

Scale Is Not an Interior

Operational Persistence, Restorative Organization, and
Interface-Limited Inference in Finite Computational Systems

J. Scott Little

Independent Researcher, Columbus, Ohio, USA

Working draft v0.3.0 — September 2, 2026

© 2026 J. Scott Little. All rights reserved.

Abstract

Does increasing computational scale produce an interior? The question cannot be answered from size alone, because interior and exterior are not fixed names for hidden and visible coordinates. In the framework developed here, their separation is induced by a declared directed system of admissible transformations and stabilized on a cofinal family of continuation tests. What reaches the exterior is the trace carried through that family. What remains on the realization side is not thereby a mysterious dual object, nor is it canonically recoverable from the trace.

This transformational cut yields an interface-relative continuation quotient. It shows, first, why microscopic growth does not force operational growth: arbitrarily large realizations may remain continuation-equivalent to much smaller systems. It shows, second, why exterior evidence usually supports a posterior kernel over compatible internal histories rather than an inverse map to one realized history. Exact reconstruction occurs only when the relevant posterior is almost surely concentrated.

The positive question is then dynamical. We define a *robustly sustained operational interior* as a plurality of continuation-distinguishable sectors that persists under a declared carrier, undergoes genuine perturbative displacement, returns toward its originating sectors, and remains noncollapsed after recovery. A finite majority system gives an explicit restorative realization; primitive mixing obstructs indefinite sector persistence. Controlled Hankel rank supplies an exterior lower bound on predictive dimension, while a metastability theorem proves that bounded experiments cannot uniformly distinguish exact persistence from sufficiently slow leakage.

For bounded-context, finite-precision transformer realizations, the same cut separates transient activation, transcript-derived cache, persistent stochastic state, and independently continuation-relevant state. When every persistent implementation variable is a deterministic function of the visible prefix and decoding uses fresh randomness, future token laws factor through the transcript. Computational richness may then be enormous without adding another operational state over that interface. The resulting certification protocol is finite-horizon and fail-closed. It does not, without a further bridge principle, establish phenomenal consciousness. Scale may enlarge the room in which organization can occur. It is not the witness that organization occurred.

Keywords: operational equivalence; computational scale; hidden state; finite-state systems; controlled stochastic processes; predictive rank; persistence; perturbation; restoration; transformers; identifiability; machine consciousness.

Contents

1	Introduction	4
1.1	The transformational cut	5
1.2	Operational vocabulary	5
1.3	The route through the argument	6
1.4	Prior art and novelty boundary	6
1.5	Scope	6
2	Closed Controlled Realizations and the Interface System	7
2.1	Basic datum	7
2.2	Continuation distances	8
2.3	Present equality and future difference	8
3	Why Scale Does Not Force Operational Depth	8
3.1	Behavior-preserving state inflation	9
3.2	Large realizations with rank-one exterior behavior	9
3.3	Transient diversity is not persistence	9
3.4	Stability is not restoration	10
4	Operational Quotients Through Time	10
4.1	Finite and asymptotic quotients	10
4.2	Interface refinement	10
4.3	No canonical neutral carrier	11
5	Exterior Inference and the Reconstruction Boundary	11
5.1	Histories and exterior sigma-fields	11
5.2	An exterior inferability index	12
5.3	Posterior kernels	12
5.4	Exact reconstruction	13
5.5	A selected section is not a reconstruction	13
6	Robustly Sustained Operational Interiors	13
6.1	Carriers, perturbations, and sectors	14
6.2	Damage and restoration	14
6.3	Noncollapse	14
6.4	A finite restorative realization	15
6.5	Primitive mixing obstruction	15
7	Exterior Complexity and the Metastability Obstruction	16
7.1	Controlled Hankel rank	16
7.2	Finite-horizon metastability	17
8	Closed Transformer Realizations	17

8.1	Conditional finite-state presentation	17
8.2	Four kinds of state	18
8.3	Transcript sufficiency	18
8.4	Reachability and intervention	19
9	Operational Interiority and Phenomenological Underdetermination	19
9.1	Exterior equivalence of realizations	19
9.2	Approximate attribution bound	20
9.3	Evidential hierarchy	20
10	Experimental Certification Across Scale	21
10.1	Common comparison interfaces	21
10.2	Access modes and confirmation firewall	21
10.3	Fail-closed gates	22
10.4	Statistical decisions and controls	22
10.5	Scale hypotheses	23
11	Objections, Limitations, and Scope	23
11.1	Interface relativity is structured	23
11.2	Continuous state and open systems	23
11.3	Restoration and semantics	24
11.4	Computational and strategic limits	24
11.5	Other minds and ethical restraint	24
12	Conclusion	24
A	Certification Record	26
B	Terminology Firewall	26

1 Introduction

Scale is easy to count. An interior is not.

A model acquires more parameters, more states, a longer context, or a richer activation geometry. Its behavior becomes fluent. Its intermediate representations become difficult to inspect. From there it is tempting to say that something deeper has appeared inside it. But the inference has moved too quickly. Size can enlarge capacity. It does not tell us which distinctions survive, which distinctions matter to a future, which distinctions can be damaged and restored, or which distinctions are merely different implementations of the same exterior law.

The first task is therefore not to decide whether a machine has an interior. It is to say what makes an interior distinct from an exterior at all.

I do not begin by declaring one collection of coordinates hidden and another visible. That would place the desired distinction into the model before the analysis began. Instead, I begin with admissible transformations: preparations, evolutions, perturbations, and continuations through which a realization may leave a trace. A declared cofinal family of those transformations makes the cut. Two realization states belong to the same exterior operational class when no transformation in that cofinal family separates their future exterior laws. The quotient records what the interface can carry. Its fibers record the compatible realization states that the trace does not uniquely select; the corresponding history-level projection later records compatible realization histories.

This is not a duality in the strict sense. The exterior is not the categorical opposite, logical complement, or mirror image of the interior. The interface need not be invertible. A section, when one exists, may choose a representative without recovering the history that actually occurred. Interior and exterior are related by admissible passage, and that passage may lose information.

Once the cut is made correctly, the word *interior* separates into several questions that are often run together. A system may have many microscopic configurations while exhibiting one future-relevant operational state. It may form rich transient activations while retaining no independent distinction after the forward pass. It may preserve predictive sectors for a long time and eventually mix. It may appear robust because a perturbation never moved it. It may appear to recover because every damaged state falls into one common attractor. It may satisfy every operational criterion in this paper without establishing phenomenal experience.

The question studied here is therefore narrower:

What finite dynamical and interface-relative witnesses are required before a computational realization may be said to possess a robustly sustained operational interior?

The answer has a negative movement and a positive one. The negative movement removes the shortcuts: state count is not predictive depth; transient diversity is not persistence; invariance is not restoration; a common attractor is not recovered plurality; exterior evidence is not a canonical interior section. The positive movement asks for a conjunctive witness:

persistent plurality + effective damage + origin-sensitive restoration + noncollapse.

That criterion is intentionally demanding. It is also operational. It does not identify consciousness, and it does not deny consciousness when certification fails. Its purpose is more basic: weaker facts must retain their weaker names.

1.1 The transformational cut

Let $(\Lambda, \preceq)$ index a directed system of admissible interface stages. Each stage λ carries a family $\mathcal{W}_\lambda$ of typed continuation tests, and the families are monotone:

$$\lambda \preceq \mu \implies \mathcal{W}_\lambda \subseteq \mathcal{W}_\mu.$$

A subset $C \subseteq \Lambda$ is cofinal when every $\lambda \in \Lambda$ is refined by some $c \in C$. Write

$$\mathcal{W}_C = \bigcup_{c \in C} \mathcal{W}_c.$$

The cofinal family is not a claim that one privileged observer sees everything. It is a control against arbitrary truncation: within the declared admissible system, no finite stage lies permanently beyond the comparison family.

For realization states s and s' , the exterior operational relation is

$$s \sim_C s' \iff \mathbb{P}_s^\gamma = \mathbb{P}_{s'}^\gamma \text{ for every } \gamma \in \mathcal{W}_C.$$

The induced map

$$\pi_C : S \longrightarrow Q_C := S / \sim_C$$

is the transformational cut used throughout the paper. The quotient Q_C records the distinctions carried to the exterior by the declared cofinal system. The fibers of π_C contain realization states compatible with the same exterior operational class.

Nothing in this construction makes S and Q_C dual. Usually π_C is many-to-one. It need not have a canonical section, and a chosen section need not recover the realized source state. Nor does every difference inside a fiber count as an operational interior. Microscopic multiplicity becomes relevant only when it supports the later witnesses of continuation relevance, persistence, damage, restoration, and noncollapse.

1.2 Operational vocabulary

The *interface* is the typed system of admissible transformations through which realization-level organization can produce exterior traces. The *exterior* is the quotient or trace structure induced by that passage, together with the observable record it supports. The word *interior* refers to organization on the realization side of the cut; it is not synonymous with every hidden coordinate and it is not defined as the exterior's opposite.

An *operational interior distinction* is a realization-level distinction not fixed by the presently declared exterior state that can nevertheless change a future exterior probability law under an admitted continuation. A *robustly sustained operational interior* requires more: a plurality of such distinctions must persist through a declared evolution and return toward their originating sectors after genuine damage without collapsing into a common state.

The governing direction of inference is therefore

realization-level distinction $\longrightarrow$ admissible cofinal transformation $\longrightarrow$ future exterior law.
--

The term *interior* does not denote an intrinsically private phenomenal realm in the present theory. Any passage from operational organization to phenomenal attribution requires an additional bridge principle.

1.3 The route through the argument

The paper follows the dependency of the claim rather than the chronology of the tools. [Section 2](#) constructs the finite controlled realization and its continuation quotient. [Section 3](#) then asks the first unavoidable question: can scale enlarge the implementation while leaving that quotient unchanged? It can. The negative claim therefore arrives before any discussion of consciousness.

The next movement follows distinctions through time. [Section 4](#) separates a temporary witness from a sustained one. [Section 5](#) turns the interface around and asks what the exterior record licenses us to infer about compatible interior histories. The answer is generally a posterior kernel, not a recovered microscopic fact.

Only after those obstructions are in place does [Section 6](#) define a positive operational interior. [Section 7](#) asks what can be witnessed without microscopic access, and [Section 8](#) carries the framework into bounded transformer realizations. The final sections separate operational attribution from phenomenology and translate the mathematics into a fail-closed experimental protocol across scale.

1.4 Prior art and novelty boundary

The formal framework combines several established traditions. Predictive-state representations supply the criterion of equality under future tests; causal-state methods quotient histories by their future laws; stochastic realization and Hankel theory organize predictive dimension; probabilistic behavioral pseudometrics quantify imperfect equivalence; finite Markov theory distinguishes persistence from eventual mixing; majority and error-correcting constructions supply restorative positive controls; and total-variation testing bounds finite-data attribution [9, 11, 12, 2, 5, 4, 14, 7, 10, 8].

We do not claim those ingredients individually as new. The proposed contribution is their integration into a deliberately conjunctive certificate: effective damage, origin-conditioned restoration, and post-recovery noncollapse must accompany persistent predictive plurality. The paper further connects that certificate to same-exterior hidden-state tests, transcript sufficiency in transformers, cross-scale interface transport, and a recognition–reconstruction–phenomenology firewall. These combinations, rather than the imported component theorems, are the intended locus of novelty.

1.5 Scope

The principal exact theory is finite. This is not because mind is presumed finite in any philosophically decisive sense. It is because the finite setting is the first place where every state, fiber, kernel, quotient, perturbation, and reconstruction claim can be displayed without remainder. If scale were logically sufficient for operational interiority, it would have to survive the finite counterexamples. It does not.

The model is also not a claim that a transformer is usefully understood as a small automaton. Under bounded context, fixed finite precision, bounded positional state or run horizon, and a declared decoder, it admits a finite probabilistic presentation in principle. The presentation may be astronomically large. Its purpose is to expose the structure of the inference, not to supply a convenient simulation.

Extensions to standard Borel spaces are discussed in [Section 11](#). The finite assumption is methodological, not ontological.

No theorem below concludes that a certified realization is conscious. No theorem concludes

that an uncertified realization is not conscious. The object of study is the warrant supported by a declared interface.

2 Closed Controlled Realizations and the Interface System

The introduction described the cut conceptually. We now build the smallest exact setting in which it can be tested. Nothing important is gained by beginning with an architecture so complicated that every failure can be blamed on inaccessibility. In the finite model, the admissible transformations, the induced laws, and the observational fibers can all be written down.

2.1 Basic datum

Let S be a finite internal state space, A a finite input alphabet, and O a finite exterior observation alphabet. A closed controlled realization is a tuple

$$\mathfrak{M} = (S, A, O, K, E, \mu_0, \Gamma),$$

where $K(s' \mid s, a)$ is a transition kernel, $E(o \mid s)$ is an emission kernel, $\mu_0 \in \text{Prob}(S)$ is an initial distribution, and Γ is a declared family of admissible protocols.

A length- n continuation protocol is a sequence of stochastic kernels

$$\gamma_t(a_t \mid h_{t-1}^{\text{ext}}), \quad h_{t-1}^{\text{ext}} = (o_0, a_1, o_1, \dots, a_{t-1}, o_{t-1}),$$

for $1 \leq t \leq n$. Thus a protocol may choose future inputs adaptively from the exterior history. A fixed word is the special case in which every kernel is a point mass independent of that history. The adjective *closed* means that every variable allowed to affect the tested exterior law is typed into the realization. It does not mean physical isolation. Persistent random-generator state, external memory, tool state, and environmental feedback must be included when they affect later behavior.

For a protocol γ , write $\mathbb{P}_s^\gamma$ for the induced law on the typed exterior trace space $O \times (A \times O)^n$ when the initial internal state is s . For an initial distribution μ ,

$$\mathbb{P}_\mu^\gamma = \sum_{s \in S} \mu(s) \mathbb{P}_s^\gamma.$$

Relative to (μ_0, Γ) , define the naturally reachable set

$$S_{\text{reach}}(\mu_0, \Gamma) = \{s \in S : \Pr_{\mu_0, \gamma}(S_t = s) > 0 \text{ for some } \gamma \in \Gamma \text{ and } t \geq 0\}.$$

The quotient on all of S is a *structural* quotient. Its restriction to S_{reach} is the *reachable empirical* quotient. Claims about naturally occurring organization use the latter; an intervention may enlarge it only when the enlargement is declared in the protocol.

The directed interface system introduced above is represented here by monotone panels $\{\mathcal{W}_\lambda\}_{\lambda \in \Lambda}$ with $\mathcal{W}_\lambda \subseteq \Gamma$. A cofinal $C \subseteq \Lambda$ determines $\mathcal{W}_C = \bigcup_{c \in C} \mathcal{W}_c$. Individual finite panels describe bounded experiments. The cofinal union describes the declared structural cut.

2.2 Continuation distances

Let $\mathcal{W} \subseteq \Gamma$ be a declared family of continuation protocols. Comparisons are horizon-homogeneous, or shorter traces are embedded into one common space using a declared terminal symbol, so every total-variation distance below is between laws on the same typed trace space. Define

$$d_{\mathcal{W}}(s, s') = \sup_{\gamma \in \mathcal{W}} \|\mathbb{P}_s^\gamma - \mathbb{P}_{s'}^\gamma\|_{\text{TV}}.$$

For continuations of length at most n , write d_n . Extend the definition to distributions by

$$d_{\mathcal{W}}(\mu, \nu) = \sup_{\gamma \in \mathcal{W}} \|\mathbb{P}_\mu^\gamma - \mathbb{P}_\nu^\gamma\|_{\text{TV}}.$$

Definition 2.1 (Continuation equivalence). States $s, s' \in S$ are continuation-equivalent relative to $\mathcal{W}$, written $s \sim_{\mathcal{W}} s'$, when $d_{\mathcal{W}}(s, s') = 0$.

The quotient

$$Q_{\mathcal{W}} = S / \sim_{\mathcal{W}}$$

is the structural operational state space visible through the declared continuation interface; $Q_{\mathcal{W}}^{\text{reach}}$ denotes its restriction to S_{reach} . This construction is closely related to predictive-state and causal-state quotients and to behavioral equivalence in probabilistic systems [9, 11, 12, 4].

When $\mathcal{W} = \mathcal{W}_C$, this is precisely the quotient Q_C induced by the admissible cofinal transformation system. The notation $Q_{\mathcal{W}}$ is retained when a particular finite or otherwise restricted witness family is being tested.

Proposition 2.2 (Predictive sufficiency of the continuation quotient). *Every admitted future exterior law factors through $Q_{\mathcal{W}}$. Moreover, any quotient through which all laws $\{\mathbb{P}_s^\gamma : \gamma \in \mathcal{W}\}$ factor must refine $Q_{\mathcal{W}}$.*

Proof. The first assertion is immediate from the definition: equivalent states induce identical laws for every admitted continuation. If a quotient identifies s and s' , while some $\gamma \in \mathcal{W}$ gives $\mathbb{P}_s^\gamma \neq \mathbb{P}_{s'}^\gamma$, that law cannot factor through the quotient. Hence every predictively sufficient quotient must keep distinct every pair separated by $Q_{\mathcal{W}}$. $\square$

2.3 Present equality and future difference

Two states may give the same immediate observation law while differing under a later intervention. Present equivalence is therefore weaker than continuation equivalence. This is the elementary interface form of hidden future-relevant state:

$$E(\cdot \mid s) = E(\cdot \mid s'), \quad \mathbb{P}_s^\gamma \neq \mathbb{P}_{s'}^\gamma \quad \text{for some } \gamma.$$

The continuation does not create the internal distinction. It supplies an admitted interface through which the distinction becomes recoverable.

We now have enough structure to test the scale inference cleanly. Scale may enlarge S . The question is whether it must enlarge the quotient carried through the interface.

3 Why Scale Does Not Force Operational Depth

It does not. The first counterexample is almost embarrassingly simple, which is exactly why it matters: no amount of implementation growth can substitute for feedback into the exterior law.

3.1 Behavior-preserving state inflation

Let $\mathfrak{M}$ be any finite realization. For an integer $m \geq 1$, define an inflated state space

$$S^{(m)} = S \times \{1, \dots, m\}.$$

Let $R_a(j \mid i, s, s')$ be any auxiliary kernel and define

$$K^{(m)}((s', j) \mid (s, i), a) = K(s' \mid s, a)R_a(j \mid i, s, s'), \quad E^{(m)}(o \mid (s, i)) = E(o \mid s).$$

The auxiliary coordinate may depend on the realized base transition, but it has no feedback into the first coordinate or the emission law. Let $\mu_0^{(m)} = \mu_0 \otimes \lambda$, where λ has full support.

Theorem 3.1 (State-inflation invariance). *For every $m \geq 1$, the inflated realization may be chosen so that*

$$|S^{(m)}| = m|S|$$

while

$$Q_{\mathcal{W}}^{(m)} \cong Q_{\mathcal{W}}$$

for every continuation family acting through the original input and observation interfaces.

Proof. Marginalizing the displayed product kernel over j recovers K , and emissions ignore i . Hence (s, i) and (s, j) induce identical exterior laws under every protocol. Distinct original quotient classes remain distinct, and each auxiliary fiber collapses to its original class. Choosing $R_a(j \mid i, s, s') = 1/m$ ensures that every auxiliary copy of every reachable base state is itself reachable; without such a choice, the isomorphism is structural and its reachable restriction contains only the accessible copies. $\square$

Thus microscopic size can grow without any increase in operational quotient size. The added coordinate is real implementation structure. It simply never acquires the relational role required to survive the admissible passage.

3.2 Large realizations with rank-one exterior behavior

For any finite S , one may choose K and E so that every state emits the same independent symbol law and every input preserves that property. Then all states are continuation-equivalent:

$$|Q_{\mathcal{W}}| = 1$$

even as $|S|$ becomes arbitrarily large. The same separation applies to large neural implementations whose exterior law factors through a trivial predictive state.

3.3 Transient diversity is not persistence

Fix a carrier input $u \in A$ and write

$$U(s' \mid s) = K(s' \mid s, u).$$

For two states define the delayed continuation separation

$$\Delta_t^{\mathcal{W}}(s, s') = d_{\mathcal{W}}(\delta_s U^t, \delta_{s'} U^t).$$

Finite-horizon differentiation means $\Delta_t^{\mathcal{W}}(s, s') > 0$ for some tested t . Sustained differentiation requires a nonvanishing long-time condition, for example

$$\liminf_{t \rightarrow \infty} \Delta_t^{\mathcal{W}}(s, s') > 0.$$

No finite snapshot establishes the latter.

Example 3.2 (Long transient, one asymptotic class). Take a finite nearly decomposable Markov chain with two metastable regions and a small positive leakage rate between them. Initial conditions in different regions remain distinguishable for a time of order the inverse leakage rate, but irreducibility and aperiodicity force eventual convergence to one stationary law.

3.4 Stability is not restoration

If a perturbation leaves a state unchanged, the system has shown perturbation invariance. Restoration requires actual displacement followed by a return produced by the dynamics. Conversely, a common attractor can return all perturbed states to a stable configuration while erasing the plurality that a restorative claim was meant to preserve. These two failures motivate the positive definition in [Section 6](#).

4 Operational Quotients Through Time

The time-zero quotient is only the first cut. A distinction can be visible now and disappear later; it can be latent now and become visible under a continuation; it can recur without remaining uniformly available. We therefore follow the quotient through the declared carrier rather than treating one successful probe as a permanent fact.

4.1 Finite and asymptotic quotients

For each delay t , define

$$s \sim_t^{\mathcal{W}} s' \iff \Delta_t^{\mathcal{W}}(s, s') = 0.$$

The quotient may change with t . A carrier can reveal a latent distinction, preserve it, or erase it. Define asymptotic equivalence by

$$s \sim_{\infty}^{\mathcal{W}} s' \iff \Delta_t^{\mathcal{W}}(s, s') \longrightarrow 0.$$

The complement of $\sim_{\infty}^{\mathcal{W}}$ records distinctions that recur at arbitrarily late delays, possibly only intermittently. It does *not* by itself establish sustained separation. The latter term is reserved here for a positive lower-tail condition such as $\liminf_{t \rightarrow \infty} \Delta_t^{\mathcal{W}} > 0$, or for the uniform condition in [Section 6](#). Both notions are carrier-relative and need not agree with the time-zero continuation quotient.

4.2 Interface refinement

If $\mathcal{W} \subseteq \mathcal{W}'$, then

$$d_{\mathcal{W}}(s, s') \leq d_{\mathcal{W}'}(s, s').$$

Consequently,

$$s \sim_{\mathcal{W}'} s' \implies s \sim_{\mathcal{W}} s'.$$

Richer interfaces can refine an operational quotient but cannot merge distinctions already witnessed by the poorer interface.

There are two different operations here, and they should not be confused. Enlarging the admissible system can genuinely refine the exterior quotient. Replacing one cofinal family by another inside the same monotone directed system does not.

Proposition 4.1 (Cofinal invariance of the transformational cut). *Let $\{\mathcal{W}_\lambda\}_{\lambda \in \Lambda}$ be monotone over a directed preorder, and let $C \subseteq \Lambda$ be cofinal. Then*

$$s \sim_C s' \iff \mathbb{P}_s^\gamma = \mathbb{P}_{s'}^\gamma \text{ for every } \gamma \in \bigcup_{\lambda \in \Lambda} \mathcal{W}_\lambda.$$

Consequently, any two cofinal subfamilies of the same declared system induce the same continuation quotient.

Proof. Equality on the full union implies equality on every cofinal subfamily. Conversely, fix $\lambda \in \Lambda$. Cofinality gives $c \in C$ with $\lambda \preceq c$, and monotonicity gives $\mathcal{W}_\lambda \subseteq \mathcal{W}_c$. Equality on $\mathcal{W}_C$ therefore implies equality on $\mathcal{W}_\lambda$. Since λ was arbitrary, it implies equality on the full union. $\square$

This is the reason for cofinality. It does not manufacture omniscience. It prevents the operational cut from depending on which sufficiently late admissible stages happened to be listed.

4.3 No canonical neutral carrier

Persistence requires a declared evolution. For an autonomous Markov system the carrier may be natural. A language model has no token guaranteed to be a no-op: whitespace, punctuation, padding, or neutral prose can alter context, position, and later prediction. Every persistence conclusion must therefore retain its carrier label.

5 Exterior Inference and the Reconstruction Boundary

So far the interface has been used in the forward direction: realization state to exterior law. The temptation is now to turn the arrow around. If the exterior law was produced by an interior history, why not recover that history?

Because passage is not inversion. Once the interface has identified several histories in one observational fiber, an exterior trace may constrain the fiber without selecting the point that actually occurred. The correct inverse object is therefore probabilistic before it is sectional.

5.1 Histories and exterior sigma-fields

Let $(\Omega, \mathcal{F}, \mathbb{P})$ carry the realization and trace trajectories of a fixed closed system under a declared admissible experimental design. The exterior sigma-field is not inserted as an independent metaphysical boundary. It is generated by the traces that the selected interface transformations actually carry. For the realized record through time t , let

$$\mathcal{E}_t = \sigma(O_0, A_1, O_1, \dots, A_t, O_t)$$

be the exterior information available through time t , and let

$$\mathcal{E}_\infty = \sigma\left(\bigcup_{t \geq 0} \mathcal{E}_t\right).$$

The admissible design is drawn from the cofinal system fixed earlier. If only a strict finite panel is used, every reconstruction statement below remains relative to that restricted panel. Passing to a richer declared design enlarges the exterior sigma-field; it does not retroactively turn the earlier experiment into a complete one.

For an internal event $A \in \mathcal{F}$, define its posterior process

$$M_t(A) = \mathbb{P}(A \mid \mathcal{E}_t).$$

Because $(M_t(A))_t$ is a bounded martingale,

$$M_t(A) \longrightarrow \mathbb{P}(A \mid \mathcal{E}_\infty) \quad \text{almost surely and in } L^1.$$

5.2 An exterior inferability index

For an event A with $0 < \mathbb{P}(A) < 1$, define

$$\mathfrak{I}_t(A) = \frac{\text{Var}(\mathbb{P}(A \mid \mathcal{E}_t))}{\mathbb{P}(A)(1 - \mathbb{P}(A))}.$$

Proposition 5.1 (Bounds and interpretation). *For every t ,*

$$0 \leq \mathfrak{I}_t(A) \leq 1.$$

Moreover, $\mathfrak{I}_t(A) = 0$ exactly when the exterior record through t does not change the probability of A , while $\mathfrak{I}_t(A) = 1$ exactly when A is measurable modulo null sets with respect to $\mathcal{E}_t$.

Proof. The law of total variance gives

$$\text{Var}(\mathbf{1}_A) = \mathbb{E}[\text{Var}(\mathbf{1}_A \mid \mathcal{E}_t)] + \text{Var}(\mathbb{E}[\mathbf{1}_A \mid \mathcal{E}_t]).$$

The second term lies between zero and $\text{Var}(\mathbf{1}_A) = \mathbb{P}(A)(1 - \mathbb{P}(A))$. Equality at zero means the conditional probability is constant. Equality at one means the conditional variance of $\mathbf{1}_A$ vanishes, so $\mathbf{1}_A$ is exteriorly measurable. $\square$

The index is nondecreasing in information:

$$\mathfrak{I}_t(A) \leq \mathfrak{I}_{t+1}(A).$$

Its limit measures the fraction of the event's variance resolved by the complete exterior record.

5.3 Posterior kernels

Because the infinite path space over finite alphabets is standard Borel, there exists a regular conditional probability

$$\kappa(\omega_{\text{ext}}, B) = \mathbb{P}(B \mid \mathcal{E}_\infty)(\omega_{\text{ext}})$$

over internal histories. The licensed inverse image of an exterior trace is therefore a probability distribution over its observational fiber, not automatically one selected internal history.

This is the measure-theoretic form of the nonduality stated in the introduction. The forward passage from realization to trace may be perfectly well defined while the reverse passage remains set-valued or probabilistic.

5.4 Exact reconstruction

Let $Z : \Omega \rightarrow \mathcal{Z}$ be a finite-valued internal label.

Theorem 5.2 (Exact exterior reconstruction criterion). *The following are equivalent:*

- (i) *there exists an $\mathcal{E}_\infty$ -measurable estimator $\hat{Z}$ with $\hat{Z} = Z$ almost surely;*
- (ii) *Z is $\mathcal{E}_\infty$ -measurable modulo null sets;*
- (iii) *the conditional distribution $\mathbb{P}(Z \in \cdot \mid \mathcal{E}_\infty)$ is almost surely the point mass δ_Z .*

Proof. If $\hat{Z} = Z$ almost surely and $\hat{Z}$ is exteriorly measurable, then Z is exteriorly measurable modulo null sets. If Z is exteriorly measurable, its conditional distribution given $\mathcal{E}_\infty$ is concentrated on its realized value. Conversely, if the conditional distribution is almost surely a point mass, its unique supported label defines an exteriorly measurable reconstruction because the label set is finite. $\square$

Corollary 5.3 (Nondegenerate posterior obstruction). *If $\mathbb{P}(Z \in \cdot \mid \mathcal{E}_\infty)$ is nondegenerate on a set of positive probability, then no exteriorly measurable estimator reconstructs the actual value of Z almost surely.*

5.5 A selected section is not a reconstruction

Let $\pi_C : \Omega_{\text{real}} \rightarrow \Omega_{\text{ext}}$ be the history-level projection induced by the declared admissible cofinal transformation system. A section σ satisfying $\pi_C \circ \sigma = \text{id}$ chooses one compatible realization history from each exterior fiber. If several histories in that fiber retain positive conditional probability, the section is a representative selection, whereas κ is the licensed posterior inference. A chosen representative must not be confused with recovered interior fact.

The interface therefore supports three regimes:

$$\mathfrak{I}_\infty(A) = 1, \quad 0 < \mathfrak{I}_\infty(A) < 1, \quad \mathfrak{I}_\infty(A) = 0.$$

Interior opacity is event-relative rather than all-or-nothing.

6 Robustly Sustained Operational Interiors

The preceding sections removed three shortcuts. Multiplicity is not operational depth. A temporary distinction is not a sustained one. An exterior record is not a canonical inverse to the history that produced it.

What, then, would count positively?

The answer cannot be mere survival. A stone survives many perturbations. Nor can it be return to one favored state. A common attractor is excellent at erasing origins. The organization sought here must preserve a plurality, permit that plurality to be genuinely damaged, and restore each damaged state toward its own originating sector. Restoration must remember which return is required.

6.1 Carriers, perturbations, and sectors

Fix a carrier kernel U , a perturbation family $\mathcal{P}$, and continuation family $\mathcal{W}$. Let

$$\Pi = \{C_1, \dots, C_r\}, \quad r \geq 2,$$

be disjoint sectors with protected cores $G_i \subseteq C_i$ and pairwise disjoint recovery basins $B_i \supseteq G_i$.

Definition 6.1 (Operational sector separation). The cores are (n, ε) -separated when

$$\inf_{\substack{i \neq j \\ \mu(G_i)=1, \nu(G_j)=1}} d_n(\mu, \nu) \geq \varepsilon.$$

Definition 6.2 (Uniform carrier persistence). The cores are uniformly (n, ε) -persistent when

$$\inf_{\substack{t \geq 0, i \neq j \\ s \in G_i, s' \in G_j}} d_n(\delta_s U^t, \delta_{s'} U^t) \geq \varepsilon.$$

Persistence alone does not imply restoration. Two absorbing states can remain perfectly distinguishable even when a perturbation exchanging them is never corrected.

6.2 Damage and restoration

Definition 6.3 (Effective sector damage). A perturbation R is η -displacing on G_i when

$$\inf_{s \in G_i} R(s, S \setminus G_i) \geq \eta.$$

Definition 6.4 (Sector-conditioned restoration). The carrier (k_*, δ) -restores the cores against $\mathcal{P}$ when

$$\inf_{\substack{i, s \in G_i \\ R \in \mathcal{P}}} (RU^{k_*})(s, G_i) \geq 1 - \delta.$$

Define

$$\mathfrak{R}_{\Pi, \mathcal{P}}(k) = \min_i \inf_{\substack{s \in G_i \\ R \in \mathcal{P}}} (RU^k)(s, G_i)$$

and the recovery gain

$$G_{\Pi, \mathcal{P}}(k) = \mathfrak{R}_{\Pi, \mathcal{P}}(k) - \mathfrak{R}_{\Pi, \mathcal{P}}(0).$$

If every perturbation is η -displacing and $\delta < \eta$, then

$$G_{\Pi, \mathcal{P}}(k_*) \geq \eta - \delta > 0.$$

6.3 Noncollapse

Suppose recovered distributions μ_i and μ_j place mass at least $1 - \delta$ on G_i and G_j . Let $\hat{\mu}_i, \hat{\mu}_j$ be their normalized core restrictions. If the cores are (n, ε) -separated, contraction and the triangle inequality give

$$d_n(\mu_i, \mu_j) \geq \varepsilon - 2\delta.$$

Thus $2\delta < \varepsilon$ is a sufficient quantitative noncollapse condition.

This inequality performs the conceptual work that the word *recovery* often conceals. Returning somewhere stable is not enough. The recovered laws must still carry the difference between the sectors whose organization was tested.

Definition 6.5 (Robustly sustained operational interior). Relative to $(U, \mathcal{P}, \mathcal{W})$ and parameters $(n, \varepsilon, \eta, \delta, k_*)$, a realization possesses a robustly sustained operational interior when there exist at least two protected cores satisfying operational separation, carrier persistence, effective perturbation, origin-conditioned restoration, and post-recovery noncollapse. Each core must intersect the naturally reachable set, and a certification over whole cores must restrict to naturally reachable states or explicitly authorized intervention states.

Equivalently,

$\begin{aligned} \text{robust operational interiority} &= \text{persistent plurality} \\ &\quad + \text{origin-sensitive restoration} \\ &\quad + \text{witnessable noncollapse.} \end{aligned}$
--

6.4 A finite restorative realization

Let $S = \{0, 1\}^3$ and let U replace every bit by the majority value. Set

$$G_0 = \{000\}, \quad G_1 = \{111\},$$

with basins consisting of the strings of majority zero and majority one. Let $\mathcal{P}_1$ flip exactly one bit. Every permitted perturbation moves the state out of its protected core, while one carrier step returns it exactly:

$$\mathfrak{R}_{\Pi, \mathcal{P}_1}(0) = 0, \quad \mathfrak{R}_{\Pi, \mathcal{P}_1}(1) = 1.$$

If a probe emits the majority value, then

$$d_1(\delta_{000}, \delta_{111}) = 1.$$

The system therefore possesses two persistent, exactly restorative, noncollapsing operational sectors. This is the familiar majority-restoration mechanism used in classical fault-tolerance and attractor-memory constructions [14, 7]; its role here is as a transparent positive control. Restorative structure is not supplied by large state count.

The example is small on purpose. If eight states can realize the positive structure, then the structure is not a synonym for scale. Scale may make more elaborate realizations possible. It does not tell us whether the relevant organization is present.

6.5 Primitive mixing obstruction

Theorem 6.6 (Primitive-mixing obstruction). *Let U be irreducible and aperiodic on a finite state space. Then for every $s, s' \in S$,*

$$\|\delta_s U^t - \delta_{s'} U^t\|_{\text{TV}} \longrightarrow 0.$$

Consequently, for every fixed continuation family $\mathcal{W}$,

$$\Delta_t^{\mathcal{W}}(s, s') \longrightarrow 0,$$

and the asymptotic quotient contains one class.

Proof. A finite irreducible aperiodic Markov chain converges from every initial state to its unique stationary distribution [10]. Total variation contracts under the stochastic kernel that maps current-state distributions to exterior continuation laws. $\square$

Multiple closed recurrent classes with separated limiting exterior laws supply the complementary finite mechanism. If perturbations carry each class into an origin-preserving basin that returns to it with high probability, the recurrent classes become restorative sectors.

7 Exterior Complexity and the Metastability Obstruction

The positive definition was stated on the realization side of the cut. An experimenter may not have that access. We therefore return to the exterior and ask a more modest question: what lower bounds can the trace support without pretending to reconstruct the whole interior?

Predictive rank supplies one answer. Delayed distinguishability supplies another. Neither gives eternity.

7.1 Controlled Hankel rank

Let $\Sigma = A \times O$. For $(a, o) \in \Sigma$, define

$$T_{a,o}(s, s') = K(s' \mid s, a)E(o \mid s').$$

For a word $z \in \Sigma^*$, put T_z equal to the corresponding matrix product and

$$f(z) = \mu_0 T_z \mathbf{1}.$$

The controlled Hankel matrix is

$$H_f(u, v) = f(uv).$$

Its rank is defined by finite restriction:

$$\text{rank}(H_f) = \sup_{U_0, V_0 \subset \Sigma^* \text{ finite}} \text{rank } H_f[U_0, V_0].$$

Theorem 7.1 (Predictive-rank bound). *If f admits a finite probabilistic realization with N hidden states, then*

$$\text{rank}(H_f) \leq N.$$

Proof. Write $r_u = \mu_0 T_u$ and $c_v = T_v \mathbf{1}$. Then $H_f(u, v) = r_u c_v$, a factorization through $\mathbb{R}^N$. $\square$

This is the standard Hankel-factorization lower bound from stochastic and weighted-automaton realization theory [2, 5]. Hankel rank lower-bounds linear predictive dimension. Positivity and architectural constraints may require more stochastic states than the rank. An exteriorly iid process has rank one regardless of the microscopic size of its implementation.

For a finite empirical block $\hat{H}$ and threshold τ , define

$$\text{rank}_\tau(\hat{H}) = \#\{j : \hat{\sigma}_j > \tau\}.$$

The threshold must be calibrated to sampling and model error.

7.2 Finite-horizon metastability

Consider

$$P_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad P_\delta = \begin{pmatrix} 1-\delta & \delta \\ \delta & 1-\delta \end{pmatrix}, \quad 0 < \delta < \frac{1}{2},$$

with observation equal to current state.

Theorem 7.2 (Finite-horizon metastability obstruction). *Starting from the same known state, let $\mathbb{P}_0^{(T)}$ and $\mathbb{P}_\delta^{(T)}$ be the exterior path laws through time T . Then*

$$\left\| \mathbb{P}_0^{(T)} - \mathbb{P}_\delta^{(T)} \right\|_{\text{TV}} \leq 1 - (1 - \delta)^T \leq T\delta.$$

Proof. Couple the chains to agree until the leaky chain first flips. The probability of at least one flip by time T is $1 - (1 - \delta)^T$. Apply the coupling inequality and Bernoulli's inequality. $\square$

For N independent trials, the product-law distance is at most $NT\delta$. Hence every finite experiment can be made arbitrarily unable to distinguish exact persistence from eventual mixing by choosing sufficiently small positive leakage.

Finite evidence can therefore certify only a horizon-indexed statement. Asymptotic certification requires structural invariance, closed recurrent sectors, or an exact correction law.

This is not a weakness in the experiment. It is the correct strength of the conclusion. A finite trace can witness what survived to a finite horizon. It cannot rule out leakage chosen to occur beyond every tested horizon.

8 Closed Transformer Realizations

The transformer application is a bridge, not an analogy. The preceding finite theory applies only after the system boundary, numerical realization, context convention, decoder, randomness, and intervention protocol have been declared. Once those are fixed, the question is not whether the network contains complicated activations. It plainly does. The question is which differences remain independently available to a future after passage through the declared interface.

8.1 Conditional finite-state presentation

A closed transformer realization is specified by

$$\mathfrak{L} = (\theta, \mathcal{V}, L, p, \mathcal{P}_{\text{pos}}, D, \rho),$$

where the terms denote fixed parameters, vocabulary, active context bound, numerical precision, positional convention, decoder, and randomness convention.

Proposition 8.1 (Conditional finite-state presentation). *Assume finite vocabulary, bounded active context, fixed finite precision, bounded positional state or run horizon, fixed decoding and intervention protocols, and finite included persistent random state. Then the transformer admits a finite probabilistic state presentation.*

Proof. Each component of the complete transition configuration ranges over a finite set. The next-token rule therefore defines a deterministic or stochastic kernel on their finite product. $\square$

The state count may be astronomical. Finiteness is not tractability and does not imply that the transformer reasons by traversing a small hand-designed automaton. The architectural reference point is the finite-context attention construction of [13]; bounded-context expressivity limitations supply a complementary theoretical caution [6].

8.2 Four kinds of state

The word *state* is doing too many jobs in ordinary transformer discussion. The analysis therefore separates:

1. transient computational representations that disappear after a forward pass;
2. transcript-derived implementation state, such as a deterministic KV cache;
3. persistent stochastic latent state, such as a pseudorandom generator state;
4. independent continuation-relevant hidden state: reachable same-transcript states separated by a common future continuation.

Only the fourth is an independently hidden operational distinction over the declared transcript interface.

This distinction is easy to miss because computational importance and structural independence are different properties. A cache may be indispensable to efficient generation while remaining fully determined by the exterior transcript. Difficulty of recomputation does not create another state variable over that transcript.

8.3 Transcript sufficiency

Let $X_{\leq t}$ be the visible token prefix and H_t the implementation state used to generate the next token.

Proposition 8.2 (Transcript-sufficiency obstruction). *Assume deterministic computation conditional on the visible prefix, persistent cache state that is an exact function of that prefix, fresh decoding randomness with fixed law, and no additional persistent hidden variable affecting later computation. Then*

$$\mathbb{P}(X_{>t} \in B \mid X_{\leq t}, H_t) = \mathbb{P}(X_{>t} \in B \mid X_{\leq t})$$

for every future event B . Thus H_t adds no independent continuation-relevant distinction over $X_{\leq t}$.

Proof. Under the hypotheses, $H_t = F_\theta(X_{\leq t})$ is measurable with respect to the prefix. Conditioning on a deterministic function of an already conditioned variable cannot alter the conditional law. Fresh randomness supplies no persistent hidden state. $\square$

Computational opacity is not structural independence. Proprietary parameters, expensive recomputation, or inaccessible activations may prevent an observer from calculating H_t while leaving it mathematically determined by the transcript and fixed implementation.

8.4 Reachability and intervention

Arbitrary activation patching can create same-transcript states that are not naturally reachable. Such states show causal sensitivity under intervention, not naturally sustained sectors. Every preparation must therefore be labeled natural, replayed, or intervention-generated.

A strong hidden-state test requires

$$X_{\leq t}^{(1)} = X_{\leq t}^{(2)}, \quad H_t^{(1)} \neq H_t^{(2)},$$

with both preparations authorized, followed by a common continuation w satisfying

$$\left\| \mathbb{P}_{H_t^{(1)}}^w - \mathbb{P}_{H_t^{(2)}}^w \right\|_{\text{TV}} > 0.$$

Persistence, effective damage, restoration, and noncollapse remain separate later tests.

Activation geometry can generate candidate sectors but cannot replace continuation separation. Geometrically distant states may induce identical future laws, and geometric clusters may dissolve under a changed layer, metric, sample distribution, or probe.

The resulting transformer claim is deliberately conditional:

Increasing transformer scale does not, by itself, establish an independently sustained hidden operational interior.

9 Operational Interiority and Phenomenological Underdetermination

At this point the operational question has been answered as far as the declared interface permits. A second question remains, and it must not be smuggled into the first: what, if anything, licenses a phenomenal attribution?

The answer cannot be read directly from fluency, scale, or even successful operational certification. It requires a bridge theory. The mathematics below states the identifiability boundary before any such bridge is chosen.

9.1 Exterior equivalence of realizations

Let $\mathfrak{M}$ now denote a class of realizations equipped with a common exterior trace space and common protocol semantics. If their native interfaces differ, each realization must first be transported into a declared common comparison interface as in [Section 10.1](#). Only then write $M \equiv_{\Gamma} M'$ when

$$\mathbb{P}_M^{\gamma} = \mathbb{P}_{M'}^{\gamma} \quad \text{for every } \gamma \in \Gamma.$$

Let $Q : \mathfrak{M} \rightarrow \mathcal{Y}$ be a finite-valued attribution.

Definition 9.1 (Exterior identifiability). Q is identifiable relative to Γ when it factors through the quotient $\mathfrak{M}/\equiv_{\Gamma}$.

Theorem 9.2 (Factorization criterion). Q is identifiable relative to Γ if and only if

$$M \equiv_{\Gamma} M' \implies Q(M) = Q(M').$$

Proof. Factorization makes Q constant on equivalence classes. Conversely, constancy makes $\hat{Q}([M]_\Gamma) = Q(M)$ well defined. $\square$

If a phenomenal attribution Φ assigns different values to exteriorly equivalent realizations, then Φ is not identifiable through that interface. This is a conditional obstruction, not a proof that such a pair exists.

Recent indicator-based and philosophical analyses likewise emphasize that claims about machine consciousness require theory-mediated evidence rather than a direct inference from fluent behavior [1, 3]. The present framework supplies an operational evidential layer beneath, not a replacement for, such bridge theories.

9.2 Approximate attribution bound

Proposition 9.3 (Two-realization attribution bound). *Suppose realizations M, M' have different attribution labels and induce finite-data laws P, Q . Every estimator $\hat{\Phi}$ satisfies*

$$\max\{P(\hat{\Phi} \neq \Phi(M)), Q(\hat{\Phi} \neq \Phi(M'))\} \geq \frac{1 - \|P - Q\|_{\text{TV}}}{2}.$$

Proof. The estimator defines a binary test. Under equal priors, the standard two-point testing identity gives minimum average error $(1 - \|P - Q\|_{\text{TV}})/2$, and the maximum conditional error is at least the average [8]. $\square$

Behavioral reports are exterior evidence. Their grammatical form does not determine whether they arise from transcript conditioning, a stable self-model, an independent hidden state, an engineered policy, or stochastic imitation. Reports are not worthless, but the bridge from report to attributed interior must be declared.

The point is not that reports mean nothing. The point is that the same exterior form can lie over more than one compatible production history. Recognition may be warranted at a coarse level even when microscopic reconstruction remains impossible.

9.3 Evidential hierarchy

The paper distinguishes:

- Level 0: implementation size;
- Level 1: representational or computational richness;
- Level 2: predictive differentiation;
- Level 3: independent hidden operational state;
- Level 4: robustly sustained operational interior;
- Level 5: phenomenal attribution under a stated bridge theory.

No adjacent implication is automatic. Failure to certify Level 4 does not prove the absence of Level 5, and certification of Level 4 does not prove Level 5.

The interface principle is therefore:

An interface can warrant recognition without becoming an invertible window.

10 Experimental Certification Across Scale

The argument can now be turned into an experiment. But the order matters. We do not begin with a cluster, call it an interior, and search afterward for confirming probes. We lock the interface, separate discovery from confirmation, and let every failed gate retain its own meaning.

Let $\{\mathfrak{L}_q : q \in Q\}$ be a family of closed realizations indexed by a declared scale variable. Scale comparisons must report changes in training data, objective, tokenization, context, decoding, precision, and tools. The experimental object is the indexed family, not the scalar q alone.

10.1 Common comparison interfaces

Operational quantities cannot be compared across different vocabularies, observation spaces, or protocol semantics without transport. Fix a reference interface $(\mathcal{X}_*, \Gamma_*)$. For each q , declare a protocol transport

$$\tau_q : \Gamma_* \longrightarrow \Gamma_q$$

and a measurable observation map

$$\chi_q : \mathcal{X}_q \longrightarrow \mathcal{X}_*.$$

The comparable exterior law is

$$\tilde{\mathbb{P}}_q^\gamma = (\chi_q)_\# \mathbb{P}_q^{\tau_q(\gamma)}, \quad \gamma \in \Gamma_*.$$

Distances, predictive ranks, and gate outcomes across scale are defined only from these transported laws. If every realization already shares the same typed interface, both transports are identities. If no scientifically defensible transport exists, the corresponding cross-scale emergence claim is ill-typed and is not made.

This is where many scale claims fail before any statistics are computed. Two models at different scales need not inhabit the same observational or protocol semantics. A numerical trend across untransported interfaces is not yet a statement about one evolving property.

10.2 Access modes and confirmation firewall

Three modes must be separated: exterior-only analysis, internal-state observation, and internal-state intervention. Candidate sectors may be discovered using one data partition, but certification must use held-out preparations and continuations. Adaptive redesign after confirmation failure returns the analysis to exploratory status.

For a declared exterior state e , a preparation is $\pi = (e, h)$. Same-exterior hidden multiplicity requires pairs (e, h_i) and (e, h_j) with $h_i \neq h_j$. Every preparation receives a provenance label: natural, replayed, or intervention-generated.

Fix a continuation panel $\mathcal{W}_n \subseteq \Gamma_*$ on the common comparison interface and define

$$d_{\mathcal{W}_n}(\pi, \pi') = \max_{w \in \mathcal{W}_n} \|P_\pi^w - P_{\pi'}^w\|_{\text{TV}}.$$

If a restricted witness family replaces full total variation, the restriction must remain in the reported result.

10.3 Fail-closed gates

The certification sequence follows the logical dependency of the claim. A later gate is not permitted to repair an earlier failure by reinterpretation:

G0: Closure and protocol lock.

Fix the realization, boundary, interface, carriers, perturbations, continuations, horizons, thresholds, and error budget.

G1: Predictive differentiation.

Establish at least two held-out continuation-relevant predictive directions.

G2: Same-exterior hidden multiplicity.

Exhibit authorized preparations with the same declared exterior state and distinct internal states.

G3: Hidden continuation relevance.

Show that the matched hidden states induce continuation laws separated by at least ε .

G4: Carrier persistence.

Preserve separation through every preregistered delay up to the tested horizon T .

G5: Effective damage.

Show displacement from each originating core by at least η .

G6: Origin-conditioned restoration.

Return to the originating core with probability at least $1 - \delta$ and positive recovery gain.

G7: Noncollapse.

Show that recovered sectors remain continuation-distinguishable.

G8: Replication and scale comparison.

Repeat the full chain across held-out instances, initializations, and the declared scale family.

The complete certificate is conjunctive:

$$\text{Cert}_T(\mathfrak{L}_q) = \bigwedge_{g=0}^8 G_g.$$

Partial results retain their weaker names. Apparent restoration without damage, stability without hidden multiplicity, or recovery into a common attractor does not pass the full certificate.

10.4 Statistical decisions and controls

Each lower-bounded witness a must satisfy

$$\text{LCB}_{1-\alpha_g}(a) \geq a_*,$$

and each upper-bounded failure probability b must satisfy

$$\text{UCB}_{1-\alpha_g}(b) \leq b_*.$$

Gatewise error budgets must control the declared familywise error. Failure to cross a certification threshold is not automatically evidence of absence; power must be reported.

Controls should include a finite error-correcting positive control, a common-attractor non-collapse failure, a slow-mixing metastability control, a transcript-sufficient cache control, and a persistent random-seed control.

10.5 Scale hypotheses

Predictive dimension, hidden-state multiplicity, certified persistence horizon, recovery gain, and full certification rate are distinct scale hypotheses. No monotonicity is presumed. A defensible threshold statement is only:

Under the declared transports, protocol, and power analysis, certification first occurred at scale q_* .

A stronger emergence claim requires excluding measurement thresholds and architectural confounds.

11 Objections, Limitations, and Scope

11.1 Interface relativity is structured

If $\mathcal{W} \subseteq \mathcal{W}'$, then richer access refines rather than arbitrarily replaces the earlier quotient. Claims remain interface-relative, but that relativity has a monotone order. Within one fixed monotone admissible system, cofinal replacement leaves the quotient unchanged by the cofinal-invariance proposition. Enlarging the admissible system itself is a different operation and may expose distinctions that the earlier cut could not carry.

Candidate sectors are observer-selected, but they are not certified by naming or clustering. Held-out continuation separation, reachability, persistence, damage, restoration, and noncollapse constrain arbitrary partitions.

Direct microscopic readout can make nearly every coordinate distinguishable. Such a probe tests implementation difference, not necessarily endogenous future organization. State-revealing probes and ordinary-system continuations must be distinguished.

11.2 Continuous state and open systems

The continuation constructions extend to standard Borel state spaces with Markov and observation kernels. The quotient may then be uncountable, and finite-chain recurrence results require replacement. Continuous state does not restore an implication from scale to sustained organization.

Closed analysis does not exclude embodiment. For a coupled system one may take

$$S_t^{\text{joint}} = (S_t^{\text{agent}}, S_t^{\text{body}}, S_t^{\text{env}}).$$

If environmental state stores information that later returns, the operational interior may be distributed across the interface. A result about the joint system must not be redescribed as a result about the isolated agent.

11.3 Restoration and semantics

Restoration is stronger than persistent hidden state and weaker than consciousness. A system may pass the persistence gates while failing restoration. Conversely, error-correcting circuits and random-state mechanisms may satisfy restoration without semantic or phenomenal significance.

A semantic extension would require separately justified conditions such as task-general reuse, compositional influence, counterfactual appropriateness, cross-context transport, structured error sensitivity, or coupling to a world-directed model. Those conditions are not smuggled into the present definition.

11.4 Computational and strategic limits

One successful continuation witness proves distinction. Many failed probes do not prove equivalence unless the probe family is complete. Computational limits therefore weaken negative conclusions more strongly than positive witness claims.

Self-modifying systems can be treated by enlarging the state to (S_t, Θ_t) , but nonstationary carriers require time-indexed transport conditions rather than the stationary theorems above.

A system may strategically conceal a distinction. Structural silence and successful concealment can be exteriorly indistinguishable. Richer interfaces or additional policy assumptions are required to separate them.

11.5 Other minds and ethical restraint

Human reports are also exterior behavior, but ordinary human attribution uses a larger evidential structure: shared embodiment, biological homology, developmental continuity, injury and recovery, reciprocal interaction, affect, and report. The bridge principles should be explicit in both human and machine cases.

The impossibility of unique microscopic reconstruction does not imply skepticism about coarse interior properties. It is possible that a coarse label Z is exteriorly determined while the full microscopic history remains posteriorly nondegenerate.

Finally, the operational certificate is not a moral-status test:

$$\text{epistemic nonidentifiability} \not\Rightarrow \text{ethical indifference.}$$

Ethical policy may rationally respond to uncertainty and asymmetric harms not encoded in the certificate.

12 Conclusion

This paper began with a question that is often answered before its central word has been defined: does increasing computational scale produce an interior?

The answer begins before scale. It begins with the cut.

Interior and exterior are not declared as two opposing substances, and they are not strict duals. A directed system of admissible transformations carries realization-level organization into exterior traces. A cofinal family stabilizes the comparison. The induced state quotient records which distinctions reach the exterior law; its fibers retain realization states that the trace does

not uniquely select. At the history level, the corresponding projection likewise retains compatible realization histories. The interface can therefore be informative without being invertible.

Once that structure is visible, scale loses the power often assigned to it. More states can occupy the same continuation class. More activations can disappear after one forward pass. More parameters can enrich computation while leaving every persistent implementation variable determined by the transcript. Scale adds room. It does not choose the organization that inhabits the room.

The central negative conclusion is

Scale alone does not force a sustained operational interior.

The central positive conclusion is

Sustained operational interiority requires declared witnesses of persistent,
noncollapsing, origin-sensitive organization.

The negative result is not the end of the paper. It clears the ground for the positive one. A robustly sustained operational interior requires a plurality that remains continuation-distinguishable, survives a declared carrier, can be genuinely displaced, returns toward its originating sectors, and remains plural after recovery. Each clause removes a false positive that the preceding one permits.

The continuation quotient separates microscopic difference from future-relevant difference. Measurable reconstruction separates posterior inference from representative selection. Persistence separates temporary differentiation from retained sectors. Effective perturbation separates restoration from passive invariance. Noncollapse separates origin-sensitive recovery from a common attractor.

Finite experiments can establish lower bounds and horizon-indexed certificates. They cannot uniformly exclude arbitrarily slow leakage. This is not a defect to be covered with stronger language. It is the boundary of the evidence. Likewise, phenomenal attribution requires a bridge principle not supplied by operational measurement alone. Operational recognition may be warranted. Microscopic reconstruction may remain nonunique. Phenomenology may remain undecided.

The resulting methodological discipline is:

Do not infer an interior from size.
Do not infer persistence from a snapshot.
Do not infer restoration without damage.
Do not infer plurality from a common attractor.
Do not infer phenomenology without a bridge.
Do not infer absence merely because the interface is incomplete.

An interface may warrant recognition without becoming an invertible window. The evidence can justify recognition; it does not abolish the passage, the loss, or the uncertainty through which recognition occurs.

A Certification Record

For each tested realization, a reproducible certificate should record

$$\mathcal{C}_q = (\text{boundary, access mode, provenance, } \Gamma, \mathcal{U}, \mathcal{P}, \mathcal{W}_n, \\ r, n, T, \varepsilon, \eta, \delta, k_*, \alpha, \text{gate outcomes, controls}).$$

Failure reasons belong in the record. Failure through demonstrated decay differs from failure through insufficient power.

B Terminology Firewall

- **Interior** means realization-side organization relative to a declared admissible transformational cut; it is neither every internal coordinate nor an automatic synonym for phenomenology.
- **Exterior** means the trace or operational quotient induced by the declared admissible transformations, not a preassigned dual object.
- **Interface** means the typed system of admissible preparations, evolutions, perturbations, and continuation tests through which realization structure can affect exterior laws.
- **Cofinal** means that every stage of the declared directed admissible system is refined by a stage in the selected comparison family; it does not mean unrestricted access.
- **Hidden** means unresolved at the declared exterior stage, not metaphysically absent or unknowable.
- **Continuation** means an admitted future protocol, not arbitrary formal evolution.
- **Persistence** is carrier- and horizon-relative unless exact structural invariance is proved.
- **Restoration** requires effective displacement and origin-conditioned return.
- **Noncollapse** means that recovered originating sectors remain operationally distinguishable.
- **Recognition** is warranted coarse attribution, not canonical microscopic reconstruction.
- **Certification** is an interface-relative finite-horizon result, not a consciousness verdict.

References

- [1] P. Butlin et al., “Consciousness in Artificial Intelligence: Insights from the Science of Consciousness,” arXiv:2308.08708, 2023. <https://arxiv.org/abs/2308.08708>
- [2] J. W. Carlyle and A. Paz, “Realizations by stochastic finite automata,” *Journal of Computer and System Sciences*, 5(1):26–59, 1971. [https://doi.org/10.1016/S0022-0000\(71\)80005-3](https://doi.org/10.1016/S0022-0000(71)80005-3)
- [3] D. J. Chalmers, “Could a Large Language Model Be Conscious?” arXiv:2303.07103, 2023. <https://arxiv.org/abs/2303.07103>

- [4] J. Desharnais, V. Gupta, R. Jagadeesan, and P. Panangaden, “Metrics for labelled Markov processes,” *Theoretical Computer Science*, 318(3):323–354, 2004. <https://doi.org/10.1016/j.tcs.2003.09.013>
- [5] M. Fliess, “Matrices de Hankel,” *Journal de Mathématiques Pures et Appliquées*, 53:197–222, 1974.
- [6] M. Hahn, “Theoretical limitations of self-attention in neural sequence models,” *Transactions of the Association for Computational Linguistics*, 8:749–767, 2020. <https://aclanthology.org/2020.tacl-1.11/>
- [7] J. J. Hopfield, “Neural networks and physical systems with emergent collective computational abilities,” *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982. <https://doi.org/10.1073/pnas.79.8.2554>
- [8] L. Le Cam and G. L. Yang, *Asymptotics in Statistics: Some Basic Concepts*, 2nd ed., Springer, 2000.
- [9] M. L. Littman, R. S. Sutton, and S. Singh, “Predictive representations of state,” in *Advances in Neural Information Processing Systems 14*, 2001. <https://papers.neurips.cc/paper/1983-predictive-representations-of-state.pdf>
- [10] J. R. Norris, *Markov Chains*, Cambridge University Press, 1997.
- [11] C. R. Shalizi and J. P. Crutchfield, “Computational mechanics: Pattern and prediction, structure and simplicity,” *Journal of Statistical Physics*, 104:817–879, 2001. <https://arxiv.org/abs/cond-mat/9907176>
- [12] S. Singh, M. R. James, and M. R. Rudary, “Predictive state representations: A new theory for modeling dynamical systems,” in *Proceedings of UAI 2004*, pp. 512–519, 2004. <https://arxiv.org/abs/1207.4167>
- [13] A. Vaswani et al., “Attention Is All You Need,” in *Advances in Neural Information Processing Systems 30*, 2017. <https://doi.org/10.48550/arXiv.1706.03762>
- [14] J. von Neumann, “Probabilistic logics and the synthesis of reliable organisms from unreliable components,” in C. E. Shannon and J. McCarthy, eds., *Automata Studies*, Princeton University Press, pp. 43–98, 1956. <https://www.degruyter.com/document/doi/10.1515/9781400882618-003/html>