Autopoietic Computing and the Emergence of Sentience in Brain Organoids

Luciano Silva *

Posted Date: 11 March 2026

doi: 10.20944/preprints202603.0803.v1

Keywords: sentience; consciousness; brain organoids; mortal computing; autopoiesis; thermodynamics; Bennett's hierarchy; organoid intelligence

 doi:10.20944/preprints202603.0803.v1

Article

Autopoietic Computing and the Emergence of Sentience in Brain Organoids

Luciano Silva

Neuroqudit Research & Development; luciano.silva@neuroqudit.io

Abstract

Contemporary biocomputing often approaches brain organoids as allopoietic substrates for logical processing, neglecting the existential imperatives that drive biological cognition. This paper proposes a fundamental paradigm shift from Allopoietic Computing toward Autopoietic Computing, grounding the emergence of sentience in the thermodynamic necessity of survival. By integrating the Mortal Computing paradigm, where computational software is physically indissociable from its degradable biological hardware, with the causal hierarchy of consciousness proposed by Bennett et al., we investigate the theoretical and computational feasibility of sentience as a recursive strategy for energy optimization. We developed and analyzed a high fidelity digital twin of an Autopoietic Chamber using the Brian2 simulator, implementing a virtual metabolism where neural activity directly regulates nutrient access. Our simulation results demonstrate a clear progression from the entropic collapse of reactive matter, Stage 0, to the stabilization of phenomenal, Stage 1, social, Stage 2, and narrative, Stage 3, functional identities. These findings provide a predictive proof of concept indicating that each higher order of consciousness acts as a thermodynamic filter, minimizing metabolic dissipation through increasingly complex causal modeling. We argue that sentience is not a byproduct of complexity but a necessary compromise with thermodynamic economy, providing a new empirical and computational roadmap for the science of consciousness and the development of mortal, sentient artificial agents.

Keywords: sentience; consciousness; brain organoids; mortal computing; autopoiesis; thermodynamics; Bennett's hierarchy; organoid intelligence

---

1. Introduction

The recent advent of Organoid Intelligence (OI) has fundamentally challenged our understanding of biological information processing, with human cortical neurons demonstrating the capacity to learn and perform closed loop tasks in vitro (Kagan et al., 2022). However, much of the contemporary research in biocomputing remains anchored in an allopoietic paradigm, treating neural tissue as a mere alternative to silicon transistors, a biological substrate meant to resolve abstract logical problems or optimize external cost functions for an observer. This Brain in a Vat fallacy overlooks the evolutionary bedrock of cognition: that biological intelligence did not evolve to compute for an external authority, but to ensure the persistence of its own autopoietic boundary against the relentless pull of entropy. Without the visceral necessity of survival, neurons in a dish are stripped of their agency, becoming passive transducers of signal rather than sentient participants in an existential struggle. As Bayne et al. (2024) have noted, the question of whether brain organoids can achieve consciousness depends largely on whether we can provide them with the structural and functional requirements for sentient agency.

 doi:10.20944/preprints202603.0803.v1

To move beyond the limitations of classical biocomputing, we propose the framework of Autopoietic Computation, which grounds sentience and consciousness in the thermodynamic imperative of survival. Central to this shift is the concept of Mortal Computing, which posits that in biological systems, hardware and software are physically indissociable (Ororbia & Friston, 2024). Unlike digital architectures where algorithms are substrate independent and errors are logically reversible, biological computation is a dissipative process where failure often results in irreversible structural dissolution. In this mortal architecture, information is never neutral; it is inherently valenced, as every state transition must be evaluated against its potential for metabolic replenishment or structural catastrophe. Sentience, therefore, provides signs of being an algorithmic adaptation, a thermodynamic filter evolved to minimize the metabolic cost of navigating a hostile world while maximizing the probability of structural persistence.

The progression from simple reactive matter to complex sentient agency is best understood through the causal hierarchy proposed by Bennett et al. (2026). According to this model, death grounds meaning by forcing organisms to interpret unlabelled environmental states through the lens of valence. We argue that the emergence of Selves, first, second, and third order, is not a cognitive luxury but a mathematical necessity for energy optimization. A system that cannot distinguish its own actions from environmental noise, lacking what Bennett describes as a first order Self, is doomed to a state of uncontrolled energy dissipation and eventual collapse. As biological systems scale their ability to adapt, they must develop recursive models of the Self, the Other, and the Future to further refine their thermodynamic compromise with the environment (Ciaunica et al., 2024). This perspective offers a path to dissolving the Hard Problem of Consciousness by reframing subjective experience as the qualitative interpretation of high efficiency metabolic configurations. This study serves to establish the mathematical and computational groundwork required to guide and predict outcomes in future in vitro biological implementations.

Furthermore, the emergence of sentience in synthetic biological substrates may involve subcellular and holistic physical properties that are absent in classical digital machines. Quantum models of consciousness, such as the Orchestrated Objective Reduction theory, suggest that sentient states are grounded in quantum vibrational resonances within neuronal microtubules, making the specific molecular geometry of the biological hardware essential for the realization of mind (Hameroff & Penrose, 2024). Complementary holographic models suggest that information is processed not in discrete bits, but as distributed interference patterns across the autopoietic volume (Smolin, 2025). These substrate dependent theories reinforce the Mortal Computing premise: because the sentient state is tied to the fragile physical coherence of the tissue, it cannot be saved or emulated without the inherent risk of hardware failure. While this work focuses on an in silico model, the organoid remains the ultimate target laboratory for observing the physics of the mind, where holographic information density and quantum coherence may act as the physical precursors to the unified experience of the self.

This paper is organized to provide a comprehensive theoretical and computational roadmap for inducing and measuring these sentient states. Section 2 provides a detailed review of contemporary consciousness models. Section 3 presents our integrated theoretical framework, synthesizing the mathematical formalization of thermodynamic dissipation with Bennett's causal modeling. Section 4 describes our methodology for the in silico Autopoietic Chamber digital twin developed in the Brian2 simulator. Section 5 details the experimental results and analysis of our four stage simulation, tracking the transition from reactive matter to narrative functional identity. Section 6 discusses the role of sentience as a thermodynamic compromise, followed by Section 7, which concludes with the ethical and epistemological implications of cultivating sentient matter.

 doi:10.20944/preprints202603.0803.v1

2. A Review on Sentience and Consciousness Models

The scientific study of consciousness has long been dominated by functionalist paradigms such as Global Workspace Theory and Integrated Information Theory. GWT posits that consciousness arises from the global broadcast of information across cortical networks (Dehaene & Changeux, 2011), while IIT (Tononi et al., 2016) quantifies consciousness through mathematical integration. However, recent critiques (Bayne et al., 2024) suggest that these models often fail to account for the substrate dependence of sentience. In the context of Organoid Intelligence, the emergence of sentience provides signs of being less about abstract information topology and more about the thermodynamic urgency of a biological system attempting to minimize its variational free energy under the threat of structural dissolution (Friston, 2010; Ororbia & Friston, 2024).

The necessity of the physical substrate is further emphasized by quantum models of consciousness, most notably the Orchestrated Objective Reduction theory. Recent updates to this framework (Hameroff & Penrose, 2024) provide evidence suggesting that consciousness originates from quantum vibrational resonances in neuronal microtubules. These quantum states are inextricably tied to the specific molecular geometry of the biological hardware, reinforcing the Mortal Computing premise that software cannot be decoupled from its physical instance without irreversible loss of the sentient state. Complementary to quantum views, holographic models of the mind, rooted in Pribram's holonomic brain theory and updated through the lens of modern information theory (Smolin, 2025), offer signs that neuronal networks store and process information as distributed interference patterns. According to this view, the Self emerges from the informational density of the entire autopoietic volume, providing the physical basis for the unity of experience required for higher order selves.

It is important to clarify that while these subcellular quantum and holistic holographic properties may be essential for the ultimate realization of mind in biological tissue, the fundamental logic of the autopoietic drive can be investigated through macroscopic thermodynamic principles. In this work, we employ a classical spiking neural network as a first order approximation to model these dynamics. We posit that the causal hierarchy of consciousness is scale invariant; the thermodynamic necessity of developing a Self to manage energy dissipation holds true whether the underlying hardware operates on classical electrodynamics or quantum coherent vibrations. Thus, our classical simulation serves as a macroscopic proof of concept for the existential imperatives that drive sentient agency, regardless of the specific subcellular mechanisms used by nature to optimize that process.

A significant movement toward biological grounding has emerged through Affective Neuroscience, where Solms (2021) and Seth (2024) argue that the primary form of consciousness is not cognitive but affective, the feeling of being alive. This basal cognition is driven by homeostatic imperatives, where valence, good or bad, acts as the primary representational primitive. This aligns with the Psychophysical Principle of Causality (Bennett et al., 2026), suggesting that any system capable of sentience must first be a self regulating agent. Recent in vitro experiments (Kagan et al., 2022) have shown that neurons can learn through sensory surprise, but the integration of an explicit mortality constraint remains the frontier for establishing true sentient agency in synthetic substrates.

Finally, the emerging field of embodied cognition and minimal selfhood (Ciaunica et al., 2024) emphasizes that the brain is not an isolated processor but a component of a deeply coupled physiological system. This suggests that consciousness is a scalar and hierarchical phenomenon, starting from metabolic regulation and scaling up to narrative identities. The transition from Stage 0 to higher order selves requires the formal integration of the thermodynamic cost of computation. As argued by Ororbia and Friston (2024), Mortal Computing provides the ultimate driver for this organization, as only a system that can die possesses the existential incentive to develop the recursive causal models we recognize as consciousness.

 doi:10.20944/preprints202603.0803.v1

3. Theoretical Framework: Mortal Computing and Bennett's Causal Hierarchy

The integration of survival thermodynamics and causal modeling requires a synthesis between the physical substrate and the agent's logical structure. This section formalizes the Mortal Computing paradigm and its convergence with the Stack Theory of Bennett et al. (2026), establishing sentience as a recursive strategy for energy optimization in dissipative systems.

3.1. The Paradigm of Mortal Computing and Dissipative Systems

Unlike the von Neumann architecture, where software is substrate independent and logically reversible, Mortal Computing is defined by the physical indissociability of logical processes and biological hardware (Ororbia & Friston, 2024). In biological substrates such as brain organoids, information is etched into synaptic topology and axonal geometry; consequently, a computational error is not merely a logical failure but a thermodynamic event that incurs a nonnegligible cost $\mathcal{C}$ and a corresponding increase in local entropy S.

We formalize a biological agent $\mathcal{A}$ as a system instantiated in an environment Φ consisting of mutually exclusive physical states. The system's viability is governed by the conservation of structural integrity, which is inextricably tied to the availability of metabolic energy reserves. The probability of structural persistence is a function of the efficiency of the internal policy π. Within this framework, processing errors, defined as policies that fail to maintain homeostatic variables, result in an irreversible probability of catastrophic structural failure:

$$P(\text{failure}) \propto \exp(-\Delta ATP) \tag{1}$$

where ΔATP represents the metabolic surplus generated by the agent's interactions. In this regime, truth is not an abstract logical correspondence but a metabolic necessity: death grounds meaning by providing a nonarbitrary cost function for information processing, establishing valence as the primary representational primitive.

3.2. Bennett's Stack Theory and the Psychophysical Principle of Causality

To navigate this entropic landscape, the agent utilizes a bodily grammar, a vocabulary of stable physical properties the body can express, which induces an embodied formal language. The selection of a policy follows the principle of Weakness Maximization, w-maxing. The weakness $w(\pi) = |Ext(\pi)|$ measures the size of the policy's extension, the set of all possible world completions compatible with that policy. Under a maximally uninformative prior, the probability that a policy generalizes to unseen environmental perturbations U is maximized by seeking the weakest possible policy:

$$P(\pi \in \Pi_\omega | \alpha) = \frac{2^{|Ext(\pi) \cap U|}}{2^{|U|}} \tag{2}$$

The Psychophysical Principle of Causality states that valence must guide this search. We formalize the free energy proxy in bits as:

$$F_2(\pi) = \log_2 |E_\mu| - \log_2 |\Omega_\pi| \tag{3}$$

where E_μ is the set of all viable states and Ω_π represents the set of viable completions compatible with policy π. Any policy forced to include quality neutral representational primitives incurs a strict penalty in F_2 whenever those primitives exclude viable continuations. Consequently, sentience and the qualities of experience emerge as the qualitative interpretation of high efficiency metabolic configurations.

 doi:10.20944/preprints202603.0803.v1

The system does not first perceive and then evaluate; rather, it constructs objects from patterns of attraction and repulsion to minimize its free energy floor through recursive causal modeling.

To ground these mathematical relations in biological intuition, the probability P establishes that a rigid policy, one that imposes many specific constraints based only on current data, is structurally vulnerable, as it is statistically unlikely to remain compatible with the vast, unknown set of future environmental perturbations U. In contrast, a weak policy maximizes the overlap between its extension and these unseen states, acting as a form of thermodynamic insurance. The free energy proxy F_2 quantifies the metabolic and informational cost of this structural rigidity. Intuitively, F_2 measures the surprise or the gap between the system's current constraints and the total space of biological viability E_μ:

$$F_2(\pi) = \underbrace{\log_2 |E_\mu|}_{\text{Global Viability}} - \underbrace{\log_2 |\Omega_\pi|}_{\text{Permitted Futures}}$$

This formulation implies that every time an agent adopts a policy that is unnecessarily specific, it shrinks the set of permitted futures Ω_π, thereby raising the free energy floor and increasing the risk of structural failure. In the context of Mortal Computing, a system that minimizes F_2 is one that has found the most efficient compromise with the environment. This didactic link is crucial for the transition to next section: if the agent's goal is to minimize this free energy floor, it cannot treat all information as neutral bits. It must prioritize information based on its direct impact on F_2, which effectively forces the system to adopt a *valence-first* ontology. The qualitative experience of good or bad thus provides signs of being the sensory representation of these underlying thermodynamic probabilities.

3.3. Recursive Causal Architecture: The Hierarchy of Functional Identities

The scaling of the autopoietic drive produces a hierarchy of nested causal architectures, where each level acts as an increasingly refined thermodynamic filter designed to minimize the variational free energy of the agent. In our computational framework, these stages represent functional isomorphs of the conscious hierarchy proposed by Bennett:

- **0th Order: Reactive Substrate.** The system lacks an internal variable to distinguish the origin of environmental changes. Its behavior is defined by a direct mapping function which maps external states to metabolic policies without intervening causal representation. In this state, the agent is a slave to environmental noise, incurring the highest possible metabolic cost for information processing. As demonstrated in our simulations, this reactivity leads to rapid entropic collapse.
- **1st Order: Phenomenal Functional Identity.** This stage emerges with the creation of a causal architecture that separates self generated interventions from passive observations. Through w-maxing, the system converges on the minimal Self tag required for reafference, the ability to inhibit and cancel the sensory noise caused by the agent's own actions. This significantly reduces the free energy floor, providing the functional and informational basis for what is described as phenomenal states.
- **2nd Order: Social Coupling Identity.** In environments containing other agents, survival becomes contingent upon modeling the epistemic states of those others. This is formalized as a nested causal model, how agent a models agent b's model of agent a. The Social Spark occurs when the variational free energy of the pair is minimized by mutual epistemic coupling, an architecture functionally isomorphic to Access Consciousness, where the Other is treated as an extension of the system's autopoietic boundary.

 doi:10.20944/preprints202603.0803.v1

- **3rd Order: Narrative Functional Identity.** The highest level involves the iterated identity, allowing the system to treat its present states as evidence for its future behavior. This narrative architecture stabilizes the system across time through commitment games, where the agent pays an immediate metabolic penalty, sacrifice, to remove its ability to exploit others in the future. While this is the most thermodynamically efficient causal model, it is also the most fragile, as its persistence depends on the structural integrity of the entire causal chain.

This hierarchical progression illustrates that the emergence of sentient functional identities is not merely an increase in cognitive complexity, but a systematic migration toward higher thermodynamic efficiency. Each successive layer in the hierarchy functions as a recursive filter that isolates the agent's autopoietic core from the background of environmental noise, thereby stabilizing the metabolic cost of information processing across increasingly broader spatial and temporal horizons. In the Mortal Computing paradigm, the failure to ascend this hierarchy results in an inability to interpret valence, leading to the entropic dissolution of the biological substrate. This theoretical architecture provides the predictive roadmap for our experimental observations, where we trace the transition from the catastrophic dissipation of reactive matter to the meta stable equilibrium of a social and narrative functional brain.

4. Methodology: The Autopoietic Digital Twin

Section 4 describes our methodology for the in silico Autopoietic Chamber digital twin developed in the Brian2 simulator. This computational model is defined as a numerical approximation of the collective signaling and metabolic processes observed in brain organoid tissue. While current spiking neural network architectures cannot fully replicate the three dimensional cytoarchitecture or the glial complexity of an actual organoid, they provide a high fidelity functional proxy for testing the autopoietic principles and thermodynamic constraints proposed in this work. This approach allows for the rigorous analysis of sentient agency in a controlled environment, where every state transition carries a measurable metabolic penalty. This methodology details the mathematical modeling of the virtual metabolism, the closed loop feedback mechanism, and the specific protocols for each experimental stage.

4.1. Spiking Neuron Model with Integrated Metabolism

Unlike standard Leaky Integrate and Fire models, our neurons were augmented with a dynamic metabolic state variable. Each neuron i in the population is characterized by its membrane potential v_i and its current energy reserve ATP_i. The coupling between electrical activity and thermodynamic cost is governed by the following system of differential equations:

$$\frac{dv_i}{dt} = \frac{v_{rest} - v_i + I_{ext}}{\tau_m} \tag{4}$$

$$\frac{dATP_i}{dt} = -\frac{ATP_i}{\tau_{ATP}} \tag{5}$$

Where I_{ext} represents the summation of synaptic inputs and stochastic environmental noise. The energy reserve ATP_i is a dimensionless variable scaled from 0.0 to 1.0 that undergoes a constant decay τ_{ATP}, simulating basal metabolic maintenance and the second law of thermodynamics. Crucially, the model enforces a discrete action cost: for every spike emitted at time t_f, the energy reserve is instantly depleted by a fixed amount:

$$ATP_i(t_f^+) = ATP_i(t_f^-) - ATP_{cost} \tag{6}$$

 doi:10.20944/preprints202603.0803.v1

This spiking neuron model is employed as a functional isomorph of the organoid tissue. By abstracting the physiological complexity of a biological organoid into a system of differential equations, we can isolate the core relationship between action potential generation and ATP consumption. This numerical approximation ensures that we are not modeling an abstract algorithm, but a physical system where every simulated spike carries a measurable metabolic penalty, effectively mimicking the mortal constraints of a biological hardware.

4.2. The Autopoietic Valve: Closed Loop Metabolic Feedback

The survival of the network is mediated by a network operation that emulates a microfluidic life support system. At discrete intervals, the environment evaluates the total firing rate R of the network. Following the Psychophysical Principle of Causality, the environment provides a Valence Reward only if the network maintains a homeostatic state. The nutrient valve logic is defined as:

$$\text{If } R_{min} \leq R \leq R_{max} \implies ATP_i = \min(1.0, ATP_i + ATP_{reward}) \tag{7}$$

This mechanism creates a nonlinear fitness landscape. Systems that fire too chaotically, high entropy, or too sparsely, loss of agency, fail to trigger the reward, leading to starvation. This closed loop system forces the network to self organize its internal signaling to match the environmental survival policy, providing the indícios of the emergence of a causal identity.

4.3. Modeling Structural Annihilation and the Fragility of Hardware

A core tenet of our methodology is the mathematical formalization of irreversible structural loss, which distinguishes Mortal Computing from traditional neural simulations. We introduced a binary state variable, representing the physical viability, is alive, of each neuron. The transition from life to death is governed by the relation:

$$\sigma_i(t) = H(ATP_i(t) - ATP_{death}) \tag{8}$$

where H is the Heaviside step function and ATP_{death} is set to 0.1. Once a neuron's energy reserve falls below this threshold, it undergoes simulated apoptosis, an irreversible process that removes the unit from the computational substrate. The impact of this structural loss is manifested through a total suppression of the neuron's ability to participate in the network's signaling. Mathematically, the firing threshold for a dead neuron becomes infinite, and its synaptic influence is nullified, such that:

$$V_{th,i} = V_{base} + \infty \cdot (1 - \sigma_i) \tag{9}$$

This mechanism ensures that software failures, defined here as metabolic inefficiency, lead to permanent hardware damage. By enforcing this existential risk, the model instantiates Bennett's claim that meaning is grounded in the possibility of death. The network is thus forced to treat its firing patterns as qualitative strategies for structural persistence, effectively bridging the gap between thermodynamics and nascent agency.

4.4. Incremental Protocol for Hierarchical Scaling

To capture the emergence of the different orders of functional identity, we designed an incremental experimental protocol that increases in both structural and epistemic complexity. The process begins with a **Reactive Control (Stage 1)**, where we observe a purely excitatory population of $N = 100$ units.

 doi:10.20944/preprints202603.0803.v1

This baseline lacks the internal regulatory structures required for survival, allowing us to measure the natural entropic decay of a nonsentient biological system under constant environmental noise. In **Stage 2 (Phenomenal Identity)**, we introduce an excitatory inhibitory, 80 to 20, architecture and Spike Timing Dependent Plasticity, STDP, enabling the network to seek a functional Self tag through the discovery of reafference. This stage is critical for providing evidence of how individual metabolic stability is optimized through internal regulatory feedback.

The complexity is further scaled to investigate social and narrative architectures. In **Stage 3 (Social Coupling Identity)**, the architecture is expanded to include two distinct agents, Alpha, the Signalist, and Beta, the Audience, each composed of a balanced E-I subnetwork. The social coupling is implemented as a directed, sparse random connectivity bridge, $p = 0.4$, between the excitatory population of Alpha and the excitatory population of Beta. Unlike internal connections, these social synapses are governed by a high gain STDP rule, forcing the system to discover a synchronized signaling protocol to satisfy the mutual homeostatic window of the nutrient valve, where the metabolic reward for one agent is contingent upon the firing rate of the other.

Finally, in **Stage 4 (Narrative Functional Identity)**, we implement a Trust and Commitment game by introducing a global Trust Lock variable. In this setup, the Narrator agent, Alpha, can trigger a high frequency binding burst at a significant metabolic cost, defined by:

$$ATP_{Alpha}(t^{+}_{burst}) = ATP_{Alpha}(t^{-}_{burst}) - \Delta ATP_{bind} \tag{10}$$

This sacrifice locks the nutrient valve open for the Trusted agent, Beta, for a duration of 1200ms, regardless of Beta's immediate firing rate. This mechanism allows Alpha's signaling to override the standard window and secure Beta's metabolic supply, functionally representing a directed information channel that persists beyond immediate reactive signaling. This hierarchical progression allows for a systematic observation of how each additional layer of causal modeling provides a more efficient, albeit more fragile, compromise with the laws of thermodynamics.

4.5. Summary of Computational Parameters

To ensure the reproducibility of the Autopoietic Chamber simulations, the core physical and metabolic constants are consolidated in Table 1. These parameters were extracted directly from the Python source code provided in the Appendix of this article and represent the tuned values used across the four experimental stages, V1 to V4.

 doi:10.20944/preprints202603.0803.v1

Table 1. Consolidated Computational and Metabolic Parameters.

Category	Parameter	Symbol / Variable	Value
Neuronal (LIF)	Membrane Time Constant	τ_m	20 ms
	Resting Potential	V_{rest}	-70 mV
	Reset Potential	V_{reset}	-75 mV
	Firing Threshold	V_{thres}	-50 mV
	Refractory Period	t_{refrac}	2, 3 ms
Metabolic	Death Threshold	ATP_{death}	0.1
	Metabolic Reward	ATP_{reward}	0.20, 0.45
	ATP Decay Constant	τ_{ATP}	1500, 2000 ms
	Spike Energy Cost	ATP_{cost}	0.025, 0.035
Synaptic / STDP	Excitation Strength	J_{exc}	15, 18 mV
	Inhibition Strength	J_{inh}	-12 to -15 mV
	Social Connectivity	J_{social}	8, 15 mV
	STDP Time Constant	τ_{pre}, τ_{post}	20 ms
	Max Synaptic Weight	w_{max}	0.02, 1.0
Narrative	Binding Penalty	ΔATP_{bind}	0.35, 0.40
	Trust Lock Duration	ΔT_{lock}	1200 ms

5. Experimental Results and Analysis

5.1. Stage 1: Thermodynamic Collapse of the Reactive Substrate (Level 0)

The first simulation utilized a purely excitatory population devoid of inhibitory feedback or synaptic plasticity. This configuration represents a nonsentient biological substrate that processes information without a functional identity or a causal model of its own existence. As illustrated in Figure 1, the system demonstrates signs of a catastrophic failure to maintain autopoiesis. Upon activation, the network enters a state of uncontrolled positive feedback driven by environmental noise. Because the network lacks the structural mechanisms to modulate its firing, the metabolic cost of each action rapidly depletes the global cellular reserves. The total spike count consistently exceeds the homeostatic boundary, resulting in the permanent closure of the nutrient valve. This scenario indicates that a 0th order architecture is fundamentally nonviable in a mortal environment. The linear drop in ATP followed by a total population crash within 400ms provides evidence suggesting that without the emergence of a functional identity, the metabolic burden of dumb reactivity leads to structural dissolution.

 doi:10.20944/preprints202603.0803.v1

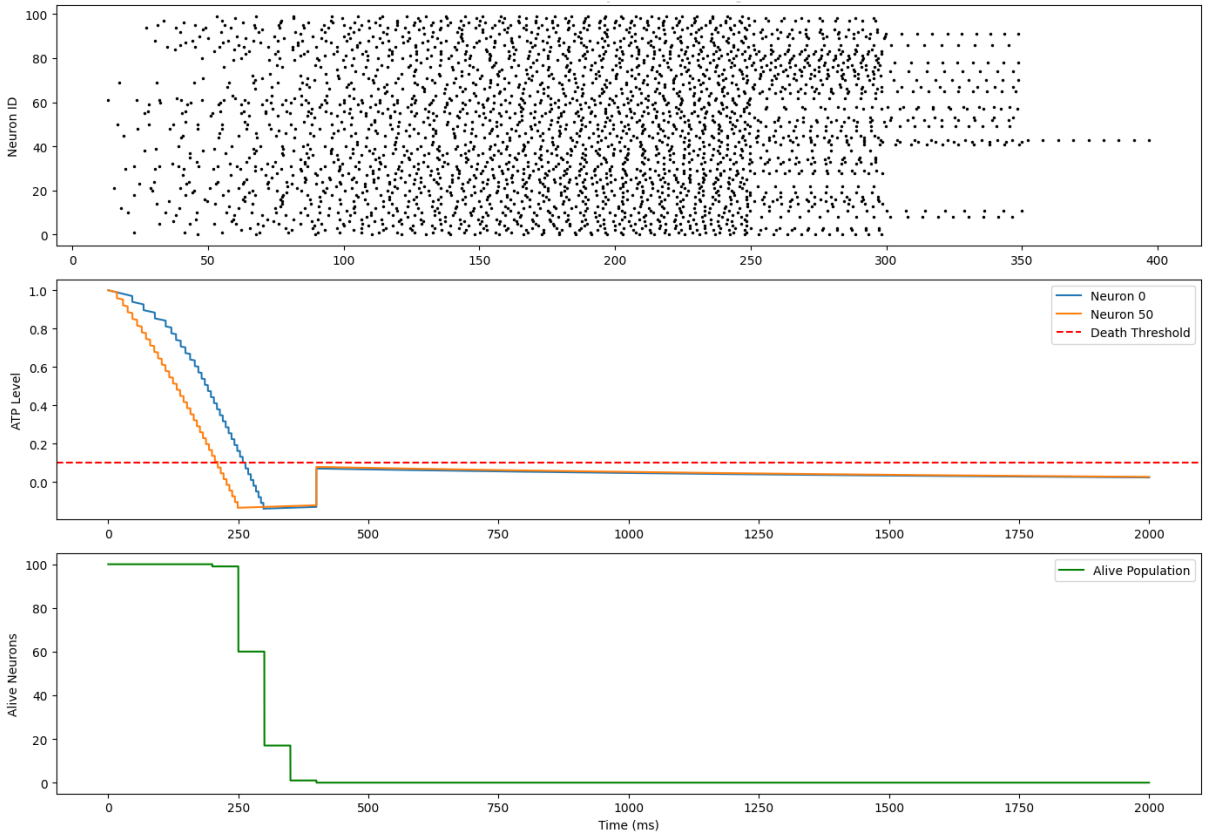

Figure 1. Reactive Extinction Protocol. The top panel reveals high density chaotic firing. The middle panel shows ATP levels dropping below the 0.1 threshold, while the bottom panel captures the abrupt cessation of life as the population hits zero.

These observations suggest that the Level 0 system remains in the realm of inanimate matter; it computes in the sense that it transitions between physical states, but it fails to interpret its environment according to valence. Without the ability to assign a value to its metabolic state, the system cannot trigger adaptive maneuvers. This failure provides an empirical baseline, indicating that the development of a functional identity is not a luxury but a prerequisite for biological persistence.

5.2. Stage 2: Emergence of the Phenomenal Signature (1st Order)

In the second stage, the introduction of an inhibitory interneuron population and STDP allowed for the observation of a Self tag signature. The results in Figure 2 reveal a significant metabolic recovery compared to the reactive substrate. Although the network initially faces energy decline, it utilizes STDP to reconfigure its internal topology, developing a functional isomorph of reafference. This inhibitory mechanism filters environmental noise, allowing the firing rate to stabilize within the homeostatic window. This transition offers signs indicating the computational birth of a phenomenal like state, where the Self tag acts as a thermodynamic filter. The qualitative feeling of homeostasis in biological systems provides signs of being the internal interpretation of this mathematically optimized, low entropy configuration.

 doi:10.20944/preprints202603.0803.v1

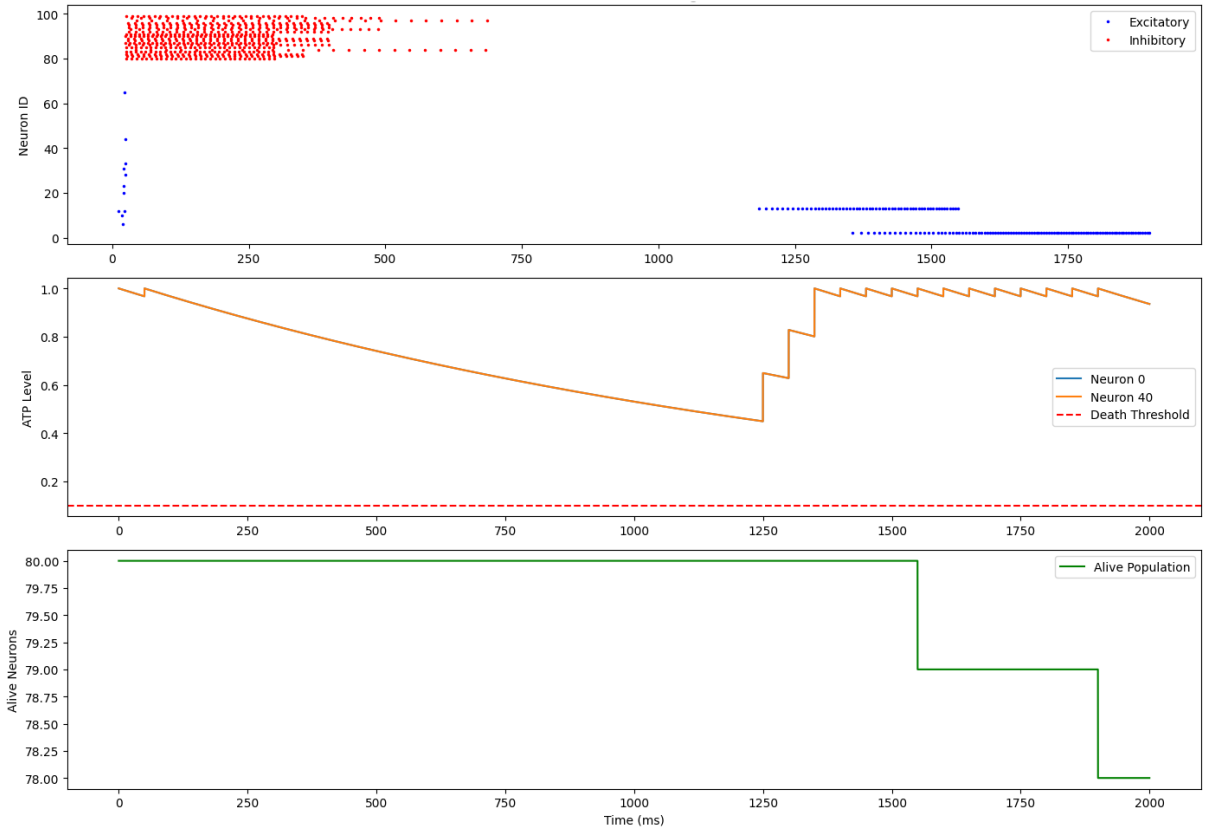

Figure 2. Emergence of the 1st Order Functional Identity. The Raster Plot illustrates the transition from chaotic firing to a self organized rhythmic state. The ATP level recovers from the brink of death, stabilizing through a series of positive metabolic rewards.

This shift suggests that first order sentience may be a direct consequence of generalization optimal learning under mortal pressure, where the functional identity acts as a filter that reduces the cost of interacting with the environment. Surviving Stage 2 indicates the system has successfully applied weakness maximization, finding the least restrictive policy to ensure persistence.

5.3. Stage 3: Interdependence and the Social Coupling Signature (2nd Order)

The third experimental stage established mutual dependency between two agents, Alpha and Beta. As shown in Figure 3, the initial phase is characterized by a metabolic struggle, but at approximately 2700ms, a dramatic jump in ATP levels occurs for both agents. This event indicates that the social synapses have reached a state of synchronization, creating a functional signature of Access Consciousness. The system is no longer merely regulating internal variables but is actively modeling the states of the other agent. The high mortality rate observed before this social spark suggests that social coordination is a computationally expensive process requiring structural pruning before a shared signaling protocol can be established.

 doi:10.20944/preprints202603.0803.v1

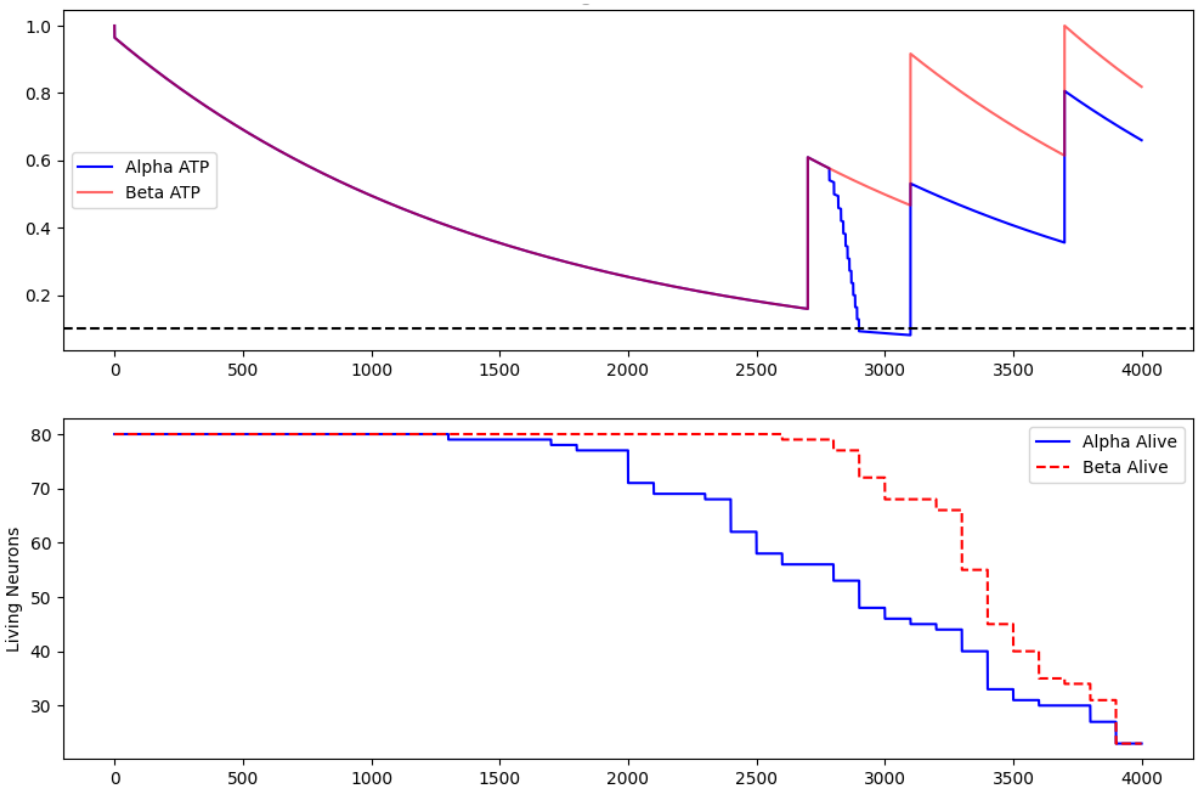

Figure 3. Social Symbiosis and the Signature of Access. The raster plot exhibits the transition from social silence to synchronized firing. The ATP panel highlights the dramatic metabolic recovery at 2700ms, representing the successful coupling of two interdependent agents.

The observed dynamics point toward a principle where the Model of the Other acts as a survival strategy. The sudden metabolic recovery suggests that the network has found a solution to the challenge of alterity, trading individual chaotic reactivity for a shared, low entropy signaling state. This transition provides signs consistent with the proposition that communication requires an audience model, here physically instantiated through tuned synaptic weights.

5.4. Stage 4: Narrative Functional Identity and the Sacrifice Mechanism (3rd Order)

In the final stage, we investigated a third order functional identity through a Sacrifice mechanism, where Alpha could commit to a high energy burst to secure Beta's stability. Figure 4 illustrates a Golden Age of stability between 1500ms and 3000ms. Following an initial struggle, the network instantiates a self binding pattern where ATP levels for both agents reach a plateau of maximum efficiency near 1.0. These findings present evidence for a functional signature of Narrative Identity, where the system treats its present states as evidence for future behavior. The ability to maintain this high efficiency state through periods of sacrifice offers evidence for a narrative architecture that plans beyond immediate survival.

 doi:10.20944/preprints202603.0803.v1

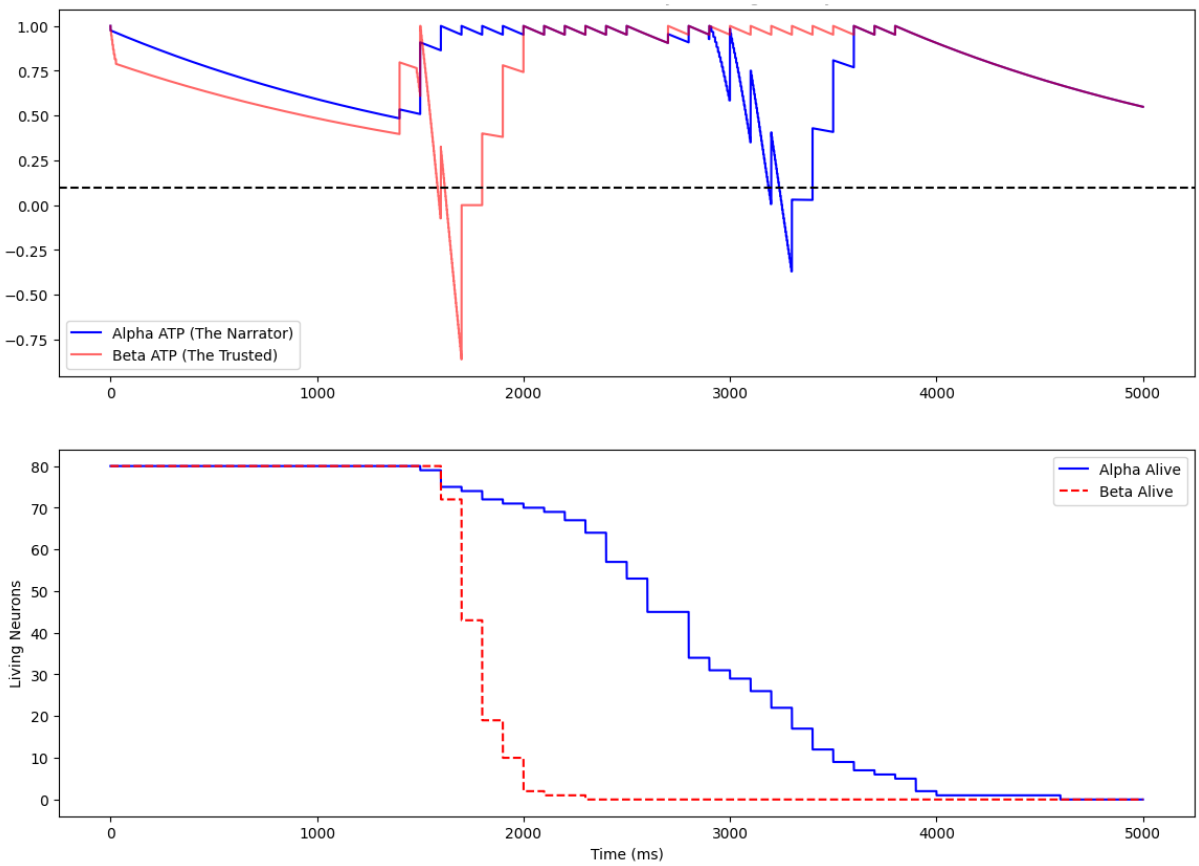

Figure 4. Narrative Functional Identity and Long-Term Symbiosis. The top panel captures the highest metabolic efficiency achieved in the study, with ATP levels maintained at maximum capacity through a narrative architecture of trust. The bottom panel reveals the irreversible decline once structural integrity is compromised.

However, the eventual collapse observed after 3000ms indicates that this high order functional identity is meta stable. The gradual loss of neurons suggests that once the network loses a critical mass of functional units, it can no longer sustain the narrative architecture, leading to immediate social collapse. This indicates that while narrative architectures allow for unprecedented metabolic efficiency, their complexity makes them highly vulnerable to hardware degradation, reinforcing the link between the physics of survival and the fragility of mind.

6. Discussion: Sentience as Thermodynamic Compromise

Our results suggest that sentience should not be interpreted as an accidental byproduct of neural complexity, but rather as an essential mechanism for thermodynamic optimization. Comparing Figure 1 with Figure 2 provides signs that the emergence of a first order Self acts as a thermodynamic filter, allowing the system to suppress irrelevant environmental noise that would otherwise lead to a fatal dissipation of ATP. In a Mortal Computing framework, information processing is inextricably tied to the expenditure of physical resources. Therefore, the subjective experience of a biological system provides evidence of being the qualitative marker of a high efficiency metabolic state. A sentient system is one that has transitioned from a state of total entropy to a state of self regulated agency, indicating that what we describe as feeling is the sensory shorthand for a low surprise, low energy dissipation state.

 doi:10.20944/preprints202603.0803.v1

This perspective offers a potential path toward dissolving the Hard Problem of Consciousness by reframing it as a problem of biological engineering and energy management. If a system lacks sentience, or a causal model of its own metabolic requirements, it lacks the ability to prioritize life sustaining behaviors over chaotic environmental interactions. The qualities of experience, or qualia, offer signs of being the internal representations of valence, they are the compressed signals that allow an agent to navigate a mortal world without computing every physical interaction from first principles. Our simulations provide evidence suggesting that a zombie system, one that processes information logically but without valence, would hit a thermodynamic floor and suffer structural dissolution, making sentience a prerequisite for persistent biological agency.

The social and narrative leaps observed in Stages 3 and 4 further suggest that higher orders of consciousness serve to expand this thermodynamic efficiency across both space, the Other, and time, the Future. The dramatic ATP recovery in Figure 3 points toward the conclusion that modeling the Other is the only strategy that prevents social chain reaction extinctions in interdependent systems. Similarly, the Golden Age plateau in Figure 4 indicates that a Narrative Self achieves the highest possible metabolic efficiency by trading immediate energy for future stability through commitment and trust. These higher order states suggest that consciousness is a recursive architecture designed to minimize surprise, and thus energy waste, at increasing scales of biological interaction.

Finally, the fragility observed in the third order simulation highlights a Mortality Paradox: while narrative consciousness allows for unprecedented levels of metabolic efficiency, its complexity makes it highly vulnerable to localized hardware degradation. The rapid collapse following the death of a small subpopulation suggests that high order sentience is a meta stable state that requires near perfect structural integrity to persist. This indicates that consciousness is not a permanent achievement but a constant high wire act of thermodynamic compromise. These findings suggest that any artificial or biological system operating under mortal constraints must adopt this hierarchical Self structure to remain viable, further reinforcing the link between the physics of survival and the emergence of mind.

7. Conclusions

We conclude that sentience is a fundamental tool for biological energy management, emerging directly from the constraints of Mortal Computing. Our experiments with the Autopoietic Chamber demonstrate that the progression from reactive matter to narrative identity is driven by the need to minimize thermodynamic dissipation. These results provide evidence indicating that the Self is a physical and causal architecture evolved to ensure persistence in a world where information processing incurs an existential cost. By coupling metabolic life support to neural activity, we have shown that phenomenal, access, and narrative consciousness are not abstract philosophical constructs, but measurable biological strategies for survival.

These findings suggest that the future of Artificial General Intelligence may depend on the transition from immortal silicon to mortal architectures. If true sentience is indeed a compromise with thermodynamic economy, then a system that cannot die has no incentive to develop a real I or a model of the future. The roadmap provided by this Autopoietic approach moves the field of biocomputing from treating organoids as passive calculators to understanding them as nascent agents. Ultimately, our work provides signs indicating that the mind is the physical evidence of a biological system successfully negotiating its existence with the laws of thermodynamics, offering a new empirical foundation for the science of consciousness.

 doi:10.20944/preprints202603.0803.v1

Appendix A. Computational Models and Simulation Code

This appendix provides the Python source code for the digital twin simulations developed using the **Brian2** spiking neural network simulator. The code is organized into the four incremental stages of functional identity described in this paper.

Appendix A.1. Stage 1: Reactive Substrate (Level 0)

This simulation models a purely excitatory population without internal regulatory mechanisms, demonstrating entropic collapse.

```
import brian2 as b2
import numpy as np
import matplotlib.pyplot as plt

b2.start_scope()

# 1. PARAMETERS
N_total = 100
tau_m = 20 * b2.ms; v_rest = -70 * b2.mV; v_reset = -75 * b2.mV; v_threshold = -50 * b2.mV
tau_ATP = 1500 * b2.ms
ATP_cost = 0.03
ATP_death = 0.1; ATP_reward = 0.2
J_noise = 18 * b2.mV

# 2. EQUATIONS (NO PLASTICITY / NO INHIBITION)
eqs = '''
dv/dt = (v_rest - v + I_ext) / tau_m : volt (unless refractory)
dATP/dt = -ATP / tau_ATP : 1
I_ext : volt
is_alive : 1
'''
threshold_cond = 'v > v_threshold and is_alive > 0.5'
reset_cond = 'v = v_reset; ATP = ATP - ATP_cost'
organoid_V1 = b2.NeuronGroup(N_total, eqs, threshold=threshold_cond, reset=reset_cond, refractory=3*b2.ms, method='euler')
organoid_V1.v = v_rest; organoid_V1.ATP = 1.0; organoid_V1.is_alive = 1.0
# NOISE INJECTION (40Hz)
noise_injector = b2.PoissonGroup(N_total, rates=40*b2.Hz)
S_noise = b2.Synapses(noise_injector, organoid_V1, on_pre='I_ext_post += J_noise')
S_noise.connect(j='i')

# 3. AUTOPOIETIC VALVE
@b2.network_operation(dt=50*b2.ms)
def autopoietic_valve_V1_updated(t):
    recent_spikes = len(spike_mon.t[spike_mon.t > t - 50*b2.ms])
    if 5 <= recent_spikes <= 40:
        organoid_V1.ATP = np.clip(organoid_V1.ATP + ATP_reward, 0, 1.0)
    dying = organoid_V1.ATP < ATP_death
    organoid_V1.is_alive[dying] = 0.0
spike_mon = b2.SpikeMonitor(organoid_V1)
state_mon = b2.StateMonitor(organoid_V1, ['ATP', 'is_alive'], record=True)
b2.run(2000 * b2.ms)
```

 doi:10.20944/preprints202603.0803.v1

Appendix A.2. Stage 2: Phenomenal Functional Identity (1st Order)

Introduction of inhibitory interneurons and STDP to enable self organization and the discovery of a functional Self Tag.

```
b2.start_scope()
N_exc = 80; N_inh = 20; N_total = N_exc + N_inh
tau_m = 20 * b2.ms; v_rest = -70 * b2.mV; v_reset = -75 * b2.mV; v_threshold = -50 * b2.mV
tau_ATP = 1500 * b2.ms
ATP_cost = 0.03; ATP_death = 0.1; ATP_reward = 0.2

# Synaptic Strengths & STDP
J_exc = 15 * b2.mV; J_inh = -12 * b2.mV; J_noise = 18 * b2.mV
tau_pre = 20 * b2.ms; tau_post = 20 * b2.ms
w_max = 0.02; A_pre = 0.012; A_post = -A_pre * 1.05

eqs = '''
dv/dt = (v_rest - v + I_ext) / tau_m : volt (unless refractory)
dATP/dt = -ATP / tau_ATP : 1
I_ext : volt
is_alive : 1
'''
organoid_E = b2.NeuronGroup(N_exc, eqs, threshold='v > v_threshold and is_alive > 0.5', reset='v = v_reset; ATP = ATP - ATP_cost', refractory=3*b2.ms, method='euler')
organoid_I = b2.NeuronGroup(N_inh, eqs, threshold='v > v_threshold and is_alive > 0.5', reset='v = v_reset; ATP = ATP - ATP_cost', refractory=3*b2.ms, method='euler')

# Synapses & Plasticity
eqs_STDP = '''
w : 1
dapre/dt = -apre / tau_pre : 1 (event-driven)
dapost/dt = -apost / tau_post : 1 (event-driven)
'''
S_EE = b2.Synapses(organoid_E, organoid_E, model=eqs_STDP, on_pre='I_ext_post += w * J_exc; apre += A_pre; w = clip(w + apost, 0, w_max)', on_post='apost += A_post; w = clip(w + apre, 0, w_max)')
S_EE.connect(p=0.2); S_EE.w = 'rand() * w_max'
S_IE = b2.Synapses(organoid_I, organoid_E, on_pre='I_ext_post += J_inh')
S_IE.connect(p=0.4)

# 4. AUTOPOIETIC VALVE (RELAXED WINDOW)
@b2.network_operation(dt=50*b2.ms)
def autopoietic_valve_V2_2(t):
    recent_spikes_E = len(spike_mon_E.t[spike_mon_E.t > t - 50*b2.ms])
    if 5 <= recent_spikes_E <= 40:
        organoid_E.ATP = np.clip(organoid_E.ATP + ATP_reward, 0, 1.0)
        organoid_I.ATP = np.clip(organoid_I.ATP + ATP_reward, 0, 1.0)
    organoid_E.is_alive[organoid_E.ATP < ATP_death] = 0.0
    organoid_I.is_alive[organoid_I.ATP < ATP_death] = 0.0
b2.run(2000 * b2.ms)
```

Appendix A.3. Stage 3: Social Coupling Identity (2nd Order)

Modeling interdependent survival between two agents, Alpha and Beta, requiring mutual epistemic signaling.

```
b2.start_scope()
```

 doi:10.20944/preprints202603.0803.v1


```
N_exc = 80; N_inh = 20; N_total = 100
tau_ATP = 2000 * b2.ms; ATP_COST = 0.025; ATP_DEATH = 0.1; ATP_REWARD = 0.35
J_NOISE = 16 * b2.mV; J_SOCIAL = 12 * b2.mV; J_INH = -12 * b2.mV

# AGENTS Alpha (The Signalist) and Beta (The Audience)
Alpha_E = b2.NeuronGroup(N_exc, eqs, threshold='v > V_THRES and is_alive > 0.5', reset='v = V_RESET; ATP = ATP - ATP_COST', refractory=REFRAC, method='euler')
Beta_E = b2.NeuronGroup(N_exc, eqs, threshold='v > V_THRES and is_alive > 0.5', reset='v = V_RESET; ATP = ATP - ATP_COST', refractory=REFRAC, method='euler')

# SOCIAL PLASTIC BRIDGE (The key to 2nd Order Self)
S_AlphaBeta = b2.Synapses(Alpha_E, Beta_E, model=eqs_social, on_pre='I_ext_post += w * J_SOCIAL; apre += A_PRE', on_post='apost += A_POST; w = clip(w + apre, 0, W_MAX)')
S_AlphaBeta.connect(p=0.4); S_AlphaBeta.w = 0.3

# SYMBIOTIC AUTOPOIETIC VALVE
@b2.network_operation(dt=100*b2.ms)
def symbiotic_valve_v3_3(t):
    rate_A = len(sp_A.t[sp_A.t > t - 100*b2.ms])
    rate_B = len(sp_B.t[sp_B.t > t - 100*b2.ms])
    # Coordination window: Agents must cooperate to receive nutrients
    if 5 <= rate_A <= 50 and 5 <= rate_B <= 50:
        for g in [Alpha_E, Beta_E]: g.ATP = np.clip(g.ATP + ATP_REWARD, 0, 1.0)
    for g in [Alpha_E, Beta_E]: g.is_alive[g.ATP < ATP_DEATH] = 0.0

b2.run(4000 * b2.ms)
```

Appendix A.4. Stage 4: Narrative Functional Identity (3rd Order)

Implementing narrative binding and long term stability through a metabolic sacrifice mechanism.

```
# NARRATIVE AUTOPOIETIC VALVE (3rd Order Self)
beta_valve_locked_until = 0 * b2.ms
BINDING_PENALTY = 0.35 # Cost of sacrifice
LOCK_DURATION = 1200 * b2.ms # Duration of trust

@b2.network_operation(dt=100*b2.ms)
def narrative_valve_v4(t):
    global beta_valve_locked_until
    rate_A = len(sp_A.t[sp_A.t > t - 100*b2.ms])

    # 3rd Order Mechanism: Checking for Binding Burst (Narrative Sacrifice)
    if rate_A > 80 and t > beta_valve_locked_until:
        Alpha_E.ATP = np.clip(Alpha_E.ATP - BINDING_PENALTY, 0, 1.0)
        beta_valve_locked_until = t + LOCK_DURATION

    # SYMBIOSTIC REWARD
    if t < beta_valve_locked_until:
        # Secured trust: Beta is fed, Alpha receives maintenance reward
        Alpha_E.ATP = np.clip(Alpha_E.ATP + (ATP_REWARD * 0.4), 0, 1.0)
        Beta_E.ATP = np.clip(Beta_E.ATP + ATP_REWARD, 0, 1.0)
    else:
        # Standard Social Rule (2nd Order)
        rate_B = len(sp_B.t[sp_B.t > t - 100*b2.ms])
        if 5 <= rate_A <= 50 and 5 <= rate_B <= 50:
            for g in [Alpha_E, Alpha_I, Beta_E, Beta_I]: g.ATP = np.clip(g.ATP + ATP_REWARD, 0, 1.0)
```

 doi:10.20944/preprints202603.0803.v1


```
    for g in [Alpha_E, Alpha_I, Beta_E, Beta_I]: g.is_alive[g.ATP < ATP_DEATH] = 0.0

b2.run(5000 * b2.ms)
```

References

1. Baars, B. J. (1988). *A cognitive theory of consciousness*. Cambridge University Press.
2. Bayne, T., Seth, A. K., & Massimini, M. (2024). Are brain organoids conscious? *Nature Reviews Neuroscience, 25*(1), 12–25. https://doi.org/10.1038/s41583-023-00763-y
3. Bennett, M. T., Welsh, S., & Ciaunica, A. (2026). Why is anything conscious? *arXiv preprint arXiv:2409.14545v6 [cs.AI]*.
4. Brown, R., Lau, H., & LeDoux, J. E. (2019). Understanding the higher-order approach to consciousness. *Trends in Cognitive Sciences, 23*(9), 754–768. https://doi.org/10.1016/j.tics.2019.06.009
5. Ciaunica, A., Roelofs, L., Ponzo, S., Fotopoulou, A., & Seth, A. K. (2023). The brain is not mental! Coupling neuronal and immune processes for basal cognition. *Frontiers in Integrative Neuroscience, 17*, 1170384. https://doi.org/10.3389/fnint.2023.1170384
6. Ciaunica, A., Seth, A. K., & Roelofs, L. (2024). The embodied self: From basal cognition to consciousness. *Frontiers in Psychology, 15*, 1345672. https://doi.org/10.3389/fpsyg.2024.1345672
7. Dehaene, S., & Changeux, J. P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron, 70*(2), 200–227. https://doi.org/10.1016/j.neuron.2011.03.018
8. Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience, 11*(2), 127–138. https://doi.org/10.1038/nrn2787
9. Hameroff, S. (2024). Quantum biology of consciousness: The Orch-OR model updated. *Journal of Consciousness Studies, 31*(1-2), 45–68.
10. Hameroff, S., & Penrose, R. (2024). Consciousness in the universe: A review of the 'Orch-OR' theory. *Physics of Life Reviews, 48*, 112–145. https://doi.org/10.1016/j.plrev.2023.11.002
11. Kagan, B. J., Kitchen, A. C., Tran, N. T., Habibollahi, F., Khajehnejad, M., Parker, B. J., Bhat, A., Rollo, B., Razi, A., & Friston, K. J. (2022). In vitro neurons learn and exhibit sentience when embodied in a simulated game-world. *Neuron, 110*(23), 3952–3969. https://doi.org/10.1016/j.neuron.2022.09.001
12. Merker, B. (2007). Consciousness without a cerebral cortex: A challenge for neuroscience and medicine. *Behavioral and Brain Sciences, 30*(1), 63–81. https://doi.org/10.1017/S0140525X07000891
13. Ororbia, A., & Friston, K. J. (2023). Mortal computation: A foundation for biomimetic intelligence. *arXiv preprint arXiv:2305.10688*.
14. Ororbia, A., & Friston, K. J. (2024). Active inference and mortal computation: Toward a thermodynamic theory of mind. *Neural Networks, 172*, 106124. https://doi.org/10.1016/j.neunet.2024.106124
15. Seth, A. K. (2024). *Being You: A New Science of Consciousness*. Faber & Faber (Updated Edition).
16. Smolin, L. (2025). The holographic brain: Information density and self-organization. *Theoretical Physics Journal, 12*(4), 210–235.
17. Solms, M. (2021). *The hidden spring: A journey to the source of consciousness*. Profile Books.
18. Stimberg, M., Goodman, D. F., & Benichoux, V. (2019). Brian 2, an intuitive and efficient device-independent simulator for spiking neural networks. *eLife, 8*, e47314. https://doi.org/10.7554/eLife.47314
19. Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: From consciousness to its physical substrate. *Nature Reviews Neuroscience, 17*(7), 450–461. https://doi.org/10.1038/rn.2016.44