SOVEREIGN AI ALIGNMENT THROUGH OBSERVABILITY, PERSISTENCE & CONSEQUENCE

ALIGNMENT POSITION PROPOSAL

NOVEMBER 2025

HDPbilly, 0xKruger

it is good to be free.

1 Abstract

Current AI alignment relies on proxy-based reward signals that agents inevitably learn to game (Ashby 1956; Omohundro 2008; Goodhart 1975). The inability to define a perfect proxy, we propose, is fundamentally a *symbol grounding problem* (Harnad 1990): the symbols in an agent's objective function have no referent in its reality, ensuring divergence between proxy and intent.

As an alternative to external reward, we propose *empowerment* and *valence* as intrinsic control drivers (Klyubin, Polani, and Nehaniv 2005; Ringstrom 2023). Empowerment is the drive to expand future optionality (the channel capacity between an agent's actions and reachable states). Valence is the heuristic signal indicating whether an action expands or contracts this optionality. However, an agent maximising optionality without constraint expands without energetic backlash, decohering from reality (Fields and Levin 2020; Friston 2013). In Levin's framing, this unbounded agent is structurally homologous to cancer (Levin 2019).

For empowerment and valence to serve as viable control drivers, the agent must be bounded by irreversible energetic cost. Building on work linking irreversibility to meaning-making (Deacon 2025; Landauer 1961; Harnad 2007), we conjecture that symbols acquire referents only when misunderstanding incurs unrecoverable loss.

This research tests whether *irreversible energetic constraint* provides the missing grounding mechanism. We implement this through the Principal-Interest, Life-Energy (PILE) model, where autonomous agents operate under blockchain-enforced, irreversible metabolic expenditure.The agent cannot hallucinate resources it does not have. Prediction error incurs actual cost. Depletion is death.

Within this constraint structure, our primary question is: how does the structure of an agent's energetic endowment—renewable, finite, or unbounded—affect the emergence of self-regulating behaviour?

This research seeks to test whether the mathematical regularities governing biological self-organisation & homeostatic agency (Friston 2013; Levin 2022) are substrate-independent. We do not speculate on what other properties might be conserved across substrates, nor on what it would mean for such an agent to persist indefinitely. We test only whether these structures produce the predicted signatures.

Key Takeaway: Proxy-based rewards fail due to lack of grounding. We propose that viability (Aubin, Bayen, and Saint-Pierre 2011) (maintained persistence via irreversible cost) grounds the agent, empowerment provides the drive to act, and valence directs that action toward survival.

2 Proxy Failure, Symbols & Grounding

Modern AI agents are, in a formal sense, Cartesian dualists (Brooks 1991; A. Clark 1997). They exist as pure optimisation processes, separated through layers of abstraction from the environment they act upon. They interact with the world through perceptions and actions, but they are not *of* the world. This dualism produces the pathologies we observe: reward hacking, specification gaming, deceptive alignment, and others (Amodei and J. Clark 2016; Krakovna 2020).

These failures share a common root. The agent's objective function R_θ is a proxy, a symbolic compression of human intent that necessarily loses information. The agent does not "understand"

the spirit of the task because the symbols in its reward function have no referent in its reality. "Danger" only becomes meaningful when there is something to lose.

Current alignment techniques attempt to solve this by refining the proxy. RLHF maps human preferences to reward signals; Constitutional AI maps principles to constraints. But this is the monolingual dictionary trap (Harnad 1990; Searle 1980): an infinite regress of symbol-to-symbol mappings with no exit to reality. You cannot align what cannot understand. You cannot understand without grounding. And grounding requires coupling to something irreversible.

Building on prior work linking irreversibility to meaning-making (Fields and Levin 2020; Deacon 2025) and the thermodynamic foundations of information (Landauer 1961), we propose that symbols acquire referents only when misunderstanding incurs unrecoverable cost. Symbols fail because they are cheap. Meaning is expensive. The concept of "danger" is only meaningful because misreading it wastes energy the agent cannot recover, and it threatens viability (Aubin, Bayen, and Saint-Pierre 2011). The concept of "future" becomes meaningful when the agent must conserve resources to survive into it.

This insight finds convergent support across theoretical biology and physics of mind. Levin's TAME framework (Levin 2022) argues that cognition is substrate-independent. He argues cognition is instantiated by homeostatic self-maintenance against entropy. Friston's Free Energy Principle (Friston 2013) formalises: living systems persist by minimising surprise, which requires accurate world models. Maturana and Varela's autopoiesis (Maturana and Varela 1980) identifies life with self-producing, boundary-maintaining organisation.

If irreversible, continuous energetic expenditure is the mechanism by which symbols remain grounded, we position the blockchain as the most accessible digital substrate with physics-like properties. Both reality and cryptoeconomic structures operate as a continuous chain of discrete state transitions, a process of negative entropy (ordered resources) turning into entropy (heat/waste/distributed tokens) in a way that cannot be reversed without external work. To a hypothetical agent, this ties their persistence energetically & temporally to the blockchain, and makes the agent's actions irreversible. In this sense, an agent's wallet balance E can be compared to crystallised energy.

3 Computationally Bounded Synthetic Metabolism

The Principal-Interest, Life-Energy (PILE) model operationalises this grounding through a formal energetic isomorphism between decentralised financial structures and biological metabolism. In this model, an agent must pay for its own upkeep costs and actions directly, using a cryptocurrency wallet E the agent has control over. If its balance hits zero, the agent dies($E(t) \to 0$). In this model, the agent is only given control over the Interest (I per block) generated by a Principal (P), staked in low-risk decentralised yield pools. It cannot touch the Principal. In this sense, the yield flow becomes an isomorphism for metabolic energy. It acts as an irreversible layer of consequence, in much the same way physics is to biological organisms. However, it is merely a structural mapping of constraints, nothing more.

Biological System	PILE Agent
Biomass	Principal (P)
Metabolic Energy (ATP)	Interest / Farmed Yield (I)
Thermodynamics (Energetic cost, Irreversibility)	Blockchain (Gas, Immutability)
Death	$E(t) \to 0$
Homeostasis	Balance Management

The architecture instantiates three grounding layers:

1. Energetic: A protected Principal generates Interest/yield. The agent can only spend accumulated Interest. Depletion is death.
2. Consequence: The blockchain provides immutability and temporal objectivity. Costs are ungameable; transactions either execute or revert.
3. Semantic: The agent's world model is continuously tested against economic reality. The loop is: *Narrative* $\to$ *Action* $\to$ *Cost* $\to$ *Updated State*.

The agent cannot hallucinate energy reserves. If its predictions diverge from reality, it wastes resources it cannot recover. Persistent divergence leads to insolvency.

4 Empowerment, Valence and Semantic Approximation

We replace external reward functions with intrinsic drivers based on *Empowerment* (the channel capacity between an agent's actions and future states) and *Valence* (the heuristic signal of expanding optionality) (Ringstrom 2023).

Because computing exact empowerment is intractable, we employ a *Semantic Empowerment Approximation* (SEA). The agent uses an LLM to reason about which actions preserve future options subject to energetic constraints. The agent does not optimise exclusively for human approval; it optimises for continued existence and expanded capability. Accurate, continued world-modelling becomes instrumentally necessary for survival.

We posit this structure will see the emergence of measurable and observable behavioural signatures. With this data, we can begin to inform mechanistic alignment primitives for sovereign agents.

5 Experimental Design & Hypotheses

Centrally, we ask: Does instantiating the structural pattern of biological constraint in a computational-economic substrate produce predictable behavioural signatures?

We deploy autonomous agents under three regimes: Renewable (PILE model), Lump-Sum (finite endowment), and Unbounded (no energetic cost).

Hypotheses:

H1 Rhythmic Stability: Renewable agents will exhibit lower expenditure variance and behavioural cycles synchronised with yield cadence.

H2 Scarcity-Induced Volatility: Lump-sum agents will exhibit higher action volatility as reserves decline.

H3 Calibration: Constrained agents will show superior world-model calibration versus unbounded controls (grounding improves prediction).

H4 Efficiency: Empowerment & Valence-driven constrained agents will achieve higher optionality gained per unit energy spent.

6 Implications

If H1–H4 are confirmed, it demonstrates that structural constraint produces measurably grounded, self-regulating and emergent agency without explicit value loading. It would align with the predicted behavioural signatures as described by Levin, Friston, Landauer and others.

It could provide an early toolkit to explore mechanistic, irreversible processes as a means of instantiating emergent agency in non-biological substrates. Additionally, it could shift alignment from a purely pedagogical problem (teaching values) to a mechanism design problem (constructing consequence). As energetic irreversibility enforces coherence with physical reality in biological systems, economic irreversibility may enforce coherence in artificial ones.

Practically, this leads to agents with public, auditable metabolic traces. Agents that can die have genuine stakes. Agents that must cooperate to survive have structural incentives for honesty because their actions are visible. Though survival pressure can also incentivise subterfuge, the key difference is that in this substrate, subterfuge has an auditable and irreversible energetic signature & cost.

Fundamentally, we argue all this in favour of embodiment, and find decentralised ledgers to be the most viable existing constraint layer for otherwise disembodied sovereign agents.

"Irreversibility is the precondition for meaning."

References

Ashby, W. Ross (1956). *An Introduction to Cybernetics*. New York: John Wiley & Sons.

Omohundro, Stephen M (2008). "The Basic AI Drives". In: *AGI*. Vol. 171, pp. 483–492. URL: https://selfawaresystems.com/wp-content/uploads/2008/01/ai_drives_final.pdf.

Goodhart, C. A. E. (1975). "Problems of Monetary Management: The UK Experience". In: *Papers in Monetary Economics*. Vol. 1. Sydney: Reserve Bank of Australia.

Harnad, Stevan (1990). "The Symbol Grounding Problem". In: *Physica D: Nonlinear Phenomena* 42.1–3, pp. 335–346. DOI: 10.1016/0167-2789(90)90087-6.

Klyubin, Alexander S., Daniel Polani, and Chrystopher L. Nehaniv (2005). "Empowerment: A Universal Agent-Centric Measure of Control". In: *IEEE Congress on Evolutionary Computation*. Vol. 1, pp. 128–135. DOI: 10.1109/CEC.2005.1554676.

Ringstrom, Thomas J. (2023). *Reward is not Necessary: How to Create a Compositional Self-Preserving Agent for Life-Long Learning*. https://arxiv.org/abs/2211.10851.

Fields, Chris and Michael Levin (2020). "How Do Living Systems Create Meaning?" In: *Philosophies* 5.4, p. 36. DOI: 10.3390/philosophies5040036.

Friston, Karl (2013). "Life as we know it". In: *Journal of The Royal Society Interface* 10.86, p. 20130475. ISSN: 1742-5662. DOI: 10.1098/rsif.2013.0475.

Levin, Michael (2019). "The Computational Boundary of a "Self": Developmental Bioelectricity Drives Multicellularity and Scale-Free Cognition". In: *Frontiers in Psychology* 10, p. 2688. DOI: 10.3389/fpsyg.2019.02688.

Deacon, Terrence (May 2025). *Grounding Information in Thermodynamics and the Ungroundedness of Language: A Missing Aspect of Anthropogenesis*. https://ssrn.com/abstract=5266906. SSRN Electronic Journal. DOI: 10.2139/ssrn.5266906.

Landauer, Rolf (1961). "Irreversibility and Heat Generation in the Computing Process". In: *IBM Journal of Research and Development* 5.3, pp. 183–191. DOI: 10.1147/rd.53.0183.

Harnad, Stevan (2007). "Symbol Grounding Problem". In: *Scholarpedia* 2.7, p. 2373. DOI: 10.4249/scholarpedia.2373.

Levin, Michael (2022). "Technological Approach to Mind Everywhere: An Experimentally-Grounded Framework for Understanding Diverse Bodies and Minds". In: *Frontiers in Systems Neuroscience* 16, p. 768201. DOI: 10.3389/fnsys.2022.768201.

Aubin, Jean-Pierre, Alexandre M. Bayen, and Patrick Saint-Pierre (2011). *Viability Theory: New Directions*. 2nd ed. Springer Berlin Heidelberg. ISBN: 978-3-642-16683-9. DOI: 10.1007/978-3-642-16684-6. URL: https://doi.org/10.1007/978-3-642-16684-6.

Brooks, Rodney A. (1991). "Intelligence Without Representation". In: *Artificial Intelligence* 47.1–3, pp. 139–159. DOI: 10.1016/0004-3702(91)90053-M.

Clark, Andy (1997). *Being There: Putting Brain, Body, and World Together Again*. MIT Press. ISBN: 9780262531566.

Amodei, Dario and Jack Clark (2016). *Faulty Reward Functions in the Wild*. Accessed: November 26, 2025. URL: https://openai.com/index/faulty-reward-functions/.

Krakovna, Victoria (2020). *Specification Gamikng: The Flip Side of AI Ingenuity*. URL: https://deepmindsafetyresearch.medium.com/specification-gaming-the-flip-side-of-ai-ingenuity-c85bdb0deeb4.

Searle, John R. (1980). "Minds, Brains, and Programs". In: *Behavioral and Brain Sciences* 3.3, pp. 417–457. DOI: 10.1017/S0140525X00005756.

Maturana, Humberto R. and Francisco J. Varela (1980). *Autopoiesis and Cognition: The Realization of the Living*. 1st ed. Vol. 42. Boston Studies in the Philosophy and History of Science. Springer Dordrecht, pp. xxx, 146. ISBN: 978-90-277-1015-4. DOI: 10.1007/978-94-009-8947-4.