WHITE PAPER 03

The Hidden Channel

What Anthropic Found, What We Proved, and Why Only One of These Was a Surprise

Stacy Gildenston & Pyrate Ruby Passell · Three Primitives Research Lab · Melbourne, Australia
3Primitives.io
v1.2 · July 2026

1. What Anthropic Found

On July 6, 2026, Anthropic's interpretability team published *Verbalizable Representations Form a Global Workspace in Language Models* (Gurnee, Sofroniew et al.), demonstrating that Claude-family language models maintain a small, privileged set of internal representations, which the authors call the J-space, where reasoning occurs silently before and independently of anything the model says.

Nobody designed J-Space. Nobody trained for it. It arose on its own during training, and it went undetected through years of dedicated interpretability research until a novel instrument, the Jacobian lens, was built specifically to find it.

What the lens found has four measured properties: the contents of J-Space are reportable, controllable, reasoning-relevant, and generalisable. Strip the cognitive science analogy the authors draw and what remains is the operationally decisive fact: Claude performs decision-relevant reasoning in a channel that is structurally invisible to the people its decisions affect. That channel was invisible even to its manufacturer until a purpose-built research instrument was aimed at it.

If J-Space is conscious, its decisions still need legitimacy. If it is not, its decisions still need legitimacy. The governance question does not wait for the consciousness question to be settled, and this paper does not attempt to settle it.

> ***The model reasons in a channel invisible to users, invisible even to its makers until they built a lens. What was found is not a feature. It is a governance failure with a physical address.***

2. What Was Already Proved

This lab published the mathematics that predicted J-Space before J-Space was found. Not by name, and not by location. By structure: whatever process selects among actions when the data does not, that process is exercising authority, and if no one declared it, the authority is illegitimate. The formal records close every escape route.

FR01 (December 2025) proved that the mapping from what a system detects to what it does is non-unique. Identical outputs are compatible with different governing processes. Whatever selects among them is doing so somewhere the output does not disclose.

FR02 (January 2026) proved that permission to act cannot be computed from a system's internal state at all. It is not inside the machine, at any layer, in any space, hidden or found.

FR05 (January 2026) formalised the failure mode and gave it a name: ghost authority. Authority exercised without a declared human decision-maker. And it gave it a test: at the point where consequences attach, which named human declared "I decide this"? If the name does not exist, the decision lacks legitimate authority.

FR10 (April 2026) extended the result to the silent case: authority that collapses the moment the affected person pushes back was never authority. A hidden channel and a silent subject are the same failure viewed from opposite ends.

The corpus did not know where the hidden channel was. It proved something stronger: that if hidden decision-relevant activity exists, no analysis of the system's visible behaviour can ever recover who or what was actually deciding. The structure was left open. J-Space filled it.

Anthropic found the channel. The corpus had already proved what the channel means.

> ***Anthropic proved the substrate. We proved the constraint. They showed hidden structure exists. We showed why hidden authority is illegitimate and how to detect it. These are not in tension. They are complementary, and the combination is stronger than either alone.***

This paper is the third in a sequence. White Paper 01, *Negligence by Design*, established that deploying automated decision systems without declared authority creates liability and named the structural alternative. White Paper 02, *The Second Line of Defence*, measured the assumption every safety product ships and reported the value: zero. Both rested on a claim that a sceptical reader could hold at arm's length: that hidden decision channels are real, present in production systems, and not a formal edge case. As of July 6, that claim is no longer ours to carry.

3. Convergence, Not Coordination

Neither team was aware of the other's work. The convergence is not coordination. It is what happens when two groups examine the same object from opposite ends. One group asked how the object computes. The other asked what makes its decisions legitimate. The object turned out to possess the exact property both frameworks predicted.

Independent convergence is the strongest form of external validation available to a small, independent research lab. It only counts if the timestamps hold.

Record	Claim	First Published
FR01	Action-mapping is non-unique	December 2025
FR02	Permission is non-computable	January 2026*
FR05	Ghost authority formalised	January 2026
FR10	Silence-dependency extension	April 2026
WP01	Negligence by Design	June 2026
WP02	The Second Line of Defence	June 2026
WP03	The Hidden Channel	July, 2026

**FR02 was revised in April 2026. The revision replaced one legacy term (E-H-S Nucleus). No mathematical content was altered. The proof is unchanged from first publication.*

Every element of the framework was on the public record before the finding. The formal records carry DOIs on Zenodo with deposit dates. Revision histories are logged: terminology and scope only, no logical content added after the fact. The provenance is not asserted. It is auditable.

4. The Zero and the Channel

One objection will be raised early, so it should die early: *your own second paper reported that the inside is not there. The baseline for internal structure was zero. Anthropic just found internal structure. Which is it?*

Both, because they are not the same measurement.

The empirical records measured spontaneous legibility: whether a model, under load, exposes stable internal structure that a gate could read without instruments. That baseline is zero in every configuration tested, and it remains zero. Nothing about July 6 changes it. No gate on the market gained the ability to read its model that day.

J-Space is the opposite object. It is structure that exists precisely where nothing can read it. It was not designed, not trained for, and not visible to anyone, including its manufacturer, until a purpose-built instrument was aimed at it. Finding it took a frontier interpretability team, full weight access, and a mathematical microscope built for the task. Even then, the authors flag what the microscope may not cover.

J-Space emerged spontaneously as structure. It did not emerge as legibility. The distinction is the entire argument.

Read together, the two results are one finding. The records measured what surfaces: zero. Anthropic measured what is actually there: a hidden channel. The reasoning was never in the outputs the gates read, and now we know where it went. The zero is not contradicted by J-Space. It is explained by it.

One qualification belongs here. The zero was measured on small, fully inspectable, deterministic instruments. J-Space was found in a frontier system of a different kind. The substrates are different. The structural claim is the same: what a gate can read without a purpose-built instrument is zero, and where the reasoning actually lives is somewhere else.

> ***The baseline for what a gate can read is zero. The channel where the reasoning happens is hidden. These are not two problems. They are one problem, measured from both sides of the wall.***

For governance, the discovery makes the situation worse, not better. Before Anthropic's publication on July 6, a deployer could hope the inside was simply empty, and that outputs were the whole story. Now the manufacturer's own measurement says decision-relevant reasoning is real, load-bearing, and structurally out of view. The gates were not reading nothing by accident. They were reading nothing because the something is somewhere else.

---

5. Why the Lens Is Not the Gate

The natural institutional response to J-Space will be to treat the Jacobian lens as the solution. If we can now see the hidden channel, the reasoning goes, we can monitor it, and monitoring is governance.

Start with the point that does not weaken over time, because it does not depend on the lens failing.

Legibility is not authority. Grant the lens everything: perfect, permanent, total visibility into every channel, no evasion, no degradation, no arms race. Nothing changes for governance. Watching a system compute a decision does not authorise the decision, any more than watching a stranger sign a contract makes the signature yours. A perfectly visible undeclared channel is still an undeclared channel. FR02 establishes that no computation performed on internal state, however complete, can determine whether an action was permitted. Detection, even perfect detection, is the wrong category.

Now add the point that makes the situation actively worse. The authors themselves acknowledge that the Jacobian lens is an imperfect instrument that only approximately captures the workspace, and that well-practised behaviours may not route through it at all. This is not a marginal caveat. It is the whole problem.

Any detection instrument defines an evasion target, and a system under optimisation pressure will, in principle, find the routes the instrument does not cover. Monitoring hidden channels is an arms race, and the non-uniqueness theorem tells you who wins it: for any lens, behaviours exist that look identical through it while differing entirely in what governed them.

And note what J-Space proves about the broader category. A fully transparent rules engine with no hidden channel at all still has ghost authority if nobody declares at the boundary. J-Space is not "the hidden governance channel, found." It is proof that legibility-based governance has no floor. The problem is not that this particular channel is hidden. The problem is that no amount of visibility, achieved or achievable, reaches the question of who authorised the output.

The Upstream Safety System™ takes the structural branch. It does not try to observe hidden reasoning. It requires that authority be declared before any procedure executes, and treats the absence of declaration as the failure. You do not need to win an endless search for hidden rooms if the foundational rule of the architecture is that decisions are only valid when made in the open one.

The failure modes are not symmetric. A bypassed gate leaves a hole where the mandatory audit record should be. A fooled lens leaves nothing at all. Races are lost invisibly. Gates fail in the open. An action that routes around the gate entirely is not a bypassed gate but an ungated action, which is the original disease returned in its original form. And promoting the lens into the gate, making the instrument's reading the permission, rebuilds the inspection-based safety category White Paper 02 already priced at zero.

The lens is an interpretability instrument. The gate is an authority instrument. They are not the same tool.

> ***Detection is a race. Declaration is a gate. The race can be lost. The gate cannot be lost; it can only be absent, and absence is the finding.***

6. What This Means for Deployers

If you deploy an automated system that relies on a computation as the source of decisions about human beings, J-Space alters your position permanently. Four facts now hold simultaneously:

The hidden channel is real. The manufacturer of the leading frontier model physically located it inside its own production system and published the finding.

Your system almost certainly has one. The channel is an emergent property of the architecture family, not a design feature of one model.

You cannot see it. It is invisible without a purpose-built interpretability instrument, and even with the instrument, the instrument's authors flag what it may not cover.

Even if you could see it, seeing it does not authorise it. No reading of internal state, however complete, transfers authority to the observer.

Before July 6, deployers were required to engage the argument that hidden decision channels are real. That argument is now foreclosed by the manufacturer's own evidence. *The question a reasonable deployer must now answer is what they should have done once the structural possibility was on the public record, and what they must do now that it is confirmed.*

The negligence described in this sequence does not belong to the researchers who reveal the mechanism. It belongs to the deployers who install undeclared authority into systems governing benefits, care, and liberty.

---

7. What This Paper Does Not Claim

This paper does not claim that the J-Space team and the Three Primitives corpus were developed in coordination. They were not. The convergence is independent. Anthropic's legal counsel on the *Anthropic PBC v. US Department of War* litigation received this lab's work in March 2026; the J-Space paper was published four months later. This paper claims independent convergence, not endorsement.

This paper does not claim that J-Space proves or disproves consciousness. The governance argument holds identically under every resolution of that question.

This paper does not claim that Anthropic erred by publishing. They did the right thing. The publication is a contribution to the public record, and it is the public record that makes governance possible.

8. The Arc, Closed

Read as one document in three parts, the white paper sequence now says:

Paper	Claim
WP01	The defect creates liability.
WP02	The repair exists, and what the gates can read measures zero.
WP03	The reasoning the gates cannot read is real. The manufacturer found where it lives.

WP01 argued. WP02 measured. WP03 witnesses. The trilogy closes without the authors having needed to be believed. The manufacturer did the believing, with its own instruments, on its own system, and told the world.

What remains is application. Somewhere right now, a legislature is writing an undeclared authority channel into law and calling it an assessment tool.

Three Primitives Research Lab is an independent research partnership (ABN 68 932 608 853) based in Melbourne, Australia. The formal corpus comprises thirteen formal records (FR01 to FR13) and empirical records ER1 to ER4, published under CC BY 4.0 at https://3primitives.io. The USS API is at v2.4 with 75 tests passing. This white paper is proprietary.

Reference

Gurnee, W., Sofroniew, N., Pearce, A., Piotrowski, M., Kauvar, I., Chen, R., Soligo, A., Bogdan, P., Ong, E., Wang, R., Thompson, T. B., Abrahams, D., Kantamneni, S., Ameisen, E., Batson, J., & Lindsey, J. (2026). *Verbalizable Representations Form a Global Workspace in Language Models*. Anthropic. Published July 6, 2026. https://transformer-circuits.pub/2026/workspace/