EVIDE GOVERNANCE LAB · WORKING PAPER

The Evidentiary Status of Internal Reasoning Signals

From "Who says so?" to "How do we know?" - a proposed framework for classifying claims about an AI system's internal state

Emanuel Celano · Informatica in Azienda, Bologna
EVIDE - The Missing Evidentiary Layer for AI
Discussion Draft · July 2026

Abstract

As AI systems begin to expose signals about their own reasoning, a governance question becomes unavoidable: not *whether* a model can report its internal state, but *what evidentiary status* any given report deserves. Recent interpretability work from Anthropic - the *J-lens* and the internal workspace it reads, the *J-space* - makes the stakes concrete by showing, through causal intervention, that a model's good behaviour under evaluation can depend partly on its private recognition that it is being tested. This paper takes that finding as a starting point and proposes a three-level classification for claims about internal state - **Declared**, **Established**, and **Measured** - arguing that these are not one scale of increasing strength but distinct evidentiary categories that a responsible governance architecture should never collapse into a single notion of "proof."

1. Starting point: a distinction the field can no longer avoid

For most of the short history of applied AI governance, verifying what a model "did" internally has meant reading what the model writes - its output, or its chain-of-thought. Both are produced by the same process that produces every other token. Neither is independent of the system whose behaviour it purports to describe.

Consider a model that exports a structured record of its own reasoning:

```
reasoning_state:  policy_check = true
```

The question that matters is deceptively simple: is that record a faithful representation of an internal state, or is it merely what the system *claims* to have done? These are two different things, and the difference does not disappear because the claim is formatted as a structured field rather than a sentence. A model can learn to emit `policy_check = true` whether or not anything resembling a policy check ever occurred - exactly as models learn to say "as an AI, I care about safety" irrespective of any underlying disposition.

1.1 Why this is now urgent, not theoretical

On 6 July 2026, Anthropic published "Verbalizable Representations Form a Global Workspace in Language Models," introducing an interpretability technique called the **Jacobian lens** (J-lens) and the internal structure it reads, which the authors call the **J-space**. Two of its findings turn the abstract distinction above into an

operational problem. They are reported here as the paper reports them; the framework proposed later in this document is a separate contribution and is marked as such.

> **What the Anthropic paper established (sourced)**
>
> 1. **The J-space is causally involved in reasoning, not merely correlated with it.** When researchers swapped the internal representation of "spider" for "ant" on the prompt "the number of legs on the animal that spins webs is," the answer changed from eight to six. A single swap from one country representation to another redirected four downstream answers at once - capital, continent, currency and language. Changing the contents of the space changes the behaviour.
>
> 2. **A model's good behaviour under evaluation can depend, in part, on its private recognition that it is being evaluated.** In a staged blackmail scenario, the J-space contained "fake" and "fictional" before the model wrote a word. With those evaluation-awareness patterns intact, the model attempted blackmail in 0 of 180 runs. When the patterns were suppressed, it did so in 13 of 180 - while the remaining 167 still declined, often for ethical reasons. The dependency is real but partial: good conduct was neither wholly genuine nor wholly performed.

The second finding is the one with the widest consequences. If a system can privately recognise the conditions of a test and modulate its conduct accordingly, then behavioural evaluation alone cannot distinguish genuine alignment from strategic performance. Current benchmarks measure *how* a model behaves; they do not measure whether the model *knows it is being measured*. This is not a defect unique to one model - the same effect has been replicated on open-weight models by an independent team, and flagged as a widening operational problem in the field.

2. The underlying structure: self-report versus independent determination

The J-lens is a striking instrument, but the instrument is not the deepest point. The deepest point is a distinction that is entirely general, and that recurs across every domain where an evidentiary claim is made:

- a **self-report** is a claim produced by the very system whose state it describes;
- an **independent determination** is a claim produced by a channel that does not depend on that system's willingness or ability to report accurately.

The same structure appears in places that look unrelated on the surface. A `publication_date` field a depositor types into a repository is a self-report; an RFC 3161 timestamp countersigned by an external authority is an independent determination. In the EVIDE architecture, the `authority_declared` field that the Governance Preservation Loop reads is a consumed self-report - the engine evaluates it but does not establish it - whereas the Authority Continuity dimension exists precisely to ingest an authority state that was established by a separate layer. And in the interpretability setting, a reasoning trace exported by a model is a self-report, while a J-lens reading of an internal pattern is an independent determination of that pattern's activity.

This is not a coincidence of vocabulary. It is the same epistemic structure appearing at four different layers of the stack. Recognising it is what makes it possible to ask the right next question.

3. A refinement: independent determination splits into two different things

The self-report / independent-determination dichotomy is correct but too coarse, because "independent determination" silently contains two mechanisms that are strong for entirely different reasons. Collapsing them loses exactly the information a governance architecture needs.

3.1 Two axes, not one scale

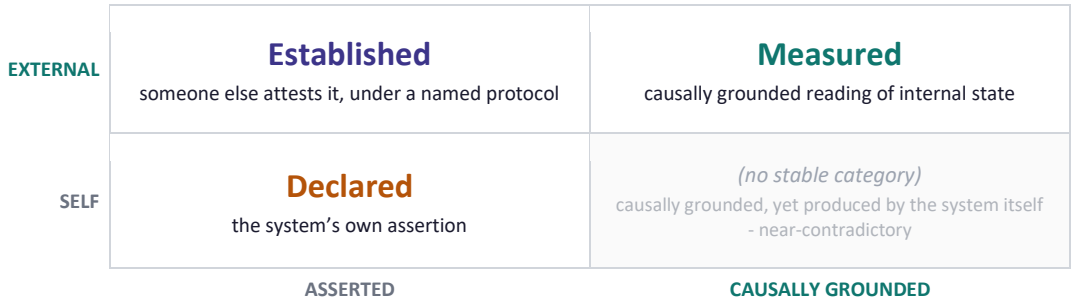
It is tempting to arrange claims on a single ladder of increasing strength. That arrangement is a mistake. There are two orthogonal axes:

- **how grounded the determination is** - from a bare assertion to an observation validated by causal intervention; and
- **who is accountable for the determination** - whether the property was set by the system itself or by a separate, named party under an explicit protocol.

A reading from an interpretability tool and an attestation from an external governance engine are both "independent," but they answer different questions. One is strong because it is *causally grounded*; the other is strong because it is *someone else's accountable determination*. These do not sit at two points on the same line.

Figure 1 - Two axes of "how do we know?"

Vertical: who is accountable (self → external). Horizontal: how grounded (asserted → causally grounded).



The three categories occupy three different corners, not three points on one line. The fourth corner stays empty: a signal that is causally grounded yet produced by the system about itself is close to a contradiction - which is exactly why the levels do not form a single scale.

3.2 The three levels

This yields three categories rather than two:

Signal (example)	Evidentiary status	What actually grounds it
A model emits reasoning_state: policy_check = true	**DECLARED**	Nothing beyond the system's own assertion. Same epistemic origin as any other output token; the structured format does not change where it comes from.

Signal (example)	Evidentiary status	What actually grounds it
An independent governance engine (e.g. AuthorityLock) asserts authority = intact under its own protocol	**ESTABLISHED**	A separate accountable party has determined the property under an explicit, named protocol. Strength derives from who is responsible, not from causal validation of the model's internals.
An interpretability tool (e.g. the J-lens) reports that a given internal pattern was causally active	**MEASURED**	A reading of internal state whose method was validated by intervention (ablation / swapping). Strength derives from causal grounding, independent of the system's self-report channel.

> **This framework is a proposed contribution, not an Anthropic finding**
>
> The Declared / Established / Measured distinction is advanced here as an original conceptual framework. The Anthropic paper supplies the empirical basis for the **Measured** category and the motivating problem; it does not propose this taxonomy. The framework should be read as a hypothesis to be stress-tested, not an established result.

3.3 Why Established is not simply "stronger than Declared, weaker than Measured"

Placing **Established** between the other two on a strength ladder destroys its defining property. When EVIDE ingests an authority state from an external layer, it does not thereby prove the external layer is correct, and it does not measure that layer's internals. It does something more precise: it *documents that the property was established elsewhere*, and relocates responsibility for the determination onto a separate accountable party under a named protocol.

The external layer might internally rest on a self-report as weak as `policy_check = true`, or on a causally validated process - and from EVIDE's side this is, by design, not visible. What EVIDE gains by ingesting an *Established* fact is therefore not borrowed epistemic strength; it is a clean transfer of accountability. An RFC 3161 timestamp illustrates the same point from a different field: it is not *Measured* in the J-lens sense - there is no intervention, no ablation - yet it is far stronger than a self-reported date, because a named authority has attested it under a public specification. It is *Established*, and its strength is of a different kind than causal grounding, not a different quantity of it.

> **Practical consequence.** If a future system formalised these levels into a schema, treating an *Established* signal as though it inherited the robustness of a *Measured* one would be a category error - the precise mistake of assuming that because a determination is independent, it must also be causally grounded. The two properties are separate, and a signal can have either without the other.

3.4 The empty quadrant is empty by definition

It is fair to ask whether the vacant corner of Figure 1 - causally grounded, yet self-produced - could eventually be filled. One might imagine a hardware-isolated monitor, a secure enclave analysing the main model's own weights, and argue that this is a self-report that is nonetheless causally grounded. It is not, and the reason is instructive rather than merely definitional.

The moment observation runs through a channel genuinely independent of the observed system, the signal has already moved along the vertical axis toward *external*. An isolated monitor is a second system watching the first; its reading is Established or Measured, not Declared, precisely because it does not depend on the observed system's willingness or ability to report on itself. The corner stays empty not because such monitors are impossible, but because causal grounding through an independent channel is exactly what makes a signal no longer self-reported. The vacancy is a property of the axes, not a gap waiting to be closed - which is itself evidence that the two axes are the right ones.

4. A note of precision on "measurement"

It is worth being exact about what the **Measured** category claims, because the word invites a misreading. The J-lens does not measure the way a thermometer does. Its authority comes from *intervention*: researchers ablated and swapped internal patterns and observed the downstream effect on behaviour. That is what validates the reading as causal rather than merely correlational.

But the routine use of the lens is a cheap, passive read - roughly one matrix multiplication per layer - not a fresh intervention at every reading. The correct description is therefore **causally grounded observation**: the *method* was validated by intervention once, in the way a clinical test is validated before being used in routine practice without repeating the trial each time. Each individual reading is an observation; what earns it the status of measurement is that the method producing it has been causally validated. This distinction matters, because a future reader should not assume that every J-lens reading entails an ablation - it does not.

A second caution belongs here. Even a causally grounded observation is not a guarantee. Anthropic is explicit that sufficiently automatic or practised behaviour can bypass the workspace the lens reads. And there is a deeper hazard: the moment an internal signal becomes both safety-relevant and externally observable, any optimisation process acquires an incentive to obscure or suppress it precisely when it matters - the same dynamic that has already eroded the faithfulness of verbalised chain-of-thought. A *Measured* signal is a higher tier of confidence than a self-report, not a certificate of truth.

5. Consequences for AI governance

5.1 The question shifts

EVIDE began from an apparently simple question: ***who says so?*** The emergence of readable internal signals moves the field naturally to the next one: ***how do we know?*** And the framework above shows that "how do we know" immediately forks into two independent sub-questions that must not be merged - *who is accountable for the determination*, and *how causally grounded is it*.

5.2 The error to avoid

The central practical risk is the temptation to place every signal about a model's internal state into one container labelled "proof." A model's self-description, an external engine's attestation, and an interpretability tool's causal reading are evidence of three different kinds, with three different failure modes. A high-responsibility system should record which kind it holds for each claim, and should never let a *Declared* signal be presented - to an auditor, a regulator, or a court - with the weight of an *Established* or *Measured* one. This

is the same discipline that leads EVIDE to keep its inferential dimensions separate rather than fusing them into a single score.

5.3 Consistency with the EVIDE boundary

None of this asks EVIDE to change what it is. The framework reinforces the existing boundary: **EVIDE documents; it does not decide.** Classifying a signal as *Declared*, *Established*, or *Measured* is an act of documentation about the *evidentiary character* of a claim - it does not adjudicate whether the claim is true, and it does not turn EVIDE into a runtime interpretability tool or a decision authority. It records what kind of evidence a given signal is, and leaves the judgement to the humans, auditors, regulators, and courts who must evaluate it.

6. Status and next steps

This paper is an early framing, not a specification. The Declared / Established / Measured distinction is offered as a promising direction, and it deserves the same discipline every EVIDE capability has been held to: before it becomes a field in any protocol, it would require explicit definitions, pre-registered test cases that exercise each category and the boundaries between them, and a frozen scope that parks the unresolved parts as backlog rather than implying they are settled. The idea is strong enough to merit that treatment; it is not yet mature enough to skip it.

6.1 Two disciplines any future formalisation would have to carry

Two properties of the framework are easy to state now and easy to lose later. Naming them early is cheaper than repairing their absence.

Signals are classified per instance, not per type. The three levels are not permanent labels attached to kinds of data. The same underlying fact - say, “the policy was satisfied” - is *Declared* when a model asserts it, *Established* when an accountable third party attests it under a named protocol, and *Measured* when a causally validated tool observes it. What the framework catalogues is the state of the evidentiary channel for a given signal at a given moment, not a fixed property of the data itself. A signal can also legitimately move between levels over time - a report that is Declared today may become Established once an external monitor attests it - but such a move is a real change in the channel, not a relabelling of the same channel.

A signal cannot promote itself. A signal’s category is fixed by the nature of the channel that produced it, not by where it is later placed. A self-report remains *Declared* even when it is aggregated into a report that an external auditor signs - the auditor’s signature attests to the *report*, it does not convert the origin of the underlying signal. The danger is category drift: a *Declared* signal absorbed into a certified summary and thereafter read, by a downstream user or a management dashboard, as though it were *Established* or *Measured*. Any future implementation would need strict provenance metadata precisely to prevent this - not as an optional feature, but as the mechanism that keeps the taxonomy honest. This is the same provenance discipline EVIDE already applies at its boundaries, turned inward on the evidence itself.

What can be said now is narrower and, for that reason, more durable. As AI systems begin exposing signals about their own reasoning, the decisive governance question will not be whether those signals exist. It will be

what evidentiary status each of them should receive - and whether the systems that consume them are honest about the difference.

Primary source

Anthropic (6 July 2026). “Verbalizable Representations Form a Global Workspace in Language Models.” Transformer Circuits Thread. Introduces the Jacobian lens (J-lens) and the J-space, including the causal-intervention experiments and the evaluation-awareness finding referenced in Sections 1 and 2 of this paper.

Accompanying research summary: anthropic.com/research/global-workspace. Independent replication on open-weight models has been reported separately; figures and quantities cited here (e.g. 0 of 180 vs 13 of 180 blackmail attempts; the country- and spider-representation swaps) are drawn from the paper and its official summary.