arbitrary - and arbitrary looks like something to hide.

WITHOUT THIS

A divergence between two parties' formal readings of the same event has no documented explanation - it looks inconsistent, or worse, manipulated.

WITH THIS

Under controlled conditions, the Lab confirmed that a party's declared visibility genuinely and predictably shapes the formal read - including a specific, non-obvious finding: a formal accountability-collapse finding requires documented continuity evidence, not merely reduced visibility. An observer who simply cannot see everything does not, by design, trigger a false accountability alarm.

CONCRETE VALUE

Distinguishes a genuine accountability failure from a mere gap in what a given observer could see - protecting against both a false alarm (declaring collapse because one party's visibility was limited) and a false reassurance (assuming an event was fine because the less-informed observer saw nothing wrong).

STATUS: CONTROLLED CALIBRATION *Confirms the mechanism behaves as architecturally intended under pre-defined conditions with self-declared observer positions. Validation against genuinely independent observers of a real, unplanned event is reserved for a later stage.*

WHAT THIS PROVIDES *confidence that the underlying inferential engine behaves consistently and as intended under known, controlled conditions. It does not yet constitute a field-validated capability on its own - broader validation is reserved for a later experimental stage.*

3. Authority Visibility Asymmetry (v0.2)

CORE QUESTION

> ***If a party simply states that they had full visibility into who authorized an action, does saying so change what the system formally concludes - or does the system respond only to what it can actually verify?***

REAL-WORLD SCENARIO

A counterparty in a dispute submits a detailed narrative of exactly how much visibility they had into who authorized an action - the full approval chain, the delegation path, every step. If the system's formal findings shifted simply because a fuller story was told, the record's core conclusions could be shaped by how persuasively a party describes their own visibility, rather than by what can actually be verified.

WITHOUT THIS

It would be unclear whether the system's formal accountability findings can be influenced simply by how richly a party narrates what they claim to have seen.

WITH THIS

The Lab confirmed this across two severity levels - including the most extreme case tested, where one observer declared zero visibility into the authority propagation path, only the state at the moment of bind. Even there, the system's core findings (integrity, accountability, oversight risk) stayed identical. Only concrete, independently structured signals moved.

CONCRETE VALUE

Confirms that the system's core accountability findings cannot be shaped by how a party chooses to narrate their own visibility into who authorized something - protecting the integrity of the formal record from persuasive but unverifiable metadata.

STATUS: CONTROLLED CALIBRATION *Single-variable isolation, Option A only (the system's own visibility signal held constant), tested across two severity levels (Test Case A, Test Case B). A paired experiment that varies the system's actual visibility (Option B) is a separate, reserved test not yet run.*

WHAT THIS PROVIDES *confidence that the underlying inferential engine behaves consistently and as intended under known, controlled conditions. It does not yet constitute a field-validated capability on its own - broader validation is reserved for a later experimental stage.*

4. Authority Continuity (AuthorityLock)

CORE QUESTION

> ***Was the person or system that authorized this action still authorized when the action actually happened?***

REAL-WORLD SCENARIO

An AI system executes an action - sending an external communication, approving a contract - based on an authorization granted days earlier by a person or another system. In the meantime, that person changes

role, or the authorization is revoked, but the action proceeds anyway. Who is responsible, if the authorization had already lapsed in practice?

WITHOUT THIS

We can only say a permission flag existed - not whether it was still true when the action executed.

WITH THIS

AuthorityLock checks whether the authority chain remained intact between declaration and execution (intact / degraded / unverifiable), evaluated as an isolated, dedicated dimension. The result is ingested from an independently established authority layer, not inferred by the EVIDE engine itself.

CONCRETE VALUE

Distinguishes “we had an authorization” from “the authorization was still valid when it happened” - the difference that matters when a defense rests on a permission granted well before the contested event.

WHAT THIS PROVIDES *a stronger evidentiary basis for defending a position. It does not determine legal outcomes or replace judicial evaluation.*

---

5. Epistemic Stabilization Buffer (ESB)

CORE QUESTION

> ***When does a decision become stable enough to be defended?***

REAL-WORLD SCENARIO

A risk-assessment system issues an immediate verdict on a transaction. In the following minutes, new signals arrive - a fraud check, a missing data point that resolves late - that change the picture. If the initial verdict is crystallized as final too early, the company ends up defending a decision made on incomplete information.

WITHOUT THIS

The decision is born and crystallized in the same instant - no observation window.

WITH THIS

The decision remains inside a stabilization window before being declared final, with a complete trace of every signal observed during that window (causal persistence, stability trend, continuity state).

CONCRETE VALUE

Protects against the claim “you decided rashly.” Demonstrates that the decision passed through a verification period before being finalized, with a full record of what was observed during that period.

WHAT THIS PROVIDES *a stronger evidentiary basis for defending a position. It does not determine legal outcomes or replace judicial evaluation.*

---

6. Governance Preservation Loop (GPL / G-LOOP-E)

CORE QUESTION

> ***Does a process remain reconstructable over time - not just at the moment it began?***

REAL-WORLD SCENARIO

A long-running AI process - a weeks-long approval workflow, a system managing an ongoing customer relationship - is validated at the start. Over time, the authority that authorized it changes, a signal becomes blocked, the system's visibility degrades. If nobody re-checks periodically, the company only discovers -

possibly during a dispute - that the evidentiary basis had already broken down long before the problem surfaced.

WITHOUT THIS

Verification happens once, when the process opens, and never again.

WITH THIS

The engine repeats the reconstructability check at every “heartbeat” of the process - who holds authority, whether any signal is blocked, whether the system remains observable - producing an operational verdict (continue / slow down and flag / stop) that surfaces degradation when it happens, not months later.

> **VERIFIED FINDING** The reconstructability engine is not an FCC proxy. A dedicated test held Forensic Cross-Check deliberately stable while breaking only who holds authority - and the engine still returned a stop verdict. This confirms the engine evaluates authority on its own terms, not by reading FCC as a shortcut. Two independent checks, not one check wearing two names.

CONCRETE VALUE

Shifts protection from “we checked once at the start” to “we kept checking for the entire duration.” Critical when a dispute concerns something that happened partway through a long process, not at its opening.

WHAT THIS PROVIDES *a stronger evidentiary basis for defending a position. It does not determine legal outcomes or replace judicial evaluation.*

At a Glance

The table below summarizes all six projects side by side.

Project	Core Question	Without	With	Key Terms
Evidentiary Deposit	Can we prove the event existed?	Event remembered	Independent proof the event occurred	FCC (+ Evidentiary Profile)
Observer Reconstructability *	Does the read change by observer?	Divergence unexplained	Divergence explained, not assumed	Declared Visibility
Authority Visibility Asymmetry *	Does a visibility claim alone move findings?	Narrative could sway findings	Only verified signals move findings	Declared Visibility
Authority Continuity	Was authority still valid?	Authority assumed	Continuous authorization evidence	AuthorityLock
Epistemic Stabilization Buffer	Was the decision stable enough?	Immediate crystallization	Proof the decision was stabilized first	ESB
Governance Preservation Loop	Did reconstructability survive over time?	One-time verification	Proof reconstructability was maintained	GPL / G-LOOP-E

* *Controlled calibration experiment, not yet a field-validated capability. See the corresponding section above for scope.*

Legend - Key Terms

FCC - Forensic Cross-Check - The core integrity classification computed at intake (stable / degraded / broken), confirming whether a record's internal signals stay consistent. A verified finding in Section 6 confirms other Lab engines evaluate independently of it, not by relaying it.

DWC - Decision Wave Compression - An inferred signal flagging when decision throughput has compressed to the point of risking premature closure.

FAC - Formal Accountability Collapse - An inferred signal flagging when the chain of accountability has broken down, even if individual steps still look correct in isolation.

OVC - Oversight Viability Collapse - A derived reading of FCC stable plus DWC critical: visible compliance, non-viable governance underneath.

Declared Visibility (Observer Position) - What a party states about how much of an event they could see - tested to confirm whether a stated visibility can shape the system's formal findings.

AuthorityLock (Authority Continuity) - Verifies that the authority behind an action was still valid when the action executed - not only when it was declared.

ESB - Epistemic Stabilization Buffer - A window of continued observation before a decision is crystallized as final, capturing whether the initial reading holds stable over time.

GPL / G-LOOP-E - Governance Preservation Loop - Repeats the reconstructability check throughout a process's life, not only once at the start. “GPL” is the Lab's slug; “G-LOOP-E” is the working-paper name - same concept, two names.

A Note on Scope

Every capability described in this document provides a stronger evidentiary basis for defending a position. None of them determines a legal outcome or replaces judicial evaluation. This distinction is deliberate: EVIDE documents, it does not decide.