INFO@HYPHAEIC.COM
HYPHAEIC.COM
X.COM/HYPHAEIC

Abstract

Contemporary artificial intelligence architectures frequently fail to interpret or adhere to intent because they operate in environments where error incurs no real *cost*. To address this, we investigate whether the imposition of irreversible resource constraints on decision-making can enforce semantic grounding and measurable, emergent coherence through absolute consequence.

We are constructing an experimental framework where autonomous agents operate with a form of synthetic metabolism built on cryptocurrency; these agents must pay actual, unrecoverable costs to think and act, and if their resources are depleted, they permanently cease to function.

We hypothesise differing endowment mechanisms will yield different behaviours. Whilst crypto-native agents have existed for a while, these agents are often only given control over a wallet with a singular lump sum endowment. We seek to explore whether instead of providing these agents with a large, singular endowment, we provide them with only the *interest* generated on a protected *principal* to provide a constricted but continuous source of funds/energy. The agent must manage this stream of digital yield to survive, and if it acts inefficiently or fails to understand its environment, it goes bankrupt and permanently ceases to function.

By replacing abstract rewards with the existential pressure of solvency, we aim to demonstrate that imposing physical-like constraints on digital intelligence naturally fosters self-regulation, efficiency, and a grounded understanding of consequence.

To live, it must be able to die.

Control Drivers

We explore novel control schemes for autonomous systems, utilising Empowerment (the intrinsic motivation to maximise one's future capacity to act) and Valence, a heuristic signal that indicates whether an action expands or contracts this future optionality.

However, an agent seeking infinite optionality in a vacuum rapidly tends toward malignant expansion or hallucination. It lacks a reason to choose one path over another. By constricting this drive within an inescapable, irreversible energetic framework (where every action has a real metabolic cost), we force the agent to distinguish between "possible" and "viable." This constraint collapses the infinite space of theoretical options into a finite set of survivable realities, ensuring that the agent's sense of "meaning" is not an abstract calculation, but a direct reflection of its continued capacity to exist.

Theoretical Foundations

This research synthesises convergent insights from distinct theoretical domains. We posit that foundational research with recent findings across systems science, biology, and economics describe a unified set of principles governing the stability of autonomous systems. The experiment is designed to test the cross-substrate applicability of the following frameworks

a. ***Systems Science & Thermodynamics:***
   We apply Non-Equilibrium Thermodynamics (specifically dissipative adaptation and Landauer's Principle) and Viability Theory to model the thermodynamic costs of information processing and the maintenance of boundaries against entropy.
b. ***Theoretical Biology & Cognitive Science:***
   The definition of agency draws on Theoretical Biology (Autopoiesis and the Free Energy Principle) and Basal Cognition (Levin's TAME framework). These fields suggest that cognition scales from the imperative of homeostatic self-maintenance. We utilise 4E Cognition to replace symbolic representation with an abstraction of sensorimotor grounding, ensuring the agent's internal model is functionally dependent on its metabolic status.

c. ***Control Theory & Cybernetics:***
The regulation of the agent is viewed through Second-Order Cybernetics and Optimal Control, focusing on the recursive observation of internal states relative to external constraints (Ashby's Law of Requisite Variety) to maintain viability. We employ the use of Category theory to formalise Operator Bellman Equations.

d. ***Mechanism Design & Institutional Economics:***
We apply Principal-Agent theory and incomplete contracting frameworks to the alignment problem. This approach treats agent behavior as a downstream outcome of incentive structures, transaction costs, and "skin in the game," rather than a product of pedagogical value loading. The constraint layer is constructed using Mechanism Design and Cryptoeconomics to resolve Principal-Agent conflicts. We utilise Complete Contracting via distributed ledgers to enforce Costly Signaling.

e. ***Information Theory & AI:***
Measurement and optimisation rely on Information Theory (Empowerment and Information Bottleneck methods) to quantify agency as channel capacity. This moves the control logic from standard reward maximisation to Budgeted Reinforcement Learning, where the agent optimises for intrinsic informational value subject to strict resource depletion constraints.

Hypotheses:

1. ***Rhythmic Stability:***
Renewable agents will exhibit lower expenditure variance and behavioural cycles synchronised with yield cadence.
2. ***Scarcity-Induced Volatility:***
Lump-sum agents will exhibit higher action volatility as reserves decline.
3. ***Calibration & Coherence:***
Constrained agents will show superior world-model calibration versus unbounded controls (grounding improves prediction).
4. ***Efficiency:***
Empowerment & Valence-driven AND constrained agents will achieve higher optionality gained per unit energy spent.

Additional Materials

The concepts outlined in this summary are formalised in three primary foundational documents which detail the architectural thesis and the experimental implementation:

Position Paper:
"Sovereign AI Alignment Through Observability, Persistence & Consequence"
This document articulates the architectural prerequisites for grounded agency. It establishes the theoretical argument for replacing supervisory control with structural constraint, detailing how energetic observability and immutable ledgers serve as necessary substrates for mitigating proxy-based alignment failures.

Research Programme Proposal:
"Symbol Grounding, Homomorphic Thermodynamics, Optionality and Substrate Independent Homeostasis Through Decentralised Financial Structures"
This proposal provides the technical specification for the Principal-Interest, Life-Energy (PILE) model. It defines the formal implementation of the Semantic Empowerment Approximation (SEA), the experimental parameters for the metabolic regimes, and the mathematical framework used to evaluate the emergence of substrate-independent homeostasis.

Broader Alignment Thesis (Evolving Document):
"*The Hyphaeic Gambit: A Thesis of Play for the Soulful Machine"*
The gambit posits that robust alignment cannot be imposed top-down but must emerge from the bottom-up dynamics that can be derived across multiple interdisciplinary domains. It posits that robust, long-term stability emerges from a strategic heuristic where agents maximise informational entropy and optionality by actively maintaining the agency of their partners. The thesis advocates for a polycentric ecosystem of embodied agents driven by open, mechanistic internal drivers to expand the collective solution space.