# Active Cognitive Pruning Controls Reportable Access under Finite Capacity: A Rate-Distortion, Network-Topological, and Maintenance-Cost Model

Yining Wu[1, *]

[1] *Independent Researcher*

(Dated: May 15, 2026)

Reportable access can fail under sustained cognitive load even when local discrimination or reflexive responsiveness remains intact. Here we develop a finite-capacity theory of reportable access based on the FDS rate-distortion capacity-deficit framework. Representational residue $\rho$ is defined as accumulated unresolved rate-distortion surplus: the minimum task-relevant coding rate $R_{\min}^{(\tau)}(\varepsilon, t)$ exceeds accessible capacity $C_{acc}(t, \tau)$, producing a capacity deficit $\Delta_R > 0$ that must be managed through pruning, compression, externalization, or task relaxation. Residue damages access-network coherence $G$ and reportability $\mathcal{R}$ unless controlled by active cognitive pruning $S$. Global access is modeled through the leading spectral mode or giant component of a residue-damaged access graph, with the saddle-node normal form recovered as a mean-field projection of a high-dimensional network transition. The framework predicts pruning-dependent reportability thresholds, rescue-window closure, leading-covariance early-warning signals, and a partial order for anesthesia emergence. It also provides a taxonomy for artificial-agent reportability and a falsification protocol distinguishing core claims from domain-specific bridges. Reduced normal-form simulations illustrate the predicted regimes: pruning-dependent reportability thresholds, rescue-window closure, and leading-covariance early-warning behavior near access collapse. High-dimensional artificial-agent benchmarks are identified as a future in-machina validation path rather than as a required premise of the present theoretical argument. The Landauer lower bound for irreversible pruning is treated only as the informational heat floor; total biological maintenance cost includes much larger physical overheads. The framework does not claim to solve phenomenal consciousness; it converts reportable access into a testable finite-system maintenance problem governed by capacity deficit, residue accumulation, active pruning, network coherence, and resource-bounded persistence.

## I. INTRODUCTION

Theories of consciousness have emphasized distinct explanatory targets. Global Workspace Theory focuses on ignition and broadcast across specialized processors [1, 2]. Integrated Information Theory treats consciousness as intrinsic cause-effect integration [3, 4]. Predictive-processing and Free-Energy formulations describe perception and action as inference under a generative model [5, 6]. Perturbational-complexity approaches operationalize conscious level through the complexity of causal responses to perturbation [7, 8]. These frameworks have generated useful empirical programs, but they leave open a control-theoretic question: under finite capacity, what must a cognitive system actively maintain in order to keep information reportable?

This paper studies that question using a finite-distinction framework. The target is not the full metaphysics of phenomenality. The target is *conscious reportability*: the maintained ability of a system to integrate task-relevant distinctions into a coherent access state that can guide report, flexible action, and self-referential updating. A system may locally discriminate stimuli, produce reflexive responses, or process masked information unconsciously while failing to maintain a globally reportable access state. Such dissociations occur in masking, attentional blink, overload, sleep onset, general anesthesia, disorders of consciousness, delirium, and task switching.

The central hypothesis is that reportable access is controlled by *active cognitive pruning*. Here pruning denotes resource-coupled suppression, erasure, inhibition, down-weighting, compression, reconsolidation, or reorganization of internal states that consume capacity without contributing to current task-relevant access. Active forgetting and directed forgetting literatures already distinguish passive decay from controlled forgetting processes [28]. The internal load removed by pruning is called *representational residue*. Residue includes obsolete working-memory items, intrusive affective states, unresolved prediction errors, stabilized but maladaptive associations, irrelevant sensory traces, and representational interference. The key point is not that all such processes share one neural mechanism. The key point is that finite systems need a selective operation that bounds nonfunctional distinction load while preserving useful distinctions.

This framing parallels a broader maintenance principle for finite systems: sustained flux generates residue; residue impairs access to functional states; active pruning controls persistence. In protocell-like systems, residue can impair transport or boundary integrity. In cognitive systems, representational residue impairs access coherence, self-model updating, and reportability. The analogy is structural rather than molecular.

The central object of this paper is not consciousness as such, but reportable access under finite capacity. In FDS terms, a cognitive system receives task-relevant distinc-

* yining.wu@alumni.upenn.edu

tion demand $R^{(\tau)}_{\min}(\varepsilon, t)$ and possesses only finite accessible capacity $C_{acc}(t, \tau)$. When task demand exceeds accessible capacity, the resulting capacity deficit $\Delta_R$ must be absorbed by compression, invariant extraction, externalization, task relaxation, or active pruning. If these mechanisms are insufficient, unresolved distinction load accumulates as representational residue $\rho$, which damages the causal-loop and access-network conditions required for flexible report.

### A. Central claim

The central claim is:

> In finite cognitive systems under sustained distinction load, active cognitive pruning can act as a control parameter separating a maintained reportability regime from overload-induced access collapse.

This is stronger than the generic claim that attention or inhibition matters, because it predicts thresholds, rescue windows, non-monotonic self-model coherence, and measurable early-warning signatures. It is weaker than the claim that consciousness has been solved. The theory is a benchmark framework for conscious access, not a complete metaphysics of experience.

### B. Contributions

This paper makes ten contributions.

1. It defines reportable access as a maintained finite-capacity regime rather than raw local discrimination.

2. It derives representational residue from FDS rate-distortion capacity deficit $\Delta_R = R^{(\tau)}_{\min}(\varepsilon, t) - C_{acc}(t, \tau)$, rather than from unresolved variational free energy.

3. It treats active inference / FEP as one implementation-level estimator of unresolved distinction load, not as the foundational derivation.

4. It derives a pruning lower bound for keeping residue below a reportability ceiling.

5. It separates functional retained structure $m$ from nonfunctional residue $\rho$, resolving the non-monotonic access window into distinct mechanisms.

6. It replaces a purely one-dimensional saddle-node story with a network-topological access model in which residue damages access edges and pruning restores functional connectivity.

7. It derives early-warning signatures from both the normal form and the leading covariance eigenmode of the access network.

8. It provides a total maintenance-cost decomposition $\dot{Q}_{\text{maint}} = \dot{Q}_{\text{phys}} + \dot{Q}_{\text{info}} + \Gamma(S) + \Xi(E_{\text{ext}})$ with Landauer as the informational heat floor only.

9. It provides an FDS agency taxonomy and an artificial-agent benchmark protocol for future in-machina tests of reportability-like access, while keeping the present paper focused on theory, reduced simulations, and falsifiable predictions.

10. It gives a minimal falsification protocol with layered failure consequences, operational proxy maps, and differential predictions against GWT, IIT, FEP, information bottleneck, PCI/complexity, and AI agency.

### C. What is not claimed

The paper does not claim to solve the hard problem of phenomenal consciousness. It does not prove that lossy compression alone produces experience. It does not identify one neural correlate of consciousness. It does not claim that all loss-of-consciousness events are saddle-node bifurcations, nor that every transition in sleep, anesthesia, or attention shares one mechanism. It does not claim that all current large language models are unconscious by metaphysical necessity. Instead, it states an operational requirement: a candidate finite system that maintains reportable conscious access must maintain task-relevant distinctions under bounded capacity, sustained input, residue accumulation, causal-loop constraints, and active pruning or an equivalent maintenance mechanism.

### D. Falsifiable predictions

The framework can be falsified or demoted in specific ways.

1. **Pruning threshold.** If finite systems under sustained distinction load maintain stable reportability indefinitely with $S = 0$, no hidden capacity growth, no hidden dilution, and no equivalent compression/inhibition operation, the active-pruning threshold claim is falsified.

2. **Network topology.** If reportability transitions show no measurable change in access-network topology, spectral connectivity, effective connectivity, or perturbational complexity after controlling for arousal and task demands, the network-topological bridge is weakened.

3. **Functional retained structure.** If functional retained structure $m$ shows no measurable contribution to self-model coherence beyond the effect accounted for by nonfunctional residue $\rho$, the nonmonotonic self-model component is demoted.

4. **Anesthesia recovery ordering.** If high-level self-report coherence reliably recovers before causal responsiveness and access-network recovery during emergence from anesthesia, the proposed recovery-ordering claim is demoted.

5. **Rescue-window closure.** If restoring pruning after overload always recovers reportability with no dependence on delay or state, the attractor-loss interpretation is demoted.

6. **Energy bridge.** If experimentally isolated irreversible pruning of cognitive load requires no measurable energetic cost above noise floor and no lower-bound-compatible heat/metabolic signature, the Landauer-bridge interpretation is weakened.

## II. RATE-DISTORTION ORIGIN OF REPRESENTATIONAL RESIDUE

### A. Capacity deficit

Consider a finite cognitive system receiving task-relevant input over time. Let $R^{(\tau)}_{\min}(\varepsilon, t)$ be the minimum coding rate needed to maintain reportable access within distortion tolerance $\varepsilon$ over timescale $\tau$. Let $C_{acc}(t, \tau)$ be the accessible representational capacity. The FDS capacity deficit is

$$\Delta_R(t, \tau, \varepsilon) = R^{(\tau)}_{\min}(\varepsilon, t) - C_{acc}(t, \tau). \tag{1}$$

When $\Delta_R > 0$, the system cannot fully represent task-relevant distinctions within its capacity. The unresolved surplus must be managed.

### B. Representational residue

Define representational residue $\rho(t)$ as the accumulated unresolved rate-distortion surplus that cannot be encoded, compressed, integrated, externalized, or pruned within the accessible capacity and resource budget:

$$\begin{aligned}\dot{\rho} = \zeta\big[&\Delta_R(t, \tau, \varepsilon) \\ &- C_{inv}(t) - G_{ext}(t) - G_{relax}(t)\big]_+ \\ &- d\rho - S\,\rho/(K + \rho),\end{aligned} \tag{2}$$

where $C_{inv}(t)$ is capacity saved through invariant compression, $G_{ext}(t)$ is effective gain from externalization, $G_{relax}(t)$ is reduction via task relaxation, $d$ is passive decay, and $S\rho/(K + \rho)$ is saturating active pruning.

Residue operates at three levels:

- **Computational residue:** unresolved rate-distortion surplus (compression error, bottleneck loss, predictive-information shortfall).

- **Behavioral residue:** task-irrelevant interference impairing report (lapses, intrusions, perseveration, attentional-blink cost).

- **Network residue:** load-dependent degradation of access topology (leading covariance mode, effective connectivity loss).

These levels are not assumed identical. The theory predicts lawful covariation under controlled load, pruning, capacity, and externalization manipulations.

### C. Active-inference implementation as a special case

The FDS rate-distortion derivation does not require Bayesian inference or variational objectives. However, in systems that encode observations through a recognition density $q_\phi(s_t, z_t|o_{\le t})$ and minimize variational free energy $\mathcal{F}_{\text{var}}$, the unresolved free energy or KL load can serve as an estimator of $\Delta_R$. When channel capacity $C_{\text{eff}}$, update energy $E_{\text{avail}}$, and information-erasure cost $\Delta Q_{\text{erase}}$ are finite, the constrained objective

$$\mathcal{J} = \mathbb{E}\big[\mathcal{F}_{\text{var}} + \lambda_C\,(I(o; z) - C_{\text{eff}})_+ \tag{3}$$

$$+ \lambda_E E_{\text{update}} + \lambda_Q Q_{\text{erase}}\big] - \beta_Y I(z; y) \tag{4}$$

reproduces the residue dynamics of Eq. (2) under the identification $\Delta_R \approx \mathcal{F}_{\text{var}} - \mathcal{F}_{\text{cap}}(C_{\text{eff}}, E_{\text{avail}})$. This makes active inference one possible implementation of FDS capacity-deficit dynamics, not the foundational derivation.

## III. NETWORK-TOPOLOGICAL ACCESS MODEL

### A. Access graph

A one-dimensional saddle-node model is useful but fragile if interpreted literally. Real brains and artificial agents are high-dimensional networks. We therefore define an access graph with nodes representing modules, cortical parcels, memory buffers, latent subspaces, or agent subsystems. Let $W_{ij}(t)$ be functional or effective coupling and let $\rho_i(t)$ be local residue. The effective access matrix is

$$M_{ij}(t) = a_i(t) W_{ij}(t) a_j(t) - \delta_{ij} \chi_\rho \rho_i(t), \tag{5}$$

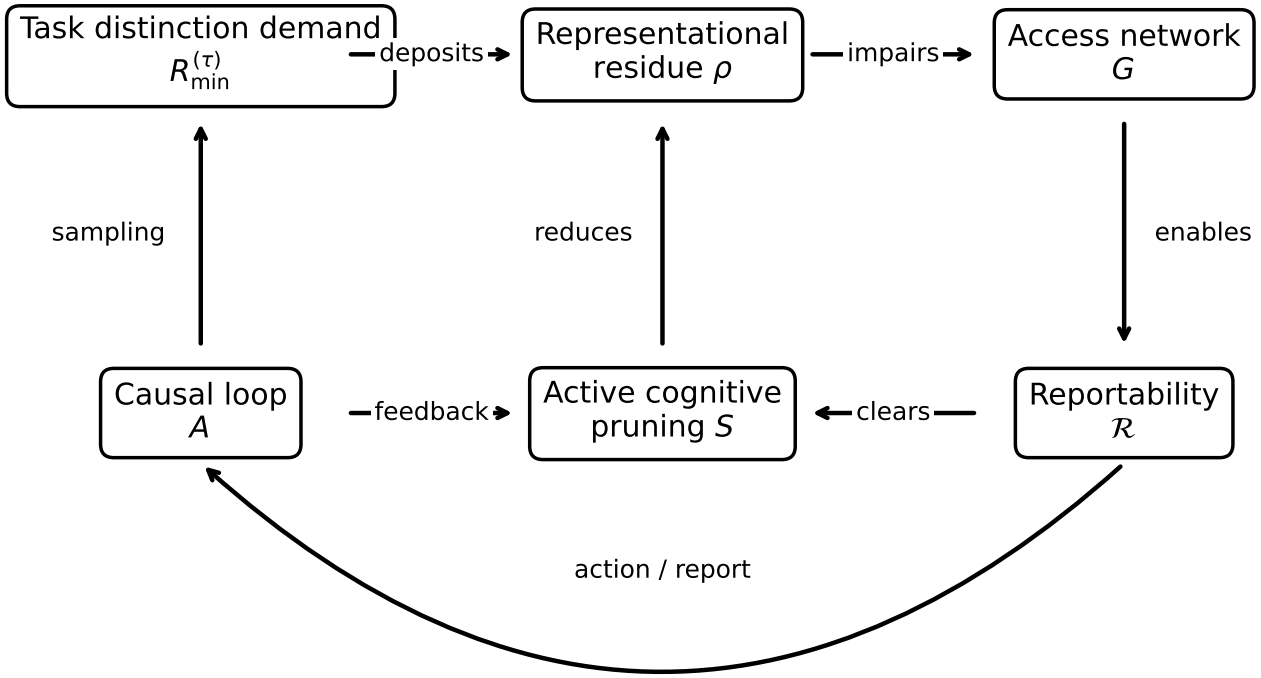


FIG. 1. Capacity-deficit maintenance loop. Task-relevant rate-distortion demand $R_{\min}$ and limited accessible capacity $C_{acc}$ produce capacity deficit $\Delta_R$, which deposits representational residue $\rho$. Residue impairs access coherence $G$ and reportability $\mathcal{R}$. Active cognitive pruning $S$ removes residue but consumes resources. Reportability persists while $\rho$ is bounded and pruning exceeds the threshold $S_c$.

where $a_i$ is local availability and $\chi_\rho$ is residue damage. Active pruning updates residue and connectivity:

$$\dot{\rho}_i = L_i - d_i\rho_i - S_i\frac{\rho_i}{K_i+\rho_i}, \tag{6}$$

$$\dot{W}_{ij} = \Gamma^{\text{task}}_{ij} - \lambda_\rho(\rho_i+\rho_j)W_{ij} - \lambda_S S^{\text{cut}}_{ij} + \lambda_R S^{\text{repair}}_{ij}. \tag{7}$$

The exact graph dynamics depend on substrate. The access criterion is more general:

$$G(t) = \sigma\left(\beta_G[\lambda_1(M(t)) - \lambda_c]\right), \tag{8}$$

where $\lambda_1$ is the leading eigenvalue or an equivalent giant-component/order parameter. Reportability is maintained only when an access component remains large enough and coherent enough for global report.

Residue can damage access in two topologically distinct ways. It can remove nodes from the giant access component, analogous to percolation. Or it can leave nodes connected while degrading spectral integration, producing a collapse of the leading eigenmode or algebraic connectivity. Both are compatible with the same coarse-grained $G$ variable.

### B. Mean-field reduction and the saddle-node projection

Let $m(t)$ be a mean access order parameter, such as normalized leading eigenvalue or giant-component size. Under homogeneous mean-field assumptions, the network dynamics reduce to

$$\dot{m} = F(m; S, C_{\text{eff}}, H, \rho). \tag{9}$$

If $F$ has a fold in the control parameter

$$r = a_S(S - S_c) + a_C(C_{\text{eff}} - H) - a_\rho\rho, \tag{10}$$

then the center-manifold reduction gives

$$\dot{x} = r - x^2 + O(x^3, rx), \tag{11}$$

where $x$ is the dominant access mode. The associated potential is

$$U(x) = -rx + \frac{x^3}{3}. \tag{12}$$

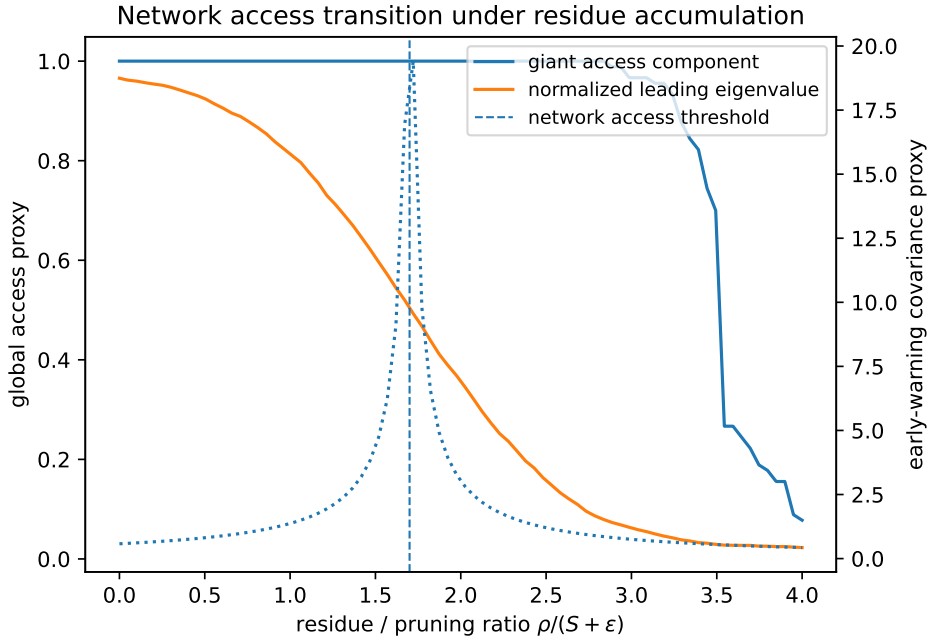


FIG. 2. Network access transition under residue accumulation. As the capacity-deficit-driven residue increases, the leading spectral proxy and giant access component decline. The leading covariance proxy rises near the transition, providing a high-dimensional early-warning signal consistent with FDS covariance-based collapse prediction.

Thus the saddle-node is not assumed as the complete brain dynamics. It is the local normal form when a high-dimensional access network loses reportability through one dominant critical mode. Other transitions are possible: Hopf, noise-induced escape, synchronization loss, or discontinuous percolation. The framework predicts that if the transition is fold-like, rescue windows and early warnings should appear.

### C. Optional topological strengthening: a cognitive two-kink test

A stronger but optional topological prediction can be derived by analogy with the FDS physical bridge TP-3. If reportability loss is mediated by a topology-changing access-network transition, then the theory predicts aligned finite-size-smoothed slope changes in an operational forgetting rate $\kappa_{\text{access}}$ and a maintenance-cost or entropy-production proxy $\dot{\Sigma}_{\text{maint}}$ near the same control parameter $\gamma_c$. This paired nonanalytic behavior is the cognitive analogue of the TP-3 two-kink prediction for physical systems.

This prediction is conditional. The core reportability model predicts thresholds, rescue-window closure, and early-warning behavior without requiring a topological transition. The two-kink prediction applies only if an access-network topological transition or invariant-supported reorganization is independently established. Failure of the two-kink prediction would demote the topological bridge, not the core capacity-deficit theory of reportable access.

## IV. LOAD-PRUNING-ACCESS DYNAMICS WITH FUNCTIONAL RETAINED STRUCTURE

### A. Separating functional structure from nonfunctional residue

A purely passive system with no persistent internal structure does not support rich reportability. But not all internal load is harmful. We therefore distinguish two variables:

$$m(t) = \text{functional retained structure}, \tag{13}$$

$$\rho(t) = \text{nonfunctional representational residue}. \tag{14}$$

Functional retained structure includes task-relevant working content, self-model, and causal-loop representations. Nonfunctional residue is the accumulated capacity-deficit surplus from Eq. (2).

### B. Core dynamics

Functional structure is maintained by task-relevant input and reorganized by pruning:

$$\dot{m} = \alpha_m I_{\text{task}}(t) - \beta_m m - \chi_\rho \rho m + \Gamma_{\text{reorg}}(S, m, \rho), \tag{15}$$

where $I_{\text{task}}$ is task-relevant input, $\beta_m$ is passive decay of retained structure, $\chi_\rho$ is damage from residue, and

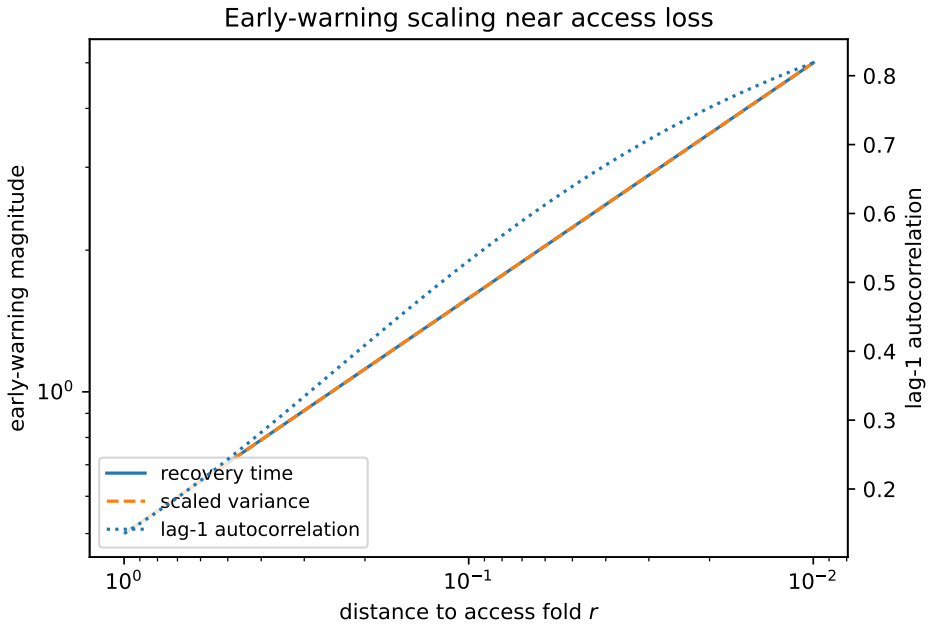


FIG. 3. Early-warning scaling near access loss. The reduced normal form predicts increasing recovery time, variance, and lag-one autocorrelation as the distance to the fold decreases. In high-dimensional data, the leading covariance eigenvalue and recovery time of the dominant access mode are preferred over scalar variance.

$\Gamma_{\text{reorg}}$ captures reorganization during pruning (preserving functional structure while removing nonfunctional load).

Nonfunctional residue dynamics follow the capacity-deficit equation:

$$\dot{\rho} = \zeta\,[\,\Delta_R - C_{inv} - G_{ext} - G_{relax}]_+ - d\rho - S\frac{\rho}{K+\rho}. \tag{16}$$

Global access coherence is damaged by residue and supported by functional structure:

$$\frac{dG}{dt} = \gamma_G\,[\sigma(\beta_G[\lambda_1(M(t)) - \lambda_c]) - G] - \beta_\rho \rho G, \tag{17}$$

where $\lambda_1(M)$ is the leading eigenvalue of the effective access matrix (defined in Sec. III).

Causal-loop availability evolves separately:

$$\frac{dA}{dt} = \gamma_A(A_0 - A) - \beta_A \rho A + u_A(t). \tag{18}$$

Reportability couples all components:

$$\mathcal{R}(t) = \frac{A(t)G(t)F_m(m)}{1+\rho(t)/\rho_0}, \tag{19}$$

where $F_m(m)$ is the contribution of functional retained structure. A simple form is $F_m(m) = 1 - e^{-m/m_0}$, representing the saturating benefit of persistent structure. For self-model coherence, a non-monotonic form $F_m(m) = me^{-\lambda m}$ captures the cost of excessive rigidity. Throughout this paper, $\mathcal{R}(t)$ denotes reportability and $R^{(\tau)}_{\min}(\varepsilon, t)$ denotes the minimum rate-distortion demand, which are distinct quantities.

### C. Pruning threshold theorem

Stated in FDS terms:

**Proposition 1 (Pruning lower bound)** *Let $\rho_c$ be the maximum residue compatible with reportable access. Let*

$$L^*_R = \zeta\,[\Delta^*_R - C^*_{inv} - G^*_{ext} - G^*_{relax}]_+ \tag{20}$$

*be the effective unresolved distinction-load deposition rate over a relevant interval. If passive decay is $d\rho$ and pruning is $S\rho/(K+\rho)$, then maintaining $\dot{\rho} \le 0$ at $\rho = \rho_c$ requires*

$$S \ge S_c(\rho_c) = (L^*_R - d\rho_c)_+ \frac{K+\rho_c}{\rho_c}. \tag{21}$$

*Interpretation.* If $S < S_c$, residue cannot be bounded at $\rho_c$ without additional capacity growth, externalization, invariant compression, or task relaxation. This connects the reportability model to the FDS prune–externalize–collapse trichotomy.

This section does not answer why there is something it is like to be a system. It does, however, make a structural claim about the geometry of reportable experience, placed in the appendix for those interested in the geometric extension.

## V. THERMODYNAMIC MAINTENANCE COST

If active pruning irreversibly erases or reorganizes residual distinctions, it must have an energetic cost. The

FDS framework decomposes total maintenance cost into physical and informational components:

$$\dot{Q}_{\text{maint}} = \dot{Q}_{\text{phys}} + \dot{Q}_{\text{info}} + \Gamma(S) + \Xi(E_{\text{ext}}), \tag{22}$$

where $\dot{Q}_{\text{info}}$ is the informational heat floor from logically irreversible pruning, $\dot{Q}_{\text{phys}}$ is the biological or hardware physical overhead, $\Gamma(S)$ is the control cost of active pruning, and $\Xi(E_{\text{ext}})$ is the cost of externalization, retrieval, memory routing, or tool use.

The informational term follows Landauer's principle:

$$\dot{Q}_{\text{info}} \geq \frac{k_B T \ln 2}{\tau} H(M_t | M_{t+1}, Y_t), \tag{23}$$

for logically irreversible state updates. This does not mean neural tissue operates near the Landauer limit. Biological overheads are many orders of magnitude larger. The importance of Eq. (23) is directional and constraining: pruning cannot be free. Neural measures such as BOLD, $CMRO_2$, glucose uptake, electrophysiological power, or local heat should be interpreted as proxies for $\dot{Q}_{\text{maint}}$, not for $\dot{Q}_{\text{info}}$ alone.

Operational proxies include local field potential power, BOLD/$CMRO_2$ coupling, glucose metabolism, heat-sensitive micro-measurements in artificial systems, and energy-per-bit estimates in neuromorphic or robotic agents. EEG/MEG complexity measures such as Lempel-Ziv complexity and perturbational complexity provide complementary information-theoretic proxies [7, 13].

## VI. SIMULATION BENCHMARKS

The numerical simulations are not intended as fitted neural models. They are benchmark demonstrations showing how the theoretical quantities behave under controlled assumptions. The simulation figures are generated from released Python code with fixed random seeds.

### A. Recovery ordering under anesthesia emergence

The model predicts a partial order, not a fixed universal timing. Causal responsiveness $A$ can return before full global access $G$; access-network recovery should precede stable self-model coherence; inhibitory/pruning rhythms should recover during the transition. This ordering is motivated by propofol EEG signatures and causal-complexity approaches to consciousness [7, 12]. Recent propofol studies also support a dynamical and network-level interpretation: propofol destabilizes neural dynamics across cortex [29] and disrupts predictive routing and local field phase modulation [30]. It can be tested with time-resolved EEG/MEG/fMRI during loss and recovery.

### B. Rescue window

A pruning-interruption protocol is used to test reversibility. A system is first maintained with $S > S_c$, then pruning is shut off. At delay $t_{\text{restore}}$, pruning is restored to $S_{\text{rescue}}$. Rescue succeeds if $\mathcal{R}(t) > \mathcal{R}_{\text{rec}}$ within an observation window. The model predicts that required rescue pruning increases with delay and that a 50% rescue boundary exists.

### C. Phase diagram

The model predicts a phase diagram over distinction load and pruning capacity. Low pruning and high load produce non-reportability; sufficient pruning maintains access; very low load or excessive rigid filtering can also reduce rich reportability. Figure 6 shows the central threshold structure.

### D. Future in-machina benchmark program

The reduced simulations above test the normal-form structure of the FDS reportability model. A stronger validation path is to embed the same variables in high-dimensional artificial agents operating under nonstationary task drift, bounded memory, prediction-error accumulation, and selective pruning. Such benchmarks should compare at least four conditions (summarized in Table I): no pruning, passive decay, random pruning, and active task-sensitive pruning. They should measure reportability-like access, residue accumulation, access-network coherence, covariance early-warning signals, and rescue-window closure after delayed restoration of pruning.

The present paper does not require a particular artificial-agent implementation. Production-level embodied-agent architectures, robotics perception stacks, memory graphs, mission systems, pruning schedulers, and deployment control loops are outside the scope of this manuscript. The role of in-machina benchmarks is to test the FDS prediction that selective pruning, unlike passive decay or random deletion, maintains reportability-like access under finite capacity and nonstationary load.

## VII. EXPERIMENTAL PREDICTIONS

We organize predictions into three primary tests and several secondary domains.

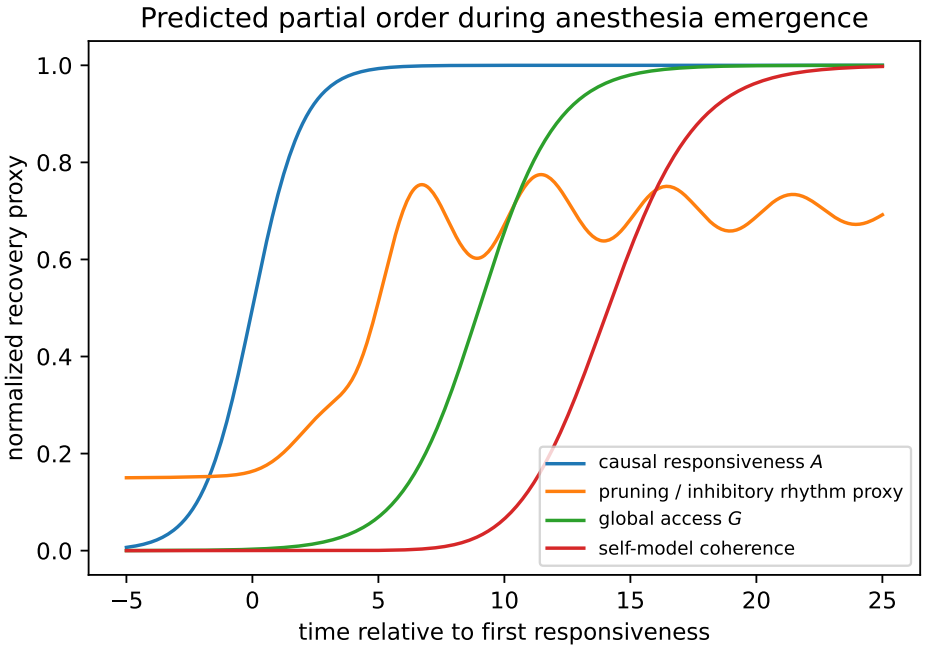


FIG. 4. Predicted partial order during anesthesia emergence. The model predicts that causal responsiveness can recover before full global access and self-model coherence. The precise timing is substrate-dependent; the falsifiable claim is the ordering constraint under controlled emergence.

TABLE I. Proposed artificial-agent benchmark conditions. This table defines a future in-machina test protocol rather than reporting unpublished benchmark results.

| Condition | Maintenance mechanism | FDS prediction |
|---|---|---|
| No pruning | No active residue control | Residue accumulation and access collapse under s |
| Passive decay | Nonselective load reduction | Partial residue reduction but poor preservation of |
| Random pruning | Random deletion | Possible load reduction but unstable reportability |
| Active pruning | Task-sensitive selective removal or compression | Bounded residue and maintained reportability-like |
| Active pruning + externalization | Selective pruning plus external memory/tool support | Strongest maintenance under high load |

### A. Primary test 1: masking and attentional blink capacity-deficit test

T1 load increases $\Delta_R$, raising $\rho$ and increasing the reportability threshold for T2. Active control, rest, distractor suppression, or sleep-like reorganization should reduce $\rho$ and improve T2 reportability. The theory predicts more than a binary report/no-report curve: recovery time, trial-to-trial autocorrelation, and leading covariance of access-network measures should rise as the system approaches the reportability threshold. Under attentional blink, residual load from target 1 should raise $S_c$ for target 2.

*Failure condition.* If T2 reportability does not depend on load-induced residue or capacity manipulation after controlling for arousal and sensory strength, the capacity-deficit bridge is weakened.

### B. Primary test 2: anesthesia emergence ordering

The strongest human test is time-resolved emergence from anesthesia. The predicted partial order is

$$A_{\text{reflex}} \to A_{\text{command}} \to G_{\text{access}} \to I_{\text{self}}, \qquad (24)$$

where $A_{\text{reflex}}$ is reflexive responsiveness, $A_{\text{command}}$ is command following, $G_{\text{access}}$ is stable global access, and $I_{\text{self}}$ is coherent autobiographical or metacognitive self-report. Access-network topology and perturbational complexity should recover before stable self-model coherence; early-warning indicators should appear near transitions when emergence is quasi-static. Perturbational complexity, EEG signatures under propofol, and signal-complexity changes under anesthesia provide direct empirical anchor points [7, 8, 12, 13].

*Failure condition.* If high-level self-report coherence reliably recovers before causal responsiveness and access-network recovery across controlled cohorts and modalities, demote the recovery-ordering claim.

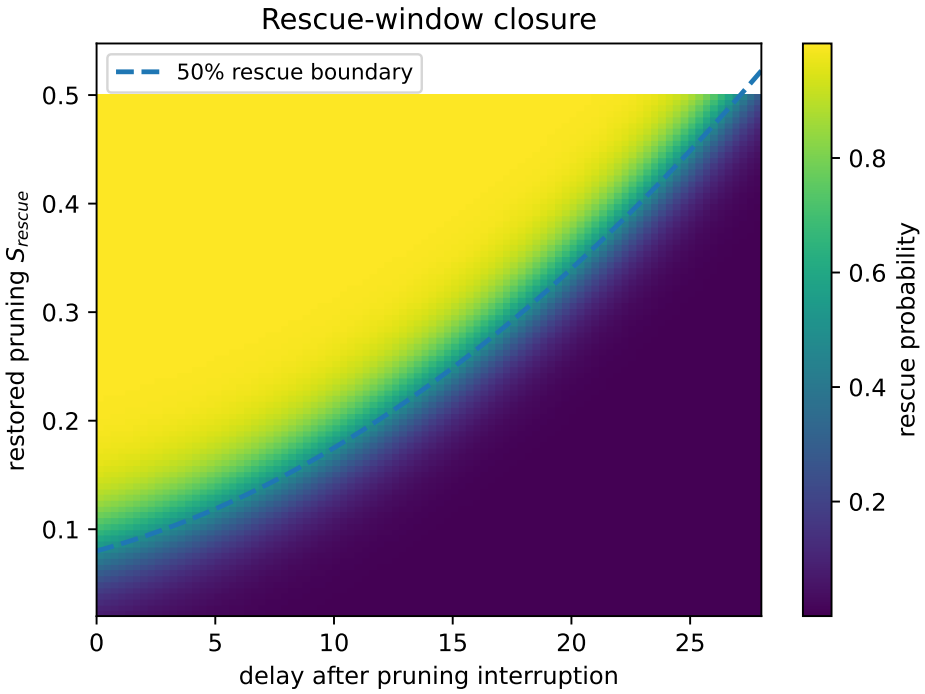


FIG. 5. Rescue-window closure. The heatmap shows rescue probability as a function of delay and restored pruning. Delayed intervention requires stronger pruning and can fail once the system crosses the basin boundary of reportability.

### C. Primary test 3: artificial-agent sleep-like pruning benchmark

A resource-limited persistent agent serves as the primary in machina test. In a long-horizon nonstationary environment with drifting rules, bounded memory, noisy observations, and delayed consequences, the theory predicts that agents without pruning accumulate residue and lose adaptability. Agent variants include stateless mapper, persistent-memory without pruning, random forgetting, active pruning, externalization-enabled, and sleep-like offline replay. The FDS prediction is that active pruning outperforms no-pruning and random-forgetting agents, especially when distribution shifts generate stale internal structure.

*Failure condition.* If active pruning or externalization provides no advantage over no-pruning or random forgetting under bounded memory and sustained task drift, demote the artificial-agent bridge.

### D. Secondary tests

Additional domains that support but are not required for the core theory: sleep deprivation (predicted increase in residue proxies and narrowing of rescue window), disorders of consciousness (predicted correlation between access-network topology and reportability), no-report paradigms (predicted dissociation between local processing and global access), EEG/MEG leading covariance early warnings (predicted rise near reportability transitions), and the information-theoretic cost of cognitive pruning (predicted correlation between pruning load and metabolic proxies). An operational proxy map is provided in Appendix E.

### E. Artificial agency taxonomy and benchmark

*a. FDS agency classification.* The FDS AI agency framework distinguishes system types by their capacity-deficit management:

A stateless base model evaluated as an input-output mapper does not instantiate the full FDS reportability loop. It may transform inputs into high-quality outputs without maintaining an operational boundary, updating durable self-relevant state, or pruning accumulated residue. A coupled architecture containing writable memory, monitoring, action channels, externalization, active pruning, and causal participation in future viability conditions may instantiate system-level residue-pruning dynamics. The unit of analysis is the maintained system boundary, not the neural network alone. The taxonomy is architectural in the formal sense but not implementation-specific. The protocol is intended to define what future artificial-agent benchmarks must test; it does not specify or depend on any particular implementation. A public benchmark should expose sufficient details for reproduction, while production embodied-agent architectures and proprietary pruning mechanisms may

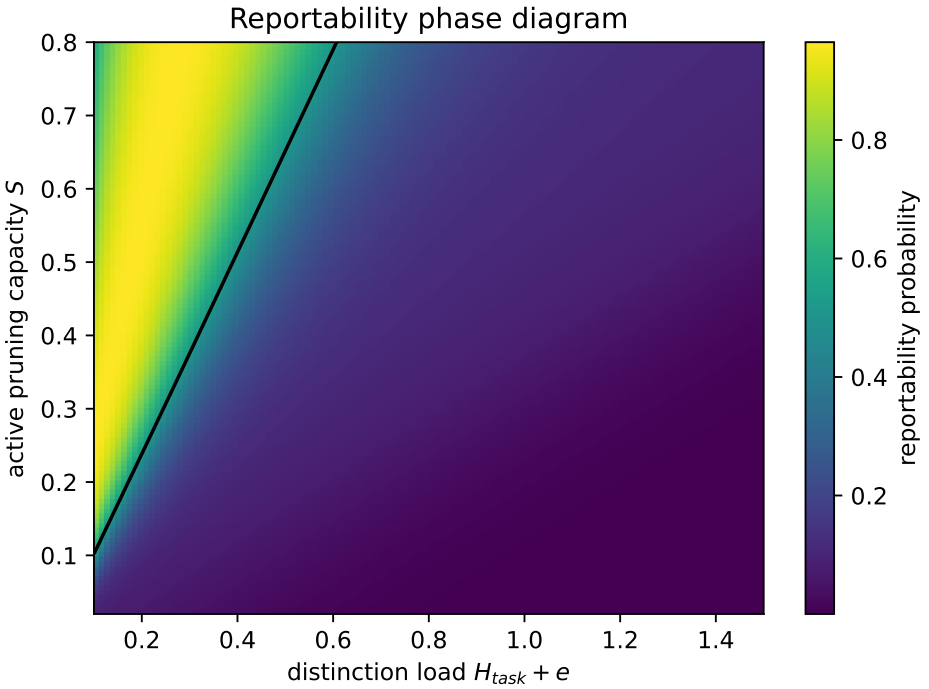


FIG. 6. Reportability phase diagram. Reportability probability is plotted over distinction load and pruning capacity. The contour marks the 50% reportability threshold.

| System type | FDS classification | Residue-pruning dynamics |
|---|---|---|
| Stateless feedforward model | Passive mapper | No persistent system-level residue |
| Base LLM with context window | Weak scaffolded mapper | Transient context load only |
| LLM + memory + tools + planner | Scaffolded system | System-level residue, externalization burden |
| Embodied adaptive controller | Adaptive controller | Persistent maintenance burden |
| Strong FDS-agent | Strong agent | Full capacity-deficit, pruning, causal loop |

TABLE II. FDS agency classification relevant to reportability-like access.

remain outside the paper's scope. Concrete embodied-agent designs, pruning policies, memory graphs, mission systems, and robotics control stacks are outside the scope of this paper.

*b. Artificial-agent benchmark protocol.* A resource-limited persistent agent (recurrent or reinforcement-learning) serves as the primary in machina test. The environment is nonstationary with drifting rules, long-horizon dependencies, bounded memory, noisy observations, and delayed consequences. Agent variants include: stateless mapper, persistent-memory without pruning, random forgetting, active pruning, externalization-enabled, and sleep-like offline replay. Recent work on LLM-agent memory highlights the need for persistent memory, update, retrieval, and evaluation beyond static context windows [31].

## VIII. DIFFERENTIAL PREDICTIONS RELATIVE TO EXISTING THEORIES

The proposed framework shares ground with several existing theories but adds distinctive predictions (summarized in Table III).

*a. Relative to GWT.* Global Workspace Theory emphasizes ignition and broadcast across specialized processors. The FDS framework agrees that global access is central but adds that broadcast alone is insufficient: under matched broadcast markers, accumulated capacity-deficit residue can impair reportability, produce rescue-window closure, and generate early-warning signatures not predicted by GWT alone.

*b. Relative to IIT.* Integrated Information Theory identifies intrinsic cause-effect integration as the substrate of consciousness. The FDS framework agrees that integration matters but predicts that high integration without adequate pruning can produce rigid or overloaded states rather than stable reportability. The leading covariance eigenvalue near reportability collapse pro-

vides a distinct testable signature.

*c. Relative to FEP.* The free-energy principle treats perception and action as inference under a generative model. The FDS framework treats the unresolved component of free energy under finite capacity as representational residue, making pruning thresholds and rescue windows explicit predictions that FEP does not derive.

*d. Relative to information bottleneck.* The information bottleneck method formalizes compression under capacity constraints. FDS agrees that compression is necessary but predicts that stateless compression is insufficient for reportability: persistent state, causal-loop embedding, and active maintenance are required.

*e. Relative to PCI/complexity.* Perturbational complexity and Lempel-Ziv complexity track conscious level empirically. FDS adds a mechanistic interpretation: these measures reflect access-network topology and leading eigenvalue dynamics. The theory predicts leading covariance early warnings near transitions that PCI markers alone do not specify.

*f. Relative to AI agency.* The FDS AI agency framework distinguishes passive mappers from scaffolded and strong agents (Table II). A stateless model evaluated as an input-output mapper does not instantiate the full reportability loop. A coupled architecture with writable memory, pruning, externalization, and causal participation can be tested for reportability-like access dynamics.

## IX. METHODS AND REPRODUCIBILITY

### A. Nondimensionalization

All simulations use nondimensional benchmark units. Define $\tilde{\rho} = \rho/\rho_0$, $\tilde{H} = H/H_0$, $\tilde{C} = C_{acc}/H_0$, $\tilde{t} = t/\tau$. Simulation figures are not fitted neural models and do not imply direct numerical equivalence between bits, joules, BOLD response, and behavioral report.

### B. Simulation protocol

All simulations are nondimensional and are intended as benchmark tests, not fitted biological models. The released code generates all figures and CSV tables. Parameters are summarized in Table IV. A run is classified as reportable if $\mathcal{R}(t) > \mathcal{R}_{\min}$ and $G(t) > G_{\min}$ over the final observation window and if $|\dot{\mathcal{R}}|$ remains below a settling tolerance. A run is classified as failed if $\mathcal{R}(t) < \mathcal{R}_{\min}$ for longer than a dwell time or if $G$ or $A$ falls below its failure threshold.

Critical pruning thresholds are estimated by bracketing and bisection over $S$. For stochastic simulations, reportability probability is estimated across random seeds and the threshold is the 50% access point obtained by monotone interpolation or logistic fit. Rescue-window boundaries are estimated as 50% rescue contours over $(t_{\text{restore}}, S_{\text{rescue}})$. Confidence intervals can be obtained by bootstrap resampling over seeds.

| Theory | Shared ground | FDS-specific addition | Distinguishing prediction |
|---|---|---|---|
| GWT | global access/broadcast | access must be maintained under capacity deficit | matched broadcast markers can fail under high residue |
| IIT | integration matters | integration can become rigid without pruning | high integration with poor pruning may impair flexible report |
| FEP | error minimization | unresolved rate-distortion surplus becomes residue | pruning thresholds and rescue windows |
| Information bottleneck | compression | compression must be embedded in causal-loop maintenance | stateless compression is insufficient |
| PCI/complexity | complexity tracks conscious level | topology and residue determine recovery/collapse | leading covariance early warnings near transition |
| AI agency | causal loop and memory | reportability-like access requires pruning under capacity deficit | passive mappers lack full residue-pruning loop |

TABLE III. Differential predictions relative to existing theories.

TABLE IV. Simulation parameters and reporting roles. Baseline values and sweep ranges are recorded in the released code.

| Symbol | Meaning | Role in model | Reporting requirement |
|---|---|---|---|
| $\alpha$ | overload-to-residue gain | residue deposition | baseline + sensitivity |
| $\eta_e$ | error-to-residue gain | unresolved prediction-error load | baseline + sensitivity |
| $d$ | passive decay | non-active residue reduction | control runs |
| $S$ | active pruning capacity | main control parameter | grid/bisection range |
| $K$ | pruning half-saturation | nonlinear pruning response | baseline value |
| $\gamma_G$ | access recovery rate | relaxation of global access | baseline + sensitivity |
| $\beta_\rho$ | residue damage coefficient | access impairment | baseline + sensitivity |
| $C_{acc}$ | effective capacity | finite-capacity constraint | sweep range |
| $\theta_G$ | access threshold | global access activation | fixed criterion |
| $\mathcal{R}_{\min}$ | reportability failure threshold | collapse classification | fixed criterion |
| $\mathcal{R}_{\mathrm{rec}}$ | rescue threshold | rescue classification | fixed criterion |
| $\sigma$ | noise amplitude | early-warning simulation | values used |
| $N$ | network size | access graph simulation | value + random seed |

## X. DATA AND CODE AVAILABILITY

Source code, parameter files, generated figures, and CSV tables are provided for the reduced illustrative simulations included in this paper. High-dimensional artificial-agent benchmarks are discussed as a future validation path and are not part of the public release. Production-level embodied-agent architectures, robotics-oriented benchmarks, proprietary pruning implementations, and deployment-specific evaluation suites are outside the scope of this manuscript.

## XI. MINIMAL FALSIFICATION PROTOCOL

The framework can be falsified or demoted in specific ways. Domain claims should fail locally without destroying the formal core.

### A. Primary tests

**Capacity-deficit residue.** Demote if sustained reportability is maintained under $\Delta_R > 0$ with no pruning, no externalization, no task relaxation, no compression improvement, and no hidden capacity growth.

**Network-topological bridge.** Demote if reportability transitions show no measurable change in access-network topology, spectral connectivity, effective connectivity, or perturbational complexity after controlling for arousal and task demand.

**Rescue-window closure.** Demote if restoring pruning always recovers reportability independent of delay, residue state, or access-network damage.

**Anesthesia ordering.** Demote if high-level self-report coherence reliably recovers before causal responsiveness and access-network recovery across controlled cohorts and modalities.

**AI pruning benchmark.** Demote if active pruning or externalization gives no advantage over no-pruning or random forgetting under bounded resources and nonstationary load.

### B. Optional bridge tests

**Topological two-kink extension.** Demote only the topological bridge if no paired kink appears under independently confirmed topology-mediated access transition. Do not propagate to core capacity-deficit claims.

## XII. LIMITATIONS

First, the model targets reportability and conscious access, not the full ontological problem of phenomenal consciousness. Second, the rate-distortion derivation is coarse-grained: it motivates the capacity-deficit equations but does not identify a unique neural implementation. Third, the network model abstracts away many details of thalamocortical, corticocortical, and subcortical dynamics. Fourth, the saddle-node normal form is a local candidate mechanism, not a universal theorem for all consciousness transitions; the model explicitly allows Hopf, noise-induced, percolation, and synchronization-loss transitions. Fifth, the thermodynamic bridge provides a lower bound, not an exact metabolic prediction. Sixth, the simulations are benchmark demonstrations and should be replaced by fits to real EEG/MEG/fMRI, behavioral, and artificial-agent data in future work. Seventh, the topological two-kink extension is an optional bridge; its failure would not falsify the core capacity-deficit theory of reportable access. Eighth, the artificial-agent benchmark protocol is proposed but not implemented as a public high-dimensional benchmark in this paper. Such benchmarks are natural future tests of the theory, but they require separate disclosure of environment design, agent architecture, seed sweeps, code, and evaluation criteria. The current paper therefore treats artificial-agent benchmarking as a proposed validation program rather than as a completed public result. Ninth, the paper does not report a full high-dimensional artificial-agent benchmark; the reduced simulations are low-dimensional and illustrative.

## XIII. CONCLUSION

The framework predicts pruning thresholds, nonmonotonic self-model coherence (via the tradeoff between functional retained structure $m$ and nonfunctional residue $\rho$), rescue-window closure, leading-covariance early warnings, anesthesia emergence ordering, sleep-like maintenance requirements, and a thermodynamic lower bound for irreversible pruning. The Landauer term is only the informational heat floor; total maintenance cost is dominated by biological or physical overheads. Reduced simulations support the internal coherence of the FDS reportability dynamics by illustrating pruning thresholds, rescue-window closure, and early-warning behavior near access collapse. The next validation step is a high-dimensional in-machina benchmark in which finite artificial agents operate under nonstationary input, bounded memory, prediction-error accumulation, and selective pruning. Such benchmarks should test whether active pruning maintains reportability-like access more robustly than no pruning, passive decay, or random deletion. The framework does not claim to solve the hard problem of phenomenal consciousness. It converts a part of the consciousness problem into a testable maintenance problem for finite distinction systems, with explicit falsification conditions, differential predictions against competing theories, and an artificial-agent benchmark protocol.

### Appendix A: Variational derivation sketch

The constrained objective in Eq. (4) induces two coupled optimization processes: online inference and maintenance. Online inference updates $q_\phi$ to reduce $\mathcal{F}_{\text{var}}$ while preserving task relevance. Maintenance updates the internal representation to keep $I(o;z)$ and update cost within budget. When $\mathcal{F}_{\text{var}}$ exceeds what can be resolved within $C_{\text{eff}}$, the residual load is deposited as $\rho$. Active pruning is the gradient-flow component that reduces $\rho$ at a finite rate and at finite energetic cost. Saturation of pruning follows from finite throughput of the maintenance channel.

### Appendix B: Network early-warning signal

Linearizing the high-dimensional access network near a fixed point gives

$$\dot{\boldsymbol{x}} = J\boldsymbol{x} + \boldsymbol{\xi}(t), \tag{B1}$$

where $J$ is the effective Jacobian and $\boldsymbol{\xi}$ is noise with covariance $Q$. The stationary covariance $\Sigma$ satisfies the Lyapunov equation

$$J\Sigma + \Sigma J^T + Q = 0. \tag{B2}$$

If the leading eigenvalue of $J$ approaches zero from below, the leading eigenvalue of $\Sigma$ diverges. Thus, in EEG/MEG/fMRI data, the preferred early-warning statistic is not scalar variance but the leading covariance eigenmode or the recovery time of the dominant access mode.

### Appendix C: Phenomenal geometry

The paper does not answer why there is something it is like to be a system. It does, however, make a structural claim about the geometry of reportable experience. Let the agent compress high-dimensional input $x$ into a low-dimensional access state $z$ through $q_\phi(z|x)$. The reportable similarity structure over experiences is not the Euclidean geometry of $x$ but the information geometry induced by $q_\phi$. A natural metric is the Fisher-Rao metric

$$g_{ab}(z) = \mathbb{E}_{x\sim p(x|z)}\left[\partial_a \ln p(x|z)\, \partial_b \ln p(x|z)\right]. \tag{C1}$$

Under capacity deficit, active pruning selects a lower-dimensional manifold that preserves task relevance at minimal cost. The theory predicts that reported qualitative similarity should track geodesic distances on this compressed manifold, while salience or intensity should track local distortion, curvature, and energetic cost.

### Appendix D: Claim status and consequence of failure

### Appendix E: Operational proxy map

The following table provides an operational mapping from model variables to candidate behavioral and neural proxies.

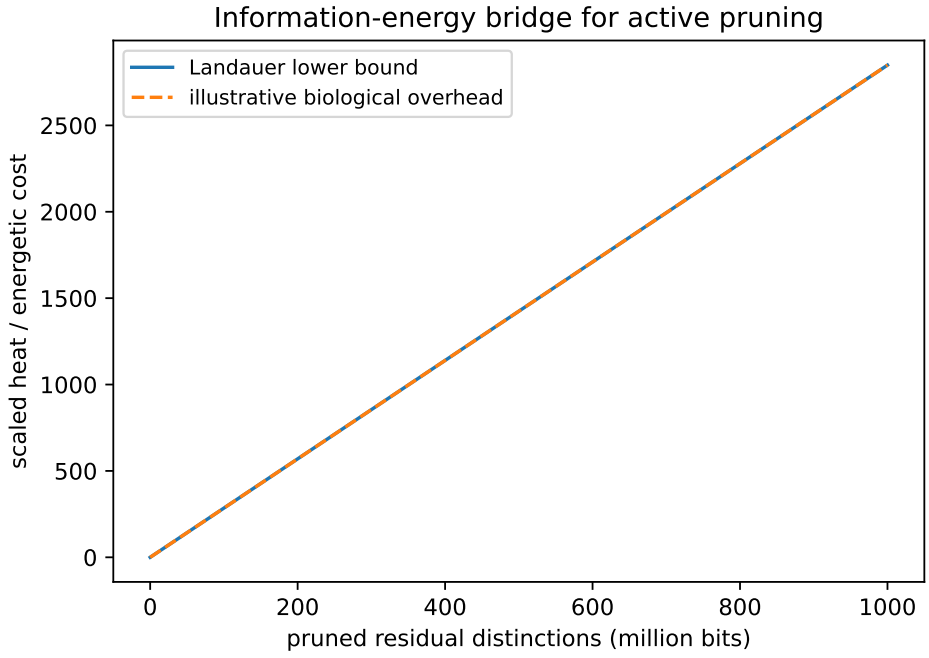


FIG. 7. Information-energy bridge. Irreversible pruning of residual distinctions implies a Landauer lower bound on heat dissipation, though biological overheads will dominate the bound. The prediction is not exact equality; it is that successful pruning under overload should not be energetically silent.

TABLE V. Claim status and consequence of failure.

| Claim | Current support | Demotion condition | Consequence |
|---|---|---|---|
| Pruning threshold | FDS capacity-deficit model + reduced simulations | pruning helps but no threshold | threshold claim weakened |
| Network topology bridge | access-graph model + spectral reduced simulation | no topology/access relation | limit to scalar load model |
| Saddle-node reduction | mean-field normal form | transition is Hopf/noise/percolation only | replace fold mechanism |
| Early warnings | normal form + covariance theory | no slowing near quasi-static threshold | remove EWS prediction |
| Reportability window | $m/\rho$ decomposition + reduced model | no non-monotonic self-model proxy | revise $\Psi$ interpretation |
| Energy bridge | Landauer lower bound | no energetic cost in isolated pruning | restrict to computational bookkeeping |
| Anesthesia ordering | reduced simulation prediction | reversed robust temporal order | revise causal-loop/access hierarchy |
| Artificial-agent sleep | proposed benchmark protocol | no pruning benefit under bounded resources | restrict artificial-consciousness claim |

TABLE VI. Operational proxy map. The table is a benchmark plan, not a claim of completed validation.

| Model variable | Behavioral proxy | Neural / physiological proxy | Candidate paradigm |
|---|---|---|---|
| $\rho$ residue/load | intrusion errors, lapse rate, perseveration | residual prediction error, EEG variability, theta/load markers | task switching, sleep deprivation, attentional blink |
| $S$ active pruning | distractor suppression, recovery speed, forgetting/reconsolidation | alpha/beta inhibition, slow-wave downscaling, LTD-like markers | sleep, TMS perturbation, cognitive control tasks |
| $C_{acc}$ effective capacity | working-memory span, dual-task capacity | PCI, Lempel-Ziv complexity, effective connectivity | masking, anesthesia, DOC |
| $G$ global access | report/no-report, confidence, flexible use | P3b/ignition, thalamocortical connectivity, leading network eigenmode | masking, attentional blink, anesthesia |
| $A$ causal-loop availability | motor responsiveness, command following | frontomotor effective connectivity, EMG responsiveness | anesthesia emergence, DOC |
| $\mathcal{R}$ reportability | verbal/manual report, confidence-calibrated access | report-linked ignition and complexity | masking, no-report variants |
| $I_{\text{self}}$ self-model coherence | agency/self-continuity reports, metacognition | DMN-FPN coupling, self-referential network dynamics | anesthesia emergence, dissociation, sleep onset |
| $\Delta Q$ pruning cost | fatigue, effort, recovery cost | BOLD/$CMRO_2$, glucose use, local heat in artificial devices | overload and active forgetting |

[1] B. J. Baars, *A Cognitive Theory of Consciousness* (Cambridge University Press, 1988).

[2] S. Dehaene and J.-P. Changeux, Experimental and theoretical approaches to conscious processing, Neuron **70**, 200–227 (2011). DOI: 10.1016/j.neuron.2011.03.018.

[3] G. Tononi, An information integration theory of consciousness, BMC Neuroscience **5**, 42 (2004). DOI: 10.1186/1471-2202-5-42.

[4] M. Oizumi, L. Albantakis, and G. Tononi, From the phenomenology to the mechanisms of consciousness: Integrated Information Theory 3.0, PLoS Computational Biology **10**, e1003588 (2014). DOI: 10.1371/journal.pcbi.1003588.

[5] K. Friston, The free-energy principle: a unified brain theory?, Nature Reviews Neuroscience **11**, 127–138 (2010). DOI: 10.1038/nrn2787.

[6] K. Friston, A free energy principle for biological systems, Entropy **14**, 2100–2121 (2012). DOI: 10.3390/e14112100.

[7] A. G. Casali et al., A theoretically based index of consciousness independent of sensory processing and behavior, Science Translational Medicine **5**, 198ra105 (2013). DOI: 10.1126/scitranslmed.3006294.

[8] M. Massimini et al., Breakdown of cortical effective connectivity during sleep, Science **309**, 2228–2232 (2005). DOI: 10.1126/science.1117256.

[9] N. Tishby, F. C. Pereira, and W. Bialek, The information bottleneck method, Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing, 368–377 (1999).

[10] R. Landauer, Irreversibility and heat generation in the computing process, IBM Journal of Research and Development **5**, 183–191 (1961). DOI: 10.1147/rd.53.0183.

[11] A. Berut et al., Experimental verification of Landauer's principle linking information and thermodynamics, Nature **483**, 187–189 (2012). DOI: 10.1038/nature10872.

[12] P. L. Purdon et al., Electroencephalogram signatures of loss and recovery of consciousness from propofol, Proceedings of the National Academy of Sciences USA **110**, E1142–E1151 (2013). DOI: 10.1073/pnas.1221180110.

[13] M. Schartner et al., Complexity of multi-dimensional spontaneous EEG decreases during propofol induced general anaesthesia, PLoS ONE **10**, e0133532 (2015). DOI: 10.1371/journal.pone.0133532.

[14] G. Tononi and C. Cirelli, Sleep and the price of plasticity: from synaptic and cellular homeostasis to memory consolidation and integration, Neuron **81**, 12–34 (2014). DOI: 10.1016/j.neuron.2013.12.025.

[15] C. B. Saper, T. E. Scammell, and J. Lu, Hypothalamic regulation of sleep and circadian rhythms, Nature **437**, 1257–1263 (2005). DOI: 10.1038/nature04284.

[16] C. Sergent and S. Dehaene, Is consciousness a gradual phenomenon? Evidence for an all-or-none bifurcation during the attentional blink, Psychological Science **15**, 720–728 (2004). DOI: 10.1111/j.0956-7976.2004.00748.x.

[17] P. Bak, C. Tang, and K. Wiesenfeld, Self-organized criticality: an explanation of 1/f noise, Physical Review Letters **59**, 381–384 (1987). DOI: 10.1103/PhysRevLett.59.381.

[18] J. M. Beggs and N. Plenz, Neuronal avalanches in neocortical circuits, Journal of Neuroscience **23**, 11167–11177 (2003). DOI: 10.1523/JNEUROSCI.23-35-11167.2003.

[19] M. Scheffer et al., Early-warning signals for critical transitions, Nature **461**, 53–59 (2009). DOI: 10.1038/nature08227.

[20] C. Kuehn, A mathematical framework for critical transitions: bifurcations, fast-slow systems and stochastic dynamics, Physica D **240**, 1020–1035 (2011). DOI: 10.1016/j.physd.2011.02.012.

[21] S. H. Strogatz, *Nonlinear Dynamics and Chaos*, 2nd ed. (Westview Press, 2015).

[22] O. Sporns, *Networks of the Brain* (MIT Press, 2011).

[23] Y. Wu, Active pruning controls boundary persistence in protocell-like systems: saddle-node attractor loss, stochastic early warnings, multiscale simulations, and wet-lab benchmark predictions, manuscript submitted (2026).

[24] C. E. Shannon, A mathematical theory of communication, Bell System Technical Journal **27**, 379–423 (1948). DOI: 10.1002/j.1538-7305.1948.tb01338.x.

[25] N. Cowan, The magical number 4 in short-term memory: a reconsideration of mental storage capacity, Behavioral and Brain Sciences **24**, 87–114 (2001). DOI: 10.1017/S0140525X01003922.

[26] L. Hasher and R. T. Zacks, Working memory, comprehension, and aging: a review and a new view, in G. H. Bower (ed.), The Psychology of Learning and Motivation **22**, 193–225 (1988). DOI: 10.1016/S0079-7421(08)60041-9.

[27] M. Overgaard and K. Sandberg, The perceptual awareness scale: a review and meta-analysis, Neuroscience of Consciousness **2021**, niab002 (2021). DOI: 10.1093/nc/niab002.

[28] J. M. Fawcett, T. L. Taylor, E. Megla, et al., Active intentional and unintentional forgetting in the laboratory and everyday life, Nature Reviews Psychology **3**, 652–664 (2024). DOI: 10.1038/s44159-024-00352-7.

[29] A. J. Eisen, L. Kozachkov, A. M. Bastos, et al., Propofol anesthesia destabilizes neural dynamics across cortex, Neuron **112**, 2799–2813.e9 (2024). DOI: 10.1016/j.neuron.2024.06.011.

[30] Y. S. Xiong, J. A. Donoghue, M. Lundqvist, et al., Propofol-mediated loss of consciousness disrupts predictive routing and local field phase modulation of neural activity, Proceedings of the National Academy of Sciences **121**, e2315160121 (2024). DOI: 10.1073/pnas.2315160121.

[31] Z. Zhang, Q. Dai, X. Bo, C. Ma, R. Li, C. Xu, J. Zhu, Z. Dong, and J. Wen, A survey on the memory mechanism of large language model-based agents, ACM Transactions on Information Systems **43**(6), Article 155 (2025). DOI: 10.1145/3748302.