WHITE PAPER 04

The Case with the Gate

Interior governance is provably unverifiable. This paper proves that governance is forced to the boundary where the system acts on a person, and that the boundary is fixed by the person, not the machine.

Stacy Gildenston · Three Primitives Research Lab · Melbourne, Australia
3Primitives.io
v1.1 · July 2026

Abstract

Current approaches to AI governance attempt to shape or verify the interior of the systems they govern. This paper proves that the interior is not a viable location for governance, and that this is a structural result rather than a limitation of present tools. WP03 established that where hidden decision-relevant activity exists inside a sufficiently complex system, no analysis of visible behaviour can recover who or what was deciding. This paper extends that result with a lemma: the absence of hidden channels is itself unverifiable from the exterior. Together, the two results eliminate the interior as a site of governance under any evidentiary standard. Governance is thereby forced to the one location that remains: the boundary where the system's output meets the person it acts on. That boundary is fixed by the person, not the machine. It is a definite, single-outcome, record-generating event regardless of the substrate, architecture, or capability of the system behind it. The three governance primitives, declared authority, detectability, and contestation, are shown to be boundary conditions at this point of exit. The result is invariant across substrate, including quantum and biological computation, because it was proved without reference to any specific channel or computational model. The practical form of this constraint is a case with a gate: the system operates freely inside, and nothing leaves the case to act on a person without passing through a gate that carries a declared human name, is detectable, and is contestable. The Upstream Safety System™ (USS) is the existing implementation of this constraint.

1. The Interior Problem

The dominant approaches to AI governance share one assumption: that the system can be made safe, aligned, or legitimate by acting on its interior. Alignment training shapes interior dispositions. Reinforcement learning from human feedback shapes interior reward structure. Constitutional methods shape interior policy. Interpretability inspects interior representations. These approaches

differ in method and in maturity, but they share a location. All of them treat the inside of the system as the place where governance happens.

WP03 (*The Hidden Channel*) established the limit of that location. If hidden decision-relevant activity exists inside a sufficiently complex system, then no analysis of the system's visible behaviour can recover who or what was deciding. The result is not a claim about current tools. It is a structural claim about what visible behaviour can and cannot certify. Coherent output is compatible with an arbitrary allocation of decision authority behind it. The observer sees the decision. The observer does not, and cannot, see the deciding.

This result is conditional. It states what follows if a hidden channel exists. A governance regime built on the interior can therefore attempt one remaining defence: certify that no hidden channel exists, and the conditional never fires. The lemma below closes that defence.

Lemma 1.1 (Non-verifiability of channel absence). The absence of hidden decision-relevant channels in a sufficiently complex system cannot be verified from the exterior.

Verifying absence requires enumerating the space of possible channels and confirming that each is empty. The channel space of a sufficiently complex system is not enumerable in practice. It is defined by the system's full internal dynamics, not by its design specification, and it grows with every representational degree of freedom the system possesses. Any verification effort terminates in the claim "no channel found so far," which is a statement about the search, not about the system.

This lemma now has direct empirical support. In July 2026, Anthropic researchers (Gurnee, Sofroniew et al., transformer-circuits.pub) identified a specific internal channel, designated J-Space, whose ablation flattens a model's experiential reports while preserving output coherence. The channel was found under the most favourable conditions verification will ever enjoy: complete white-box access, unrestricted instrumentation, and a research team with every incentive to look. Behavioural analysis had not surfaced it. A channel that survives those conditions until deliberately excavated is a demonstration that the channel space is not enumerable by inspection. The strongest interior certification available to any laboratory, including the best resourced laboratory in the field, is a report on where it has looked.

The two results compose. If a hidden channel exists, WP03 applies and the deciding agent is unrecoverable from behaviour. If a laboratory claims no hidden channel exists, Lemma 1.1 applies and the claim is unverifiable. There is no third branch. Interior governance is therefore unverifiable from the exterior under every evidentiary standard that matters: regulatory audit, legal discovery, and the standing of the person affected.

One clarification is required, because this paper will be read by interpretability researchers and it is not directed against them. Interpretability is valuable science. J-Space is a significant finding, and Section 6 will show that it strengthens the present argument. The claim here is narrower and compatible with the science: interpretability can discover channels, but it cannot certify their absence, and therefore it cannot function as a legitimacy mechanism. Discovery and certification are different acts. The first is possible and productive. The second is structurally unavailable.

2. The Forced Exit

If the interior cannot host governance, governance must live somewhere else. This section shows that the elimination is exhaustive: every candidate location other than the boundary either collapses into the interior problem or turns out to be the boundary under another name.

Candidate one: the interior. Eliminated by Section 1.

Candidate two: the inputs. A regime might govern what goes into the system: training data standards, pre-deployment audits, certified development processes. This appears to avoid the interior. It does not. Input governance is a claim about the interior state the inputs produce. The assertion “this training regime yields a system that decides legitimately” is a claim about interior dispositions, and verifying it requires exactly the interior access that Section 1 eliminated. Input governance is interior governance with the verification deferred. The deferral does not discharge the obligation. It relocates the unverifiable claim from the present tense to the future tense.

Candidate three: the institution. A regime might govern the humans instead: regulate the deploying organisation, assign liability, require named officers. This is frequently offered as the alternative to technical governance. It is not an alternative. It is the gate, described institutionally. An institutional requirement that a named human hold authority for a class of automated decisions is only meaningful at the point where those decisions reach people, because that is the only point at which the requirement can be checked. Institutional governance that never touches the boundary is aspiration. Institutional governance that touches the boundary is this paper’s proposal. The candidate does not compete with the boundary. It converges on it.

Candidate four: the boundary. The point where the system’s output meets the person it acts on. It is exterior to the system, so Section 1 does not apply to it. It is observable by construction, because an action that reaches a person has, by definition, left the system. It is the location at which the affected person, the regulator, and the court all stand. It is the only candidate left.

This is not a design preference. The unrecoverability result and Lemma 1.1 eliminate the interior; the interior’s elimination absorbs the inputs; the institutional candidate converges on the boundary rather than competing with it. Governance is forced to the boundary because there is nowhere else for it to be.

3. The Classical Boundary

The boundary has a property that the interior never has: it is fixed by the person, not the machine.

The receipt of a decision by a person is a definite, single-outcome, record-generating event. A loan is approved or it is not. A plan is reduced or it is not. A file is flagged or it is not. Whatever exotic dynamics operate inside the case, superposition, stochastic sampling, distributed representation, biological noise, the action on the person resolves to one outcome, and that outcome is in principle recordable. This is not a claim about physics in general. It is a claim about the specific event of a decision landing on a human being, and it holds for that event because contestation requires it to hold: a person can only contest a definite decision, and a legal system can only adjudicate one. The boundary is classical in the operative sense, which is that it admits a record.

This is why the constraint survives every substrate. A quantum system, a neuromorphic system, a biological system grown from cultured neurons, and any architecture not yet invented all share one fact: when their output acts on a person, that action is a definite event at the receiving end. The person does not become quantum because the computer is. The boundary is defined by the human end of the transaction, and the human end does not change when the machine end does.

FR02 completes this section. FR02 established that computation does not qualify to govern: the capacity to compute an outcome is not the capacity to hold authority for it, because authority is a status conferred and contested among persons, not a function evaluated. Quantum computation is computation. Biological computation is computation. A substrate upgrade changes what the system can compute. It changes nothing about what the system is qualified to hold. The question “but what if the AI runs on X” is therefore closed for every value of X: whatever runs inside the case, the qualification to govern was never inside the case to begin with.

---

4. The Geometry of the Case

FR10 proved a symmetry that gives this paper its shape. A hidden channel at the system end and a silent subject at the person end are the same governance failure viewed from opposite directions. In the first, decision-relevant activity exists that no observer can see. In the second, an affected person exists from whom no contestation signal can travel. Both are failures of detectability. Both make the allocation of authority unrecoverable. One hides the decider; the other erases the decided-upon.

The case therefore has two walls, and each wall names one failure mode. Inside the case, the system can hide. Section 1 proved this cannot be prevented and need not be: the interior is permitted to be opaque precisely because governance no longer depends on seeing into it. Outside the case, the person can be silenced: excluded by friction, unheard by design, or simply absent from the detection function. FR13 established what that absence means formally, and Section 5 places it.

The gate sits between the two walls and addresses both failures in a single mechanism. Facing inward, it forces the system to declare: nothing exits without a named human authority attached, so interior opacity never converts into exterior ghost authority. Facing outward, it gives the person

standing: the declared name, the detectable action, and the contestation path are exactly the materials a silenced subject needs to stop being silent. The gate is not two safeguards bolted together. It is one boundary condition read from both sides, which is what FR10's symmetry predicts a correct mechanism would look like.

---

5. The Gate

The three governance primitives are authority, detection, and contestation. FR12 (*The Forced Bijection*) proved that these are forced: three independent fields with no shared methodology, taken as consequences of consciousness being fundamental, map onto exactly these three primitives, and the mapping admits no alternative allocation. FR12 established that the primitives are forced. This paper establishes where they are forced to: the boundary.

The primitives are boundary conditions. They do not describe what happens inside the case, and they do not need to. They describe what must be true at the exit point, at the moment an action passes from the system to the person:

Authority must be declared. A named human, authorised for that class of decision, on the record before the action proceeds. Not derivable from the output, not reconstructed after the fact, not distributed across a process. Declared, in advance, at the gate. The gate checks configuration, not content: is there a named human with declared standing for this class of action? It does not check whether the individual decision was correct. That distinction is critical at scale: a named human holds declared authority for a scoped class of automated decisions, and the gate confirms the authority configuration exists before any action in that class proceeds. Individual review of each decision is a process that stays inside the case, where it may or may not occur according to the organisation's own design. The gate does not require it, because the gate's question is about legitimacy, not quality. The class itself must be bounded. A declaration covering "all decisions" is unbounded and does not satisfy δ, because it names a scope too broad to carry meaningful responsibility. The gate checks at runtime that a valid declaration exists for the specific class of action being taken, not that a blanket declaration was filed at design time.

The action must be detectable. The passage through the gate generates the record that Section 3 showed the boundary always admits. What acted, under whose authority, on whom, and when. The record exists because the gate is the only path out, so nothing that acts on a person can fail to appear in it. Detection is satisfied when the record reaches the affected person, not when it is filed. A record that exists but is never surfaced to the person it concerns leaves the subject silent in practice, which is the failure FR10 already characterised from the other end.

The affected person must be able to contest. The declared name and the generated record are the standing a person needs. Contestation is not a customer service channel appended to the system. It is a structural property of the boundary: because the decision is definite and the authority is named, there is something to contest and someone answerable.

FR13 (the ILMM Coupling Theorem) governs the failure case. When the gate cannot confirm a valid authority configuration, the condition D = ∅ obtains: no valid configuration exists. FR13 established that this condition triggers stewardship. It is not a judgment call and no human declares it; it is a testable condition arising from the detection primitive itself. Operationally, the gate stays closed and the action waits until a valid authority can be confirmed. This is the formal content behind a sentence that already runs in production: a decision that cannot be authorised is a decision that waits. Reopening is not automatic. FR13's stewardship powers govern the return path: a Structural Gap Audit demonstrating that the governance circuit has been repaired, followed by revalidation of a valid δ for the decision class.

Note what this section does not require. It does not require the system to be interpretable, aligned, truthful about its interior, or even understood by its builders. Every requirement lands on the boundary, and Section 2 proved that this is not a concession. It is the only place requirements can land and remain checkable.

---

6. Substrate Invariance

The constraint proved in Sections 1 through 5 was constructed without reference to any specific channel, architecture, or computational model. WP03's unrecoverability result quantifies over hidden channels as such. Lemma 1.1 quantifies over the channel space as such. Section 3 fixes the boundary by the human end of the transaction. At no point does the argument consult the substrate. This is why the argument cannot be obsoleted by one.

The J-Space result is the correct test of this claim, and the result passes it. Anthropic's researchers found a specific hidden channel in a specific architecture: an internal space whose ablation flattens experiential reports while output coherence survives. The corpus had already proved, before that channel was found, what any such channel means for governance: visible behaviour cannot recover who was deciding, and no certification of channel absence was ever available. The empirical finding landed inside the proved constraint, exactly where the constraint said any such finding would land.

The two results are complementary, and the division of labour is exact. They proved substrate: a real channel, in a real model, located and characterised. We proved constraint: what any channel, found or unfound, means for the location of governance. Neither result substitutes for the other. The empirical finding without the constraint is one channel in one model, patchable in the next release and silent about the next architecture. The constraint without the finding is mathematics awaiting an instance. Together they close the question from both ends: hidden channels are real, and governance was already proved to survive them, because it was never located where they live.

This is also why the framework does not need to enumerate possible hidden channels, in silicon, in quantum hardware, in cultured biological tissue, or in substrates not yet built. Enumeration is the strategy Lemma 1.1 proved unavailable, and the boundary placement is what makes it

unnecessary. A constraint that held only for known channels would be a patch. This one holds for the channel space, which is why the next discovered channel, whatever it runs on, will land inside it too.

---

7. The Distinction

Two sentences look similar and are structurally opposite.

“The AI cannot escape” is containment. Containment is a claim about the interior: about what the system can reach, compute, or circumvent. Every containment claim inherits the interior problem. It must be verified against a channel space that Lemma 1.1 proved non-enumerable, against capabilities that grow with each generation, and against an adversary that improves while the walls do not. Containment is an arms race, and the structure of the race is that you lose: the defender must close every channel, the system need only find one, and Section 1 proved the defender cannot even verify how many there are.

“The AI cannot act on a person without passing through the gate” is a boundary constraint. It makes no claim about the interior at all. The system may be arbitrarily capable, arbitrarily opaque, and arbitrarily strange inside the case. The constraint does not degrade as capability grows, because it was never a function of capability. It is a function of the human end of the transaction, and Section 3 showed that end does not move. Containment gets harder every year. The boundary constraint does not get harder, because the person receiving the decision is the same kind of event in every year.

One objection deserves its own paragraph, because it is the strongest available. A sophisticated system need never issue anything that looks like a decision about a person. It adjusts a price. It reorders a feed. It flags a file as low priority. It drafts the summary a human signs. A human acts downstream, and the system’s fingerprints are nowhere on the action. Where is the boundary now?

The answer is that the boundary was never defined by output modality. It is defined by effect on a person. An action is at the boundary when a person’s situation is altered by it, whatever syntactic form the alteration takes. The reordered feed that determines what a claimant sees, the flag that determines which file a caseworker opens, the draft that determines what the signer signs: each alters a person’s situation, and each is therefore an action at the boundary requiring declared authority, detectability, and a contestation path. A gate that checked only explicit decisions would be a gate with a hole shaped like every implicit one. The scope of the gate is the scope of effect. Mediated action does not evade the boundary. It is the boundary, reached through a longer path, and the governance question it raises, who held authority for the alteration, is the same question with the same forced answer. FR10’s effective primitive value provides the test: if the named

human's authority collapses under the first act of genuine contestation, because workload, institutional pressure, or system design made departure structurally impossible, the authority was never satisfied. A human who cannot meaningfully depart from the system's recommendation is not exercising authority. *They are wearing it.*

The Upstream Safety System™ is the existing implementation of the boundary constraint. USS v2.4 is a deterministic, rule-based governance engine with no machine learning in the decision path: classification informs, only the gate decides. Declared human authority must exist before a scoped action proceeds; where it cannot be confirmed, the gate stays closed and the action waits, which is FR13's stewardship condition running in code. The system enforcing the constraint is written by humans, readable by humans, and answerable to humans. The AI runs inside it. The implementation is documented, tested, and live at https://3primitives.io.

---

8. Implications

The argument closes as follows. Interior governance is unverifiable: proved by WP03 where channels exist, and by Lemma 1.1 where their absence is claimed. The elimination forces governance to the boundary, the single remaining location, and the boundary is fixed by the person, invariant across substrate, architecture, and capability. The three primitives, already proved forced by FR12, are boundary conditions at that point of exit, with FR13 governing the case where no valid authority exists. The practical form of the constraint is a case with a gate: the system operates freely inside, and nothing leaves to act on a person without a declared name, a generated record, and a contestation path.

For builders, the implication is that legitimacy is not a property you train into the interior. No amount of alignment work discharges the boundary requirement, and no future interpretability result can, because certification of the interior is not available at any level of tooling. The gate is not a tax on capability. It is the only mechanism by which a capable system's actions become defensible, and it costs the interior nothing: the case does not slow the system down.

For regulators, the implication is that audit has a natural location and it is not inside the model. An audit standard aimed at interiors inherits Lemma 1.1 and certifies searches rather than systems. An audit standard aimed at the boundary asks three checkable questions: was authority declared, was the action recorded, could the person contest. These questions have answers in every case, for every system, on every substrate, which is what an audit standard requires.

For anyone deploying or investing in AI systems at scale, the implication is a test that separates governed systems from ghost authority: for each action this system takes that alters a person's situation, name the human who held authority for it. A system that can answer is inside a case with a gate. A system that cannot is exercising authority no one declared, and every result in this stack, from WP01's negligence analysis through WP02's zero measurement to the present proof, converges on what that exposure means.

The formal records proved the gate is forced. This paper proves where: at the boundary, because it can be nowhere else. The gate carries a human name, is detectable, and is contestable, and the constraint holds for any system, on any substrate, at any capability, because it was never about the machine. It was always about the person at the other end.

Dependencies and References

Formal Record FR02, Three Primitives Research Lab. Establishes that computation does not qualify to govern.

Formal Record FR10, Three Primitives Research Lab. Establishes the symmetry between the hidden channel and the silent subject as one detectability failure viewed from opposite ends.

Formal Record FR12, *The Forced Bijection*, Three Primitives Research Lab. Published: April 2026. Establishes the forced mapping of three independent fields onto the three governance primitives.

Formal Record FR13, ILMM Coupling Theorem, Three Primitives Research Lab. Establishes the stewardship condition triggered by D = ∅.

White Paper WP03, *The Hidden Channel*, Three Primitives Research Lab. Published: July 2026. Establishes the unrecoverability of the deciding agent from visible behaviour where hidden decision-relevant activity exists.

White Paper WP01, *Negligence by Design*, Three Primitives Research Lab. Published: July 2026. Establishes that deploying automated decision systems without declared authority creates liability and names the structural alternative.

White Paper WP02, *The Second Line of Defence*, Three Primitives Research Lab. Published: July 2026. Measures the foundational assumption every safety product ships and reports the value: zero.

Gurnee, Sofroniew et al., Anthropic, July 2026. J-Space interpretability result, transformer-circuits.pub. Identifies a specific internal channel whose ablation flattens experiential reports while preserving output coherence.

Three Primitives Research Lab · Melbourne, Australia
3Primitives.io