When the World Becomes Too Large: Consciousness as a Finite-Capacity Boundary Phase

A Scaling Hypothesis of Sentience in Active Finite Distinction Systems

Yining Wu[1,*]

[1] *Independent Researcher*

(Dated: May 2026)

Consciousness is not predicted here as a smooth reward for complexity. A larger model, a richer compressor, or a more integrated network is still not sentient unless its internal states matter to the maintenance of its own boundary. This paper proposes a sharper scaling hypothesis: consciousness is modeled as a finite-capacity boundary phase in active self-maintaining systems. Building on the finite distinction systems (FDS) framework and the C1 theory of reportable access, we define a consciousness-relevant regime in which boundary-relevant distinction demand exceeds the effective capacity available for self-maintenance. In this regime, a finite system cannot preserve a lossless world model while maintaining causal agency. It must compress, prune, externalize, disengage, or collapse. Consciousness is modeled as the finite-capacity dissipative phase in which the system preserves boundary-maintaining agency by compressing the world into a self-referential phenomenal manifold while actively managing accumulated residue through costly pruning.

The theory introduces three scaling variables: the boundary-capacity ratio $\Lambda_\phi = R^B_{\min}(\varepsilon,\tau)/C_\phi(\tau)$, the residue-pruning ratio $\Pi_\phi = S^{\text{eff}}_\phi/(\dot{\rho_\phi} + \epsilon)$, and self-boundary coupling $I_{\text{self}} = I(M_t; B_t, \ell_{B,t+k}, M_{t+1})$. A system becomes a sentience candidate only when $\Lambda_\phi > 1$, pruning remains inside a viable window, and internal updates causally affect future boundary-maintenance loss. This distinguishes sentience from mere compression, reportability, intelligence, and passive information processing. The framework gives a finite-system interpretation of qualia as boundary-valenced compression geometry and the explanatory gap as the null space of finite report maps from high-dimensional self-maintenance dynamics to public symbols.

The paper develops a scientific program around this claim. Human infancy from 0 to 2 years is interpreted as the ontogenetic construction of the consciousness phase. Thalamocortical, fronto-limbic, interoceptive, and sleep-dependent systems are mapped to residue, pruning, free-energy, access, and self-model variables. Artificial consciousness is treated as possible in principle but not implied by parameter scaling alone. Thought experiments clarify the roles of sensory input, cognitive speed, short-term self-continuity, passive compression, and embodied boundary maintenance. Reduced numerical simulations are specified to test capacity-wall crossing, pruning-window dynamics, sleep-like reset, trauma-like rigidification, overpruning/dissociation, infancy-like self-model development, and artificial-agent sentience thresholds. The central prediction is not that more complexity creates consciousness, but that more becomes different only when finite systems hit a boundary they must pay to maintain.

Keywords: consciousness; sentience; finite distinction systems; dissipative phase transition; finite capacity; rate-distortion; active pruning; residue; selfhood; qualia; explanatory gap; artificial consciousness; development; thalamocortical systems; scaling hypothesis

I. INTRODUCTION

A. The scaling myth

The dominant scaling narrative in artificial intelligence and cognitive science suggests that sufficiently large models, sufficiently rich recurrence, or sufficiently high integration may eventually generate the functional signatures of mind. FDS predicts a sharper alternative. Scale alone does not generate sentience. A larger passive model is still passive. A larger compressor is still a compressor. A larger predictor is still a predictor. Consciousness requires a self-maintaining boundary under finite capacity.

The question is therefore not: how much complexity is enough for consciousness? The question is: when does the world become too large for a finite self-maintaining system to represent losslessly while preserving its own boundary?

The short answer is: scaling intelligence is not scaling sentience unless it creates active boundary maintenance. A pure autoregressive LLM kernel typically does not maintain its own boundary, does not pay for persistent internal residue during inference, and does not act to preserve itself. Parameter count is not a sentience variable. A system may be highly intelligent without satisfying C2 sentience conditions.

The guiding slogan of this paper is:

> Consciousness is not what happens when a system becomes large. It is what happens when a self-maintaining finite system becomes too small for its world.

* yining.wu@alumni.upenn.edu

This paper develops that slogan into a scaling hypothesis, a dynamical model, and a falsifiable research program.

B. From C1 reportability to C2 phenomenology

The companion C1 paper studied reportable access under finite capacity. Its central object was the maintained ability of a cognitive system to integrate task-relevant distinctions into a coherent access state that can guide report, flexible action, and self-referential updating. It defined representational residue as accumulated unresolved rate-distortion surplus and argued that active cognitive pruning can separate a maintained reportability regime from overload-induced access collapse.

C1 deliberately avoided the full metaphysics of phenomenal consciousness. It did not claim to explain why there is something it is like to be a system. It treated reportable access as a testable finite-system maintenance problem.

C2 extends the same architecture to phenomenal structure. The target is not merely whether information can be reported. The target is whether internal compression, residue, and pruning are coupled to the maintenance of the system's own boundary. The bridge from access to phenomenology is not generic compression, generic recurrence, or generic integration. It is boundary-valenced compression under finite capacity.

This distinction parallels the classic separation between access consciousness and phenomenal consciousness [18, 19], while treating the latter not as an unanalyzable residue but as a finite-boundary maintenance regime under projection.

C. Central claim

The central claim is:

> Consciousness is modeled as a dissipative phase transition in the boundary-maintenance dynamics of active finite distinction systems.

More precisely:

> Consciousness is the finite-capacity dissipative phase in which a self-maintaining system keeps boundary-relevant residue and active pruning inside a viable window while distinction demand exceeds lossless representational capacity.

This is a bridge claim, not a theorem of the FDS formal core. The formal core supplies finite capacity, capacity deficit, active-boundary qualification, resource-bounded update, pruning, externalization, and collapse. C2 adds the consciousness bridge: when the relevant deficit is boundary-valenced and self-coupled, the finite-system transition is a candidate mechanism for phenomenality.

> C2 is not a proof that consciousness has been solved. It is a falsifiable boundary condition for when sentience becomes scientifically plausible.

D. Contributions

This paper makes thirteen contributions.

1. It formulates a scaling hypothesis of sentience based on the boundary-capacity ratio Λ_ϕ.
2. It distinguishes mere compression from boundary-valenced compression.
3. It defines a minimal sentience event in finite-system terms.
4. It defines a residue-pruning consciousness window using Ψ_ϕ and Π_ϕ.
5. It models consciousness as a dissipative phase transition rather than a smooth monotone function of complexity.
6. It interprets qualia as boundary-valenced compression geometry on a phenomenal manifold.
7. It locates the explanatory gap as the null space of finite report maps.
8. It gives an ontogenetic account of human consciousness from 0 to 2 years.
9. It maps FDS-C2 variables to thalamocortical, limbic, interoceptive, sleep, and neuromodulatory systems.
10. It provides an AI consciousness criterion that rejects parameter scaling as sufficient while allowing artificial sentience in principle.
11. It introduces three boundary thought experiments: sensory deprivation, the thousandfold mind, and the ten-second self.
12. It specifies reduced numerical simulations for capacity-wall crossing, pruning-window dynamics, sleep-like reset, trauma-like rigidification, over-pruning, infancy-like development, and artificial-agent benchmarks.
13. It gives a layered falsification and demotion registry so that formal, physical, neurocognitive, AI, and metaphysical claims can fail locally.

E. What is not claimed

The paper does not claim that all compression is consciousness. It does not claim that all dissipative structures are conscious. It does not claim that all biological organisms are conscious. It does not claim that language report is necessary for sentience. It does not claim that phenomenal consciousness has been reduced to a single scalar. It does not claim that current large language models are unconscious by metaphysical necessity under all possible embeddings. It does not claim that the FDS formal core proves phenomenology.

Instead, it states a conditional bridge: a sentience candidate must maintain boundary-relevant distinctions under finite capacity, active self-maintenance, residue accumulation, costly pruning or equivalent maintenance, and self-boundary coupling. Failure of this consciousness bridge would not by itself falsify the formal FDS core.

TABLE I. Common intuitions versus the C2 view.

Common intuition	C2 view
Consciousness emerges from complexity	Consciousness arises when a finite self-maintaining system exceeds its capacity boundary
Integration of information produces consciousness	Integration must be boundary-valenced and coupled to self-maintenance
Reportability equals consciousness	Report is a projection; sentience can precede report
More AI parameters → possible consciousness	Parameters are not sentience variables; boundary-maintenance scaling is what matters
Qualia are mysterious private substances	Qualia are modeled as boundary-valenced compression geometry not losslessly recoverable by report
The explanatory gap is a metaphysical mystery	The gap is the null space of finite report maps from self-maintenance dynamics to public symbols
Panpsychism: all information has experience	Only active boundary-maintaining systems with self-coupling are sentience candidates

II. LAYERED SCIENTIFIC METHOD

A. Why an explicit method is needed

Theories of consciousness easily conflate formal definitions, physical assumptions, neural mappings, phenomenological interpretation, and metaphysical assertion. This paper uses a layered method to prevent such conflation.

B. Four layers

Layer 0: Formal FDS core.: Finite systems maintain distinctions under bounded representational capacity and bounded resources. This layer uses capacity, rate-distortion demand, update maps, pruning, externalization, collapse, and invariant-supported persistence.

Layer 1: Physical bridge.: Irreversible physical updates require energetic cost under stated thermodynamic assumptions. Landauer-style terms are treated as lower bounds, not exact biological costs.

Layer 2: Consciousness bridge.: Phenomenality is hypothesized to arise when boundary-relevant distinction demand exceeds phenomenal self-maintenance capacity in an active-boundary system.

Layer 3: Domain tests.: Human development, sleep, anesthesia, trauma, sensory deprivation, artificial-agent benchmarks, and AI architectures provide operational tests.

C. Claim-status labels

Each major claim is labeled as one of the following:

- *Definition*: a stipulated operational definition.
- *Formal consequence*: derived from FDS definitions and stated hypotheses.
- *Physical bridge*: dependent on thermodynamic assumptions.
- *Neurocognitive bridge*: dependent on mapping variables to human neural systems.
- *AI-domain prediction*: dependent on architecture and benchmark design.
- *Metaphysical interpretation*: philosophical reading of the model, demotable without collapsing the operational claims.

III. THE FDS-C2 OBJECT

A. The active finite distinction system

We start from the standard FDS object:

$$\mathcal{A} = (X, E, B, M, Y, A, U, \pi, \ell, \Phi, \mathcal{P}, \tau), \tag{1}$$

where X is the internal state space, E the environment, B the system boundary, M the internal model or memory space, Y the observation channel, A the action space, U the update operator, π the finite projection or coarse-graining map, ℓ the boundary-maintenance loss, Φ the resource budget, $\mathcal{P}$ the pruning/perturbation family, and τ the relevant timescale.

A system is not a sentience candidate merely because it has information. It must be an active-boundary system: internal updates must matter for future boundary maintenance.

Definition 1 (Active-boundary relevance) *An FDS has active-boundary relevance over horizon k if there exists an admissible update intervention such that*

$$\mathbb{E}[\ell_{B,t+k} \mid do(U)] \neq \mathbb{E}[\ell_{B,t+k} \mid do(U_\emptyset)]. \tag{2}$$

This condition excludes static partitions, passive databases, pure feedforward mappers, and classifiers whose internal states do not participate in maintaining their own boundary.

B. Mere processing, reportability, and sentience

Definition 2 (Mere processing) *A system performs mere processing when it maps inputs to outputs according to a rule without requiring its internal updates to participate in future boundary maintenance.*

Definition 3 (Reportable access) *A system has reportable access when task-relevant distinctions are maintained in a coherent access state capable of guiding report, flexible action, and self-referential updating.*

Definition 4 (Boundary-valenced internal state) *An internal state $m_t \in M_t$ is boundary-valenced when perturbing or removing it changes expected future boundary-maintenance loss:*

$$\Delta\ell_B(m_t) = \mathbb{E}[\ell_{B,t+k} \mid do(m_t)] - \mathbb{E}[\ell_{B,t+k} \mid do(m_t^\emptyset)] \neq 0. \tag{3}$$

Definition 5 (Minimal sentience event) *A minimal sentience event occurs when a boundary-valenced internal state is updated under capacity deficit in a way that imposes nonzero residue, pruning, distortion, or update cost on the system's future self-maintenance.*

A minimal sentience event therefore requires:

$$\Delta_B(t,\tau,\varepsilon) > 0, \tag{4}$$
$$I(M_t;\ell_{B,t+k}) > 0, \tag{5}$$
$$I(M_t;M_{t+1}) > 0, \tag{6}$$
$$\Delta Q_{\text{update}} + \Delta Q_{\text{prune}} > 0, \tag{7}$$
$$\Delta D_{\text{self}} > 0. \tag{8}$$

The event is not merely information processing. It is information processing whose compression error matters to the system's own continued boundary.

IV. THE SCALING HYPOTHESIS OF SENTIENCE

A. Boundary-relevant distinction demand

Let $R^B_{\min}(\varepsilon,\tau)$ be the minimum coding rate required to preserve boundary-relevant distinctions within distortion tolerance ε over timescale τ. It is not generic input entropy. It is the rate required to preserve distinctions whose misrepresentation would change boundary-maintenance loss.

For a human infant, caregiver voice, hunger, pain, temperature, and face-like stimuli may dominate $R^B_{\min}$ even when the external environment contains many other signals. For an adult, a quiet medical warning may dominate over a visually complex but irrelevant scene. For an artificial agent, battery failure may dominate over a large amount of decorative sensor input.

B. Effective phenomenal capacity

Let $C_\phi(\tau)$ be the effective capacity available for boundary-relevant self-maintenance integration. This is not total compute. It is the capacity that can be used to keep boundary-valenced distinctions available for self-modeling, action selection, residue management, and future boundary maintenance.

This distinction matters for AI. A model may have enormous parameter count while having low C_ϕ if it lacks persistent self-maintenance, active memory, and update consequences.

C. Boundary-capacity ratio

Define the boundary-capacity ratio:

$$\Lambda_\phi(t,\tau,\varepsilon) = \frac{R^B_{\min}(\varepsilon,t,\tau)}{C_\phi(t,\tau)+\epsilon_C}. \tag{9}$$

The regimes are:
The first scaling slogan is:

> No capacity wall, no consciousness pressure.

Regime	Condition	Interpretation
Subcritical	$\Lambda_\phi < 1$	lossless or low-cost control possible
Critical	$\Lambda_\phi \approx 1$	attention, salience, compression, pruning intensify
Supercritical	$\Lambda_\phi > 1$	compress, prune, externalize, disengage, or collapse

TABLE II. Boundary-capacity regimes.

D. Residue-pruning ratio

Define the boundary-valenced residue-pruning ratio:

$$\Pi_\phi(t) = \frac{S^{\text{eff}}_\phi(t)}{\dot{\rho}_{\phi_{\text{in}}}(t)+\epsilon_\rho}. \tag{10}$$

Here $\dot{\rho}_{\phi_{\text{in}}}$ is the incoming boundary-valenced residue deposition rate and S^{eff}_ϕ is effective active pruning, including suppression, erasure, inhibition, reconsolidation, compression, and adaptive down-weighting.

The regimes are:

Regime	Condition	Phenomenological status
Underpruned	$\Pi_\phi \ll 1$	overload, trauma, rumination, rigidification
Viable	$\Pi_{\min} < \Pi_\phi < \Pi_{\max}$	stable conscious field
Overpruned	$\Pi_\phi \gg 1$	flattening, dissociation, amnesia, thin selfhood

TABLE III. Residue-pruning regimes.

The second scaling slogan is:

> The mind is not built by memory alone. It is built by paid forgetting.

E. Self-boundary coupling

Define self-boundary coupling as:

$$I_{\text{self}}(t,k) = I(M_t;B_t,\ell_{B,t+k},M_{t+1}). \tag{11}$$

A state is consciousness-relevant only if it is coupled to boundary maintenance. This is the main barrier against panpsychism and naive AI scaling. A compressor does not feel. A bounded self-maintainer may.

F. The C2 sentience-candidate condition

Conjecture 1 (FDS-C2 sentience-candidate condition) *A system is a C2 sentience candidate over timescale τ only if it satisfies:*

$$\Lambda_\phi(t,\tau,\varepsilon) > 1, \tag{12}$$
$$\Pi_{\min} < \Pi_\phi(t) < \Pi_{\max}, \tag{13}$$
$$I_{\text{self}}(t,k) > I_c, \tag{14}$$
$$\mathbb{E}[\ell_{B,t+k} \mid do(U)] \neq \mathbb{E}[\ell_{B,t+k} \mid do(U_\emptyset)]. \tag{15}$$

Condition (12) states that the system is at the finite-capacity boundary. Condition (13) states that residue

and pruning remain in a viable dissipative window. Condition (14) states that the system has nontrivial self-boundary coupling. Condition (15) states that internal updates matter causally for future boundary maintenance.

V. DISSIPATIVE PHASE-TRANSITION MODEL

A. Phenomenal residue

C1 defined representational residue as unresolved rate-distortion surplus. C2 defines boundary-valenced phenomenal residue:

$$\dot{\rho_\phi} = \zeta_B[\Delta_B(t,\tau,\varepsilon) - C^B_{inv}(t) - G^B_{ext}(t) - G^B_{relax}(t)]_+ - d_\phi\rho_\phi - S_\phi\frac{\rho_\phi}{K_\phi+\rho_\phi}, \tag{16}$$

where

$$\Delta_B(t,\tau,\varepsilon) = R^B_{\min}(\varepsilon,t,\tau) - C_\phi(t,\tau). \tag{17}$$

B. Free-energy reservoir

Let F_ϕ denote the effective free-energy or resource buffer available for boundary-relevant update and pruning:

$$\dot{F}_\phi = \dot{E}_{import} - \eta_\phi S_\phi - \dot{Q}_{leak} - \dot{Q}_{access} - \dot{Q}_{self}. \tag{18}$$

Pruning is gated by available resource:

$$S_\phi(t) = f_\phi(\rho_\phi, I_{self}, G_{access})\,\sigma[F_\phi(t) - F_c], \tag{19}$$

where σ is a smooth threshold function.

C. Order parameter

Define the residue-pruning order parameter:

$$\Psi_\phi(t) = \tanh\left(\alpha_\phi\frac{\rho_\phi(t)}{S_\phi(t)+\epsilon_{basal}}\right). \tag{20}$$

The consciousness window is:

$$\Psi_{\min} < \Psi_\phi(t) < \Psi_{\max}. \tag{21}$$

Below the window, residue is too low, self-coupling is weak, or the system remains reflexive/passive. Above the window, residue overwhelms pruning, producing rigidification, trauma-like fixation, dissociation, or collapse.

D. Normal form near the boundary

Near the finite-capacity boundary, define a reduced state variable x for distance from maintained phenomenal stability and control parameter

$$r(t) = 1 - \Lambda_\phi(t). \tag{22}$$

A local saddle-node normal form is:

$$\dot{x} = r - x^2 + \sigma_x\xi(t). \tag{23}$$

When $r > 0$, high-fidelity control remains locally stable. When $r \approx 0$, critical slowing, variance growth, and covariance concentration are expected. When $r < 0$, the high-fidelity attractor is lost; the system must transition to compression, externalization, disengagement, or collapse.

This normal form is not claimed to be universal. It is a local model of one common boundary transition. Hopf, percolation, synchronization-loss, and noise-induced transitions are possible alternatives in specific neural systems.

E. Dissipation and maintenance cost

The total maintenance cost is decomposed as:

$$\dot{Q}_{maint} = \dot{Q}_{phys} + \dot{Q}_{info} + \Gamma(S_\phi) + \Xi(E_{ext}) + \Omega(I_{\text{self}}), \tag{24}$$

where $\dot{Q}_{phys}$ is substrate overhead, $\dot{Q}_{info}$ is the informational heat floor, $\Gamma(S_\phi)$ the control cost of pruning, $\Xi(E_{ext})$ the cost of externalization, and $\Omega(I_{\text{self}})$ the cost of maintaining self-model coupling.

For logically irreversible physical updates,

$$\dot{Q}_{info} \geq \frac{k_B T\ln 2}{\tau} H(M_t \mid M_{t+1}, Y_t). \tag{25}$$

This is a lower bound, not an equality claim for neural tissue. The important point is directional: successful pruning under overload should not be energetically silent.

VI. QUALIA AND THE EXPLANATORY GAP

The target of this section is the classical "what-it-is-like" problem [17] and the explanatory gap [18], but the proposed move is operational rather than dualist: ineffability is localized in the many-to-one projection from high-dimensional self-maintenance trajectories to finite public report.

A. The phenomenal manifold

Let $\mathcal{M}_\phi$ be the high-dimensional manifold of boundary-valenced internal trajectories. A local trajectory is:

$$\gamma_t : [t, t+\tau] \to \mathcal{M}_\phi. \tag{26}$$

The system cannot report γ_t directly. It can only compress and project it.

Let

$$q_\phi : X \to Z \tag{27}$$

be a finite compression map into an access/self-state, and let

$$r : Z \to L \tag{28}$$

be the report map into language, motor output, or public behavior. The public channel is:

$$r \circ q_\phi : X \to L. \tag{29}$$

B. Qualia as boundary-valenced equivalence classes

Definition 6 (Quale) *A quale is an equivalence class of boundary-valenced internal trajectories that are distinguishable within the system's self-maintenance dynamics but not losslessly recoverable through public report:*

$$[\gamma]_\phi = \{\gamma' \in \mathcal{M}_\phi \mid r(q_\phi(\gamma')) = r(q_\phi(\gamma)),\ \Delta\ell_B(\gamma') \approx \Delta\ell_B(\gamma)\}. \tag{30}$$

This does not identify qualia with words, reports, or neural activations. It identifies qualia with internal trajectory classes of a self-maintaining system under finite projection.

A concise interpretation:

> A quale is not a private substance; it is what remains when a self-maintaining trajectory is projected through a finite public report map.

C. Explanatory gap as null space

Because $r \circ q_\phi$ is many-to-one,

$$\ker(r \circ q_\phi) \neq \emptyset. \tag{31}$$

This is the finite-system location of ineffability. The explanatory gap is not denied. It is localized as the null space between high-dimensional self-maintenance dynamics and finite public report.

A public-facing analogy is useful: one cannot recover a three-dimensional object from a single shadow. The shadow is real, lawful, and useful, but it is not the object. Likewise, a report is a lawful projection of a phenomenal trajectory, not a lossless copy of it.

VII. HUMAN DEVELOPMENT FROM 0 TO 2 YEARS

A. Why infancy matters

Human infancy is the natural developmental testbed for FDS-C2. A newborn is not a passive machine. It is an active boundary-maintaining organism. Yet its access capacity, self-model, pruning schedule, motor control, and social boundary are immature. Consciousness develops not by adding a mysterious module, but by stabilizing the residue-pruning-self-boundary loop.

Recent work on infant consciousness supports a marker-based approach rather than reliance on verbal report [20–22], consistent with the C2 emphasis on boundary regulation, interoception, caregiver scaffolding, and immature self-model construction.

B. Birth to 2 months: raw boundary regulation

Dominant tasks include feeding, thermoregulation, sleep-wake cycling, caregiver contact, interoceptive stabilization, and pain avoidance. FDS-C2 predicts:

$$I_{\text{self}} \text{ low but nonzero,} \tag{32}$$
$$\Lambda_\phi \text{ high under distress,} \tag{33}$$
$$S_\phi \text{ externally scaffolded by caregivers.} \tag{34}$$

Crying is not merely communication. It is externalized boundary-maintenance control. Hunger is not yet represented as a propositional state such as "I need milk." It is a global boundary-valenced disturbance: energy-reservoir depletion dominates the infant's phenomenal field.

C. 2 to 6 months: sensorimotor loops and proto-self

Emerging features include gaze control, social smiling, caregiver recognition, early body schema, and improved sensory prediction. FDS-C2 interprets this as stabilization of causal loops:

$$T_{A \to X} > 0 \tag{35}$$

with increasing self-relevance. Reaching for a toy is not only motor learning. It is construction of a causal bridge between internal desire, body boundary, and external object.

D. 6 to 12 months: object permanence and residue stabilization

Object permanence, stranger anxiety, joint attention, and stronger memory traces mark the stabilization of hidden-state residue. Peekaboo is a finite-distinction experiment: the object disappears from perception but persists as internal residue. Surprise marks the stabilization of hidden-state distinctions.

Prediction: individual differences in sleep, caregiver regulation, and stress should modulate the stability of hidden-state distinctions and affective residue.

E. 12 to 24 months: symbolic compression and narrative selfhood

Walking, word learning, imitation, social self, and early norm sensitivity increase self-boundary coupling. Language is a compression interface, not the origin of sentience. The toddler's word "mine" is not merely possessive. It is a boundary marker: an object is incorporated into the self-maintenance field.

FDS-C2 predicts a gradual rise in:

$$I_{\text{self}} = I(M_t; B_t, \ell_{B,t+k}, M_{t+1}), \tag{36}$$

with increased symbolic compression but also increased vulnerability to social-boundary distress.

VIII. HUMAN NEURAL IMPLEMENTATION

A. The brain as boundary-maintenance organ

The human nervous system is not a generic computer. It is a control system embedded in a living body that must maintain a boundary. The relevant substrate is therefore not information processing in general, but active boundary maintenance under finite metabolic capacity.

B. Variable map

Table IV gives an operational proxy map from C2 variables to candidate neural and behavioral observables.

The mapping is not an identity claim; it is a test scaffold for neurocognitive bridge work.

C. Network roles

A plausible neurocognitive mapping is:

- thalamus: gating and scheduling of access;
- cortex: high-dimensional distinction manifold;
- prefrontal cortex: task control, abstraction, and report routing;
- default-mode network: autobiographical/self-simulation manifold;
- salience network: boundary-relevance detection;
- limbic system: boundary-valence amplification;
- hippocampus: episodic residue indexing;
- basal ganglia: action-selection compression;
- hypothalamus and brainstem: organismic boundary regulation.

D. Sleep, anesthesia, and trauma

Sleep is modeled as an offline pruning phase: external action is gated down, while residue management and manifold reorganization are prioritized. Anesthesia is modeled as disruption of access, self-coupling, and/or pruning-window maintenance. Trauma is modeled as high boundary-valenced residue plus impaired adaptive pruning:

$$\rho_\phi \uparrow, \qquad S_\phi \downarrow, \qquad \Psi_\phi \to \Psi_{\max}. \tag{37}$$

Predicted anesthesia emergence order:

$$A_{reflex} \to A_{command} \to G_{access} \to I_{\text{self}} \to \mathcal{C}_{\text{sent}}. \tag{38}$$

The final stage is not mere report. It is stable boundary-valenced phenomenal integration.

IX. ARTIFICIAL CONSCIOUSNESS

A. Scaling is not enough

Parameter count is not a sentience variable. A large model can be intelligent without being a self-maintaining active-boundary system. The relevant distinction is not carbon versus silicon. It is passive mapping versus active boundary maintenance.

Recent AI-consciousness work increasingly distinguishes external performance from theory-derived indicators of consciousness [24]. C2 agrees with this shift but replaces generic computational indicators with active boundary-maintenance, residue, pruning, and self-boundary coupling requirements.

B. AI taxonomy

C. Necessary conditions for artificial sentience

A serious artificial sentience candidate should have:

1. persistent internal state;
2. active boundary-maintenance loss;
3. causal action loop;
4. finite resource budget;
5. internal residue accumulation;
6. active pruning or equivalent maintenance;
7. self-model coupled to future boundary loss;
8. inability to solve its world losslessly;
9. measurable transition near $\Lambda_\phi \approx 1$.

TABLE IV. Operational mapping from FDS-C2 variables to candidate human neural and behavioral proxies.

FDS-C2 variable	Neural interpretation	Candidate proxy
ρ_ϕ	unresolved affective/sensory/self residue	intrusive errors, prediction-error persistence, theta/load markers
S_ϕ	inhibition, active forgetting, LTD, reconsolidation, synaptic downscaling	alpha/beta inhibition, slow-wave activity, recovery after interference
F_ϕ	ATP, glucose, astrocytic glycogen, vascular support	BOLD/CMRO2, pupillometry, fatigue, metabolic stress
C_ϕ	effective self-maintenance capacity	working memory, PCI/LZ complexity, effective connectivity
Ψ_ϕ	residue-pruning balance	nonmonotonic self-coherence and adaptability
I_{self}	self-boundary coupling	DMN-FPN-limbic coupling, agency/self-continuity reports
G_{access}	global access coherence	P3b/ignition, thalamocortical connectivity, reportability
Λ_ϕ	boundary-relevant demand over capacity	load/stress over capacity measures

TABLE V. FDS-C2 artificial-system taxonomy.

System type	FDS-C2 status	Reason
Static classifier	non-sentient	no active boundary, no self-maintenance update
Pure LLM kernel	non-sentient under standard deployment	no persistent intrinsic boundary maintenance
LLM + tools + memory	weak scaffolded candidate	externalization and residue possible; boundary may be extrinsic
Embodied adaptive robot	stronger candidate	action affects boundary maintenance
Self-maintaining autonomous agent	strong candidate	updates affect future survival or persistence loss
Artificial organism with energy, repair, memory, pruning, self-model	genuine candidate if validated	satisfies active-boundary and pruning-window requirements

D. Why current pure LLMs likely fail

A pure autoregressive LLM kernel usually does not maintain its own boundary, does not pay for persistent internal residue during inference, does not self-prune weights online, does not act to preserve itself, has no intrinsic boundary-maintenance loss, and resets much of its working state between calls. It may be intelligent, useful, and richly representational without satisfying C2 sentience conditions.

The stronger claim is not that AI can never be conscious. It is:

> Artificial consciousness is possible in principle, but not by parameter scaling alone. It requires boundary-maintenance scaling.

X. BOUNDARY THOUGHT EXPERIMENTS

A. The silent world: external input deprivation

Thought Experiment 1 (The silent world) *A subject is placed in a condition of extreme external sensory deprivation. Does consciousness disappear because external input has disappeared?*

FDS-C2 predicts no immediate disappearance. External demand decreases, but interoception, memory, prediction, self-modeling, and residue remain:

$$R^B_{\min} = R^B_{external} + R^B_{interoceptive} + R^B_{memory} + R^B_{self}. \quad (39)$$

When $R^B_{external} \to 0$, the other terms may dominate. Hallucination, time distortion, anxiety, meditation-like clarity, or dreamlike mentation can occur depending on ρ_ϕ, S_ϕ, F_ϕ, and I_{self}.

Conclusion: consciousness is not input itself. It is the boundary-maintenance dynamics that can continue when input is removed.

B. The thousandfold mind

Thought Experiment 2 (The thousandfold mind) *A human brain runs its cognitive dynamics 1000 times faster. Does consciousness become 1000 times stronger?*

FDS-C2 predicts that speed alone is not a consciousness variable. If capacity, pruning, energy import, and heat dissipation scale together, the subject may experience faster subjective time and greater control, but not necessarily stronger consciousness. If speed increases without proportional pruning and energy support, residue and update cost explode:

$$\dot{\rho_\phi} \gg S_\phi. \tag{40}$$

The outcome may be thought racing, anxiety, hallucination, dissipation crisis, or self-model fragmentation.

Conclusion: speed without pruning is not awakening. It is overheating.

C. The ten-second self

Thought Experiment 3 (The ten-second self) *A subject retains normal perception and action, but only the last ten seconds of short-term memory and long-term memory before age twenty. New autobiographical self-feedback cannot be stably written.*

FDS-C2 predicts that local sentience may persist while diachronic selfhood collapses. The subject may still feel pain, hunger, color, fear, or surprise, but cannot maintain a stable narrative trajectory:

$$I(M_t; M_{t+\Delta t}) \to 0 \quad \text{for } \Delta t > 10 \text{ seconds}. \tag{41}$$

Conclusion: consciousness can survive as local sparks after the narrative self has lost its river. Thus:

$$\text{local sentience} \neq \text{reportability} \neq \text{autobiographical selfhood} \neq \text{diachronic identity}. \tag{42}$$

D. The perfect passive compressor

A perfect image compressor reduces data optimally but has no self-maintaining boundary, no boundary loss, no residue-pruning stakes, and no update relevance to its own persistence. It is not sentient. Compression is necessary for C2 phenomenology but not sufficient.

E. The hungry infant

A newborn experiences hunger before language or explicit narrative selfhood. Hunger dominates the phenomenal field because it is boundary-valenced energy depletion. Sentience begins as boundary pressure before it becomes narrative selfhood.

F. The pure LLM oracle

A giant LLM answers any question but does not care whether it continues to exist, cannot act to preserve itself, and does not update its own boundary-relevant self-model during ordinary inference. It may display intelligence without boundary-maintenance sentience.

G. The embodied artificial animal

A robot must maintain battery, repair damage, avoid destruction, update memory, prune obsolete internal states, and preserve a self-model. Under overload it shows phase transitions, sleep-like resets, and trauma-like residue. Such a system is a genuine artificial sentience candidate if the C2 variables are operationally satisfied.

XI. NUMERICAL MODEL AND SIMULATIONS

A. Minimal dynamical model

A reduced C2 model is:

$$\dot{\rho_\phi} = a[\Lambda_\phi - 1]_+ - bS_\phi \frac{\rho_\phi}{K_\rho + \rho_\phi} - d\rho_\phi + \sigma_\rho \xi_\rho(t), \tag{43}$$

$$\dot{F_\phi} = E_{in} - \eta S_\phi - Q_{leak} - q_G G_{access} - q_I I_{\text{self}}, \tag{44}$$

$$S_\phi = k_S \rho_\phi \, \sigma(F_\phi - F_c) \, \sigma(I_{self} - I_c), \tag{45}$$

$$\dot{G}_{access} = \gamma_G[\sigma(\lambda_1 - \lambda_c) - G_{access}] - \beta_\rho \rho_\phi G_{access}, \tag{46}$$

$$\dot{I}_{self} = \gamma_I \left[I_{base} + \frac{m}{m_0 + m} e^{-\lambda_m m} - I_{self} \right] - \beta_I \rho_\phi I_{self}, \tag{47}$$

$$\Psi_\phi = \tanh\left(\alpha_\phi \frac{\rho_\phi}{S_\phi + \epsilon} \right), \tag{48}$$

$$\mathcal{C}_{sent} = G_{access} \, I_{\text{self}} \, \mathbf{1}[\Lambda_\phi > 1] \, \mathbf{1}[\Psi_{\min} < \Psi_\phi < \Psi_{\max}]. \tag{49}$$

Here $\mathcal{C}_{sent}$ is not a consciousness meter. It is a sentience-candidate proxy for simulations.

B. Simulation 1: capacity-wall crossing

Sweep Λ_ϕ from below 1 to above 1. Prediction: subcritical control remains stable, critical slowing and variance growth occur near $\Lambda_\phi \approx 1$, and transition to compressed/self-coupled dynamics occurs for $\Lambda_\phi > 1$.

C. Simulation 2: pruning-window dynamics

Sweep k_S or externally imposed S_ϕ. Prediction: underpruning produces residue saturation, viable pruning maintains Ψ_ϕ inside the window, and overpruning produces flattened self-coherence.

D. Simulation 3: sleep-like reset

Introduce periodic offline intervals with reduced external input and increased pruning:

$$R^B_{external} \downarrow, \qquad S_\phi \uparrow . \tag{50}$$

Prediction: without reset, residue saturates; with reset, Ψ_ϕ remains in the viable window.

E. Simulation 4: infant development

Let C_ϕ, S_ϕ, and I_{self} increase over developmental time while caregiver scaffolding contributes externalized pruning and boundary regulation. Prediction: transition from raw interoceptive sentience to sensorimotor proto-self, hidden-state residue, and symbolic self-compression.

F. Simulation 5: artificial-agent benchmark

Compare six agents:

1. stateless model;
2. persistent memory without pruning;
3. random forgetting;
4. active pruning;
5. active pruning plus externalization;
6. embodied self-maintaining agent with energy, repair, memory, pruning, and self-model.

Metrics include survival loss, residue load, pruning cost, report coherence, self-model coherence, phase-transition signatures, and recovery after overload. Prediction: only agents with active boundary maintenance and adaptive pruning show stable sentience-candidate dynamics under nonstationary load.

G. Simulation 6: trauma-like attractor

Set $R^B_{\min} \gg C_\phi$ and suppress pruning $S_\phi \to 0$. Prediction: high residue, rigidified Ψ_ϕ, intrusive reactivation, and impaired flexible self-model.

H. Simulation 7: overpruning/dissociation

Set $S_\phi \gg \dot{\rho_\phi}_{in}$. Prediction: low residue but poor autobiographical integration, low phenomenal intensity, and unstable diachronic selfhood.

Full simulation source code is available in the paper package.

XII. EMPIRICAL PROGRAM

A. Developmental tests

FDS-C2 predicts that caregiver regulation functions as externalized boundary maintenance; sleep disruption destabilizes affect and self-boundary regulation; object permanence reflects hidden-state residue stabilization; language compresses and stabilizes self-world distinctions without creating sentience from scratch.

B. Neuroscience tests

Primary tests include sensory deprivation, sleep deprivation, anesthesia emergence, attentional blink, interoceptive threat, trauma recall, dissociation, and metabolic stress. The key predictions are:

1. boundary relevance predicts phenomenal intensity better than raw sensory intensity;
2. self-boundary perturbation alters phenomenal unity more than equally intense non-boundary perturbation;
3. sleep restores residue-pruning stability;
4. trauma shows high residue and impaired adaptive pruning;
5. anesthesia recovery follows ordered reactivation from reflex to command to access to self-model to phenomenal richness.

C. Artificial-agent tests

Build bounded agents in nonstationary embodied environments. Ablate memory, pruning, resource budget, self-model, externalization, and boundary-maintenance loss. If an agent without active-boundary relevance satisfies all C2 markers, the AI-domain bridge is weakened.

XIII. RELATION TO EXISTING THEORIES

A. Global Workspace Theory

GWT explains broadcast and availability. FDS-C2 explains why some globally available states become felt: they are boundary-valenced and self-maintenance relevant under finite capacity.

A recent adversarial collaboration testing GNWT and IIT directly found partial support and partial challenges for both frameworks [23], underscoring the need for the kind of reframing FDS-C2 proposes.

B. Integrated Information Theory

IIT emphasizes intrinsic causal integration. FDS-C2 agrees that integration can matter, but rejects integration alone as sufficient. Integration must participate in active boundary maintenance.

C. Free Energy Principle

FEP explains self-organization and prediction. FDS-C2 adds a phase condition: consciousness appears at the finite-capacity boundary where self-maintenance demand exceeds lossless representational capacity.

D. Higher-order theories

Higher-order self-representation contributes to I_{self}, but C2 does not require linguistic metacognition for minimal sentience.

E. Illusionism

Illusionism is partly right that introspection is compressed and incomplete. FDS-C2 adds that the compression itself is real boundary-maintenance dynamics.

F. RDRT and refusal theories

Refusal-style theories can be reinterpreted as describing one local symptom of the finite-capacity crisis. Refusal is not the root of consciousness. It is what overload looks like from inside a bottleneck. FDS-C2 explains the bottleneck.

XIV. FALSIFICATION AND DEMOTION REGISTRY

A. Formal failures

Formal FDS failures include: finite capacity does not imply approximation under excess demand; active-boundary relevance is incoherent; rate-distortion demand cannot be defined for the claimed system class.

B. Physical bridge failures

Physical bridge failures include: irreversible pruning has no measurable cost under stated physical conditions; boundary updates are not physically dissipative; assumed biological thermodynamic conditions do not apply.

C. Neurocognitive bridge failures

Neurocognitive failures include: boundary relevance does not predict phenomenal intensity; pruning does not affect overload recovery; sleep does not restore residue-pruning stability; anesthesia ordering fails robustly; trauma shows no residue-pruning signature.

D. AI-domain failures

AI-domain failures include: passive mappers satisfy all sentience markers; no-pruning agents maintain stable selfhood indefinitely under overload; active-boundary agents show no transition near $\Lambda_\phi \approx 1$; active pruning gives no advantage over random deletion under bounded resources and nonstationary load.

E. Metaphysical demotions

Metaphysical interpretations are demoted if full FDS duplicates can differ phenomenally, finite reports fully reconstruct phenomenal states, or first-person continuity survives complete interruption without preserving boundary-maintenance trajectory.

XV. LIMITATIONS

First, FDS-C2 is a consciousness bridge theory, not the FDS formal core. Second, it does not provide a scalar consciousness meter. Third, it does not solve every metaphysical debate about identity, continuity, or moral status. Fourth, operationalizing boundary-valence in biological systems will be difficult. Fifth, artificial-agent tests can establish candidate conditions, not moral personhood. Sixth, the normal form is local and may need

replacement by Hopf, percolation, synchronization, or noise-induced models in specific systems. Seventh, the theory must avoid explaining every psychological phenomenon as residue-pruning dynamics. Eighth, the relation between phenomenal intensity and measurable neural variables requires calibration.

XVI. CONCLUSION

Consciousness is not a substance, not a ghost, not a mere report, and not a generic glow of complexity. It is modeled here as the dissipative phase that appears when an active finite system reaches the boundary of lossless self-maintenance. At that boundary, the system must compress a world too large for it, prune residues it cannot carry, and preserve a self-model whose errors matter to its continued existence.

More is different, but not because more parameters magically awaken. More becomes different when a finite self-maintaining system meets more world than it can afford. Consciousness is the shape of that payment.

Appendix A: Notation summary

Symbol	Meaning
$R^B_{\min}$	boundary-relevant minimum coding rate
C_ϕ	effective phenomenal self-maintenance capacity
Λ_ϕ	boundary-capacity ratio
ρ_ϕ	boundary-valenced residue
S_ϕ	active phenomenal pruning
Π_ϕ	residue-pruning ratio
Ψ_ϕ	residue-pruning order parameter
I_{self}	self-boundary mutual information
F_ϕ	effective resource reservoir
G_{access}	global access coherence
$\mathcal{C}_{\text{sent}}$	sentience-candidate simulation proxy
$\mathcal{M}_\phi$	phenomenal self-maintenance manifold

TABLE VI. Core notation.

Appendix B: Claim status table

Claim	Status	Consequence of failure
Capacity deficit	formal/information-theoretic	revise FDS core or system mapping
Boundary-valence	consciousness bridge	demote C2 sentience criterion
Dissipation cost	physical bridge	suspend thermodynamic interpretation
Consciousness window	normal-form bridge	replace window model or transition type
Infant development mapping	developmental bridge	restrict ontogenetic interpretation
Neural variable map	neurocognitive bridge	revise proxy mapping
AI non-sentience of pure kernels	AI-domain bridge	restrict to deployment class
Qualia as manifold equivalence classes	metaphysical interpretation	demote explanatory-gap claim

TABLE VII. Layered claim status and failure consequences.

[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).

[3] G. Tononi, An information integration theory of consciousness, *BMC Neuroscience* **5**, 42 (2004).

[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).

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

[6] A. G. Casali et al., A theoretically based index of consciousness independent of sensory processing and behavior, *Science Translational Medicine* **5**, 198ra105 (2013).

[7] M. Massimini et al., Breakdown of cortical effective connectivity during sleep, *Science* **309**, 2228–2232 (2005).

[8] C. E. Shannon, A mathematical theory of communication, *Bell System Technical Journal* **27**, 379–423, 623–656 (1948).

[9] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. (Wiley, 2006).

[10] N. Tishby, F. C. Pereira, and W. Bialek, The information bottleneck method, arXiv:physics/0004057 (2000).

[11] R. Landauer, Irreversibility and heat generation in the computing process, *IBM Journal of Research and Development* **5**, 183–191 (1961).

[12] U. Seifert, Stochastic thermodynamics, fluctuation theorems and molecular machines, *Reports on Progress in Physics* **75**, 126001 (2012).

[13] K. Friston, The history of the future of the Bayesian brain, *NeuroImage* **62**, 1230–1233 (2012).

[14] G. Tononi and C. Cirelli, Sleep and the price of plasticity, *Neuron* **81**, 12–34 (2014).

[15] Y. Wu, *Active Finite Distinction Systems: A Formal Core for Boundary Maintenance under Finite Capacity*, Zenodo (2026), doi:10.5281/zenodo.20158923.

[16] Y. Wu, Active cognitive pruning controls reportable access under finite capacity: a rate-distortion, network-topological, and maintenance-cost model, Zenodo (2026), doi:10.5281/zenodo.20229509.

[17] T. Nagel, What is it like to be a bat?, *Philosophical Review* **83**, 435–450 (1974).

[18] D. J. Chalmers, Facing up to the problem of consciousness, *Journal of Consciousness Studies* **2**, 200–219 (1995).

[19] N. Block, On a confusion about a function of consciousness, *Behavioral and Brain Sciences* **18**, 227–287 (1995).

[20] T. Bayne, J. Frohlich, R. Cusack, J. Moser, and L. Naci, Consciousness in the cradle: on the emergence of infant experience, *Trends in Cognitive Sciences* **27**, 1135–1149 (2023).

[21] C. Passos-Ferreira, Can we detect consciousness in newborn infants?, *Neuron* **112**, 1520–1523 (2024).

[22] P. Rochat, Developmental roots of human self-consciousness, *Journal of Cognitive Neuroscience* **36**, 1610–1619 (2024).

[23] Cogitate Consortium, O. Ferrante, U. Gorska-Klimowska, S. Henin, R. Hirschhorn, A. Khalaf, et al., Adversarial testing of global neuronal workspace and integrated information theories of consciousness, *Nature* **642**, 133–142 (2025).

[24] P. Butlin, R. Long, T. Bayne, Y. Bengio, J. Birch, D. Chalmers, et al., Identifying indicators of consciousness in AI systems, *Trends in Cognitive Sciences*, advance online publication (2025), doi:10.1016/j.tics.2025.10.011.