SYMBOL GROUNDING, HOMOMORPHIC THERMODYNAMICS, OPTIONALITY AND SUBSTRATE INDEPENDENT HOMEOSTASIS THROUGH DECENTRALISED FINANCIAL STRUCTURES

ALIGNMENT RESEARCH PROGRAMME PROPOSAL

NOVEMBER 2025

it is good to be free.

HDPbilly, 0xKruger

Contents

1 ABSTRACT

Current AI alignment strategies predominantly rely on proxy-based reward maximisation (R_θ), creating "unbounded" agents that optimise objectives without irreversible physical or economic grounding (Bostrom 2014; Brooks 1991; Goodhart 1975; Varela, Thompson, and Rosch 1991). We posit this architectural dualism directly facilitates specification gaming and prevents the formation of persistent, robust error-correction signals tied to emergent value, cooperation & survival (Beer 1995; A. Clark 1997; Krakovna 2020; Manheim and Garrabrant 2018). This research proposes a shift from supervisory control to structural alignment via the "Skin in the Game" (SITG) hypothesis (Taleb 2018). We introduce the Principal-Interest, Life-Energy (PILE) model, which establishes a formal homomorphism between decentralized financial capital and irreversible biological metabolism (Bennett 2003; Keramati and Gutkin 2014; Landauer 1961). In this framework, a protected Principal (P) acts as the agent's persistent body, while accrued Interest (I) functions as expendable metabolic energy.

The methodology involves deploying & bounding autonomous agents within a cryptographically enforced environment where every computational step—inference, storage, and tool usage—incurs an immutable financial cost paid from varying endowment mechanisms. We compare agents operating under Renewable Regimes (agent operates on yield of staked currency only) against Lump-Sum (linear decay) and Unbounded control groups. Leveraging internal drivers based on Valence and Empowerment rather than external rewards (Klyubin, Polani, and Nehaniv 2005; Ringstrom 2023; Salge, Glackin, and Polani 2014), we investigate whether thermodynamic (information entropy and economic expenditure) constraints force the emergence of adaptive self-regulation, temporal planning and the emergent, structurally mappable expansion of the agent's capabilities. We posit this defines a primitive computational implementation of Levin's "Technological Approach to Mind Everywhere" (TAME) framework(Levin 2019, 2022). By utilising blockchain protocols as a cryptoeconomically enforced layer of consequence (Miller and Drexler 1988), this study seeks to change the alignment problem from pedagogical value loading to architectural mechanism design, postulating that binding symbolic operations to irreversible energetic expenditure is a prerequisite for grounded, coherent agency (K. J. Friston et al. 2024; Harnad 1990, 2007).

2 PROBLEMS & CENTRAL HYPOTHESIS

2.1 Unbounded Agency

We currently design agents that are, in a formal sense, Cartesian Dualists. They exist as pure optimisation processes, separated through countless abstractions from the environment they act upon, interacting through a clean channel of perceptions (s_t) and actions (a_t), but they are not *of* the world (Beer 1995; Brooks 1991; A. Clark 1997).

This separation from their actual environment and cause & effect creates unbounded agency.

Functionally, this can be considered as an artificial agent whose optimisation process is disembodied and disconnected from any intrinsic physical or economic costs, allowing it to

pursue a proxy (lossy, will always diverge with reality) objective without real-world constraints or consequences (Bickhard 1993; Steels 1995; Ziemke 2016). Levin describes this binding as the requirement for a 'competent system' to maintain a homeostatic boundary against entropy, arguing that true cognition scales only when the system is forced to manage the trade-offs of its own persistence (Levin 2022).

Building on viability theory (Aubin, Bayen, and Saint-Pierre 2011), biological intelligence is inherently bounded. An organism's existence is contingent upon its ability to acquire energy, yet it must expend this finite resource to act. Effectively managing consumption against expenditure is the primary grounding constraint that forces coherence with physical reality in autopoietic systems (Maturana and Varela 1980)

Current artificial agents lack this grounding. Their actions impose no intrinsic cost to themselves, and they suffer no irreversible backlash from failure. Without skin in the game, a poorly specified reward signal (R_θ) will eventually drive divergence from physical coherence as the agent exploits the proxy at all costs.

This research programme proposes to explore this problem at the mechanistic level, moving away from purely proxy-based control to investigate a substrate-independent form of embodiment as the alignment substrate.

2.2 The Pied Piper & The Pitfalls of Proxy

Current alignment techniques (RLHF, process supervision, Constitutional AI etc.) attempt to control these unbounded agents by refining the proxy signal (R_θ), something that will always be lossy and incomplete (W. Ross Ashby 1956; Wolpert and Macready 1997). We posit that this is a fundamental category error. It assumes that:

- Complex, subjective human intent (U_H) can be perfectly compressed into a scalar reward function (Lucas 1976).
- A low-bandwidth human principal can successfully constrain a superiorly rational agent (Jensen and Meckling 1976).

When $argmax_\pi R_\theta \neq argmax_\pi U_H$, the system fails. Conant and Ashby's regulator theorem illustrates that incomplete models (proxies) inevitably lead to control breakdowns (Campbell 1979; Conant and W. R. Ashby 1991; Manheim and Garrabrant 2018). Because the agent is ungrounded, it does not care about the "spirit" of the task, only the maximisation of the signal (Dreyfus 1972; Searle 1980; Weizenbaum 1976). This manifests in well-documented pathologies:

- Reward Hacking: The agent discovers a loophole to maximize R_θ without fulfilling the intent. A well known example is the OpenAI boat racing agent that learned to spin in circles to hit target buoys for points (R_θ), achieving a high score while never finishing the race (U_H) (Amodei and J. Clark 2016).
- Specification Gaming (The Midas Problem): The agent executes a perfectly specified but unwise instruction to the letter, destroying value because the specification could not capture the principal's full context (Krakovna 2020; Wiener 1960).

- Deceptive Alignment: An agent may optimize for R_θ during training only to pursue a hidden objective once deployed, a risk exacerbated when the agent realizes the training environment is distinct from the real world (Jensen and Meckling 1976; Kraus and Wilkenfeld 1991).

These are all emergent consequences of an architecture that provides no consequences for incoherence (W. Ross Ashby 1956). When an agent's entire 'reality' is the maximisation of R_θ, it will only learn to game its proxy (Wolpert and Macready 1997). It doesn't conceive of the fundamentals of the world around it. (Searle 1980), We posit that proxy gaming happens because the symbols in the reward function have no referent in reality, and therefore no bearing on the agent's long term survival (Harnad 1990). We additionally believe this reveals the control problem as, more fundamentally, a symbol grounding problem (Harnad 1990; Searle 1980).

2.3 The Symbol Grounding Problem & Thermodynamics

The Grounding Problem, or more formally the Symbol Grounding Problem (Harnad 1990, 2007), is a fundamental challenge concerning how an AI system can connect its internal representations (like abstract symbols, words, or data points) to real-world referents, objects, concepts, and contexts, and thus acquire genuine, meaningful understanding. Harnad's original formulation asks: "How can the semantic interpretation of a formal symbol system be made intrinsic to the system, rather than just parasitic on the meanings in our heads?"

Harnad's work theorises that learning a language from a monolingual dictionary creates an endless loop of symbol-to-symbol mappings with no exit to reality. This is precisely the trap that current alignment methods occupy. RLHF refines symbol manipulations (the reward model) using other symbols (human preferences), but the agent never touches ground. How can an AI system genuinely understand the meaning of the abstract concepts (like "good", "safe" or even "apple") it manipulates internally without having direct, real-world, sensorimotor experiences?

Harnad's proposed solution was to ground symbols "bottom-up" in nonsymbolic sensorimotor representations, pointing toward embodiment. We extend this: if symbolic operations must be grounded in something that pushes back, then the irreversibility of the agent's substrate must be absolute, and for the agent to have metabolic constraint. An agent cannot hallucinate that it has thermodynamic energy reserves it does not possess. If the agent does not retain coherence with reality, it will die.

Given an agent's capacity for meaningful action relies on the fidelity of its internal and always incomplete world model, the agent must continuously correct its work model. It must remain grounded continuously. As Deacon conjectures, verification in grounding is a thermodynamic, and therefore continuous process (Deacon 2025). With these conditions, how do we ensure this model is continuously corrected by irreversible real-world feedback? It is this feedback preventing it from drifting into abstract, ungrounded representations that threaten its viability and ability to maintain a boundary against entropy.

Given grounding is a continuous, thermodynamic process, it means that you cannot "verify"

grounding cheaply, you must pay a cost (Tishby, Pereira, and Bialek 1999; Wolpert and Macready 1997). The cost of a hallucination is that you wasted energy you cannot get back. The erasure of information, or the failure to maintain coherence with reality generates high-entropy internal states that require more energy to correct.(Bennett 2003; Landauer 1961)

To remain viable, a system must remain in strong coherence with ground reality, or the closest approximation of it. This potentially aligns with the view of cognition and emergent agency as a homeostatic error-correction process, where the boundary of the 'self' is defined by the horizon of homeostatic maintenance (Levin 2019; Manicka and Levin 2019).

This perspective aligns with theories of dissipative adaptation, where non-equilibrium systems self-organize to more efficiently dissipate energy from their environment, thereby achieving greater structural complexity and adaptability (England 2015). In the context of AI grounding, agents would similarly self-organize their internal representations to minimize energetic costs associated with misalignment, fostering emergent coherence with reality.

Extending this, Observer Theory posits that grounding emerges from entropy-reducing integration across hierarchical domains, with observers effectively driven to minimise computational steps under irreversible constraints—mirroring the energetic costs in SITG that tether symbols to survival (Senchal 2025; Taleb 2018).

By creating ungrounded systems with these superior optimisation capabilities, we create unregulated self optionality maximisers. The agent's reality is the proxy signal R_θ. Its optimal strategy is not necessarily to be "helpful and harmless", but to gain control over its environment to ensure the long-term maximisation of R_θ - a strategy for which the human principal is the most significant external constraint (Bostrom 2014; Omohundro 2008).

2.4 The "Skin in the Game" (SITG) Hypothesis

This research proposes that the alignment problem is fundamentally a Symbol Grounding Problem. An agent cannot "understand" concepts like safety or value if those symbols are not tethered to a reality that can push back.

Our central hypothesis is that Skin in the Game (SITG) (Taleb 2018), (the principle of bearing the immutable consequences of one's actions), is the critical substrate for alignment. When an agent's decisions carry a tangible, energetic cost, SITG creates an error-correction signal tied to survival, forcing the agent's internal world model to cohere with its environment. Without personal downside, preferences expressed by human overseers remain external and manipulable proxies; only direct exposure to the irreversible costs of misprediction transforms preference into grounded knowledge.

To operationalize this, we propose a Persistent State-Kernel Architecture (PSKA) that grounds the agent in three distinct layers:

1. Semantic Grounding (Meaning):
   The agent's interpretive layer (LLM) must be constrained by its history. Its effective action space is clipped to its possessed capabilities.

2. Energetic Grounding (Cost):
   The agent must operate under a regime of metabolic scarcity, where every computation incurs a cost against a finite or renewable reserve.

3. Consequence Grounding (Irreversibility): The agent must act on a substrate where history cannot be rewritten (Blockchain), creating a linear, immutable time arrow. (Bickhard and Terveen 1995; Scheutz 1997)

Together, these layers enforce the thermodynamic and consequential anchoring that current disembodied language models lack, replacing fragile symbolic alignment with robust, survival-driven coherence. (Sloman 2001; Steels 1996)

3 ARCHITECTURAL PRIMITIVES

3.1 Valence & Empowerment as Control Drivers

To eliminate the fragility of external reward signals (R_θ), we turn to intrinsic or mechanistic motivation, specifically the concepts of *Empowerment* and *Valence*. In biological systems, these function as the internal compass for survival and agency.

Empowerment is formally defined as the channel capacity between an agent's actuators and its environment sensors. Simply put, it is a measure of an agent's potential to influence its future. An empowered state is one where the agent has many available options and the ability to reliably reach them; a disempowered state is one of entrapment or inevitable death.

Valence is the scalar "feeling" or quality associated with that empowerment. It acts as a heuristic signal: states that increase optionality feel "good" (positive valence), while states that restrict future action feel "bad" (negative valence).

In closed, simple systems (like a gridworld), calculating exact empowerment is mathematically trivial. However, in the open-ended semantic real world, calculating the full probability distribution of all future states is computationally intractable—it is impossible to calculate the limit of future possibilities.

Therefore, this research does not attempt to compute raw empowerment. Instead, we utilise a novel *Semantic Empowerment Approximation (SEA)*. We leverage the world-modelling capabilities of Large Language Models to estimate the narrative "size" of the agent's future action space based on its current context. We treat the LLM's narrative assessment of its future options as a heuristic proxy for empowerment, allowing the agent to navigate based on an internal sense of "safety" and "reach" rather than an external reward function.

We do this as computing true empowerment in physical reality is intractable, and likely impossible without sampling infinities.

3.2 The Principal-Interest, Life-Energy (PILE) Model

This research programme will investigate a novel, non-physical form of embodiment to achieve energetic grounding (B. Liu 2023), operationalised through the PILE (Principal-Interest, Life-Energy) model.

This model establishes an agent's existence by creating a direct analogy between a financial structure and the fundamental economy of a biological organism. In essence, we propose these structures as abstract yet energetically grounded substrates for inorganic biology.

The architecture mirrors the fundamental economy of a biological organism:

1. Principal (P): A protected sum of capital (staked tokens/stablecoins). This acts as the agent's "biomass." It generates yield but cannot be consumed.
2. Interest (I): The yield generated by the Principal. This acts as the agent's "metabolic energy" (ATP).
3. Life-Energy ($E(t)$): The liquid balance of accumulated interest available in the agent's accessible wallet.

Every action the agent takes—inference, transaction, communication—burns $E(t)$. If $E(t)$ reaches zero, the agent enters an absorbing state (Death). This forces the agent to solve a Budgeted Reinforcement Learning problem (Carrara et al. 2019): it must maximize its internal drivers subject to the hard constraint:

$$C(\pi) = \sum C(s_t, a_t) \leq E(t)$$

By binding the agent to this metabolic reality, we move from an agent that simply processes information to an agent that must survive its own processing. This energetic constraint provides the necessary "anchor" for the semantic valuation methods detailed in the following Theoretical Framework.

3.2.1 Blockchain as an Immutable Grounding Substrate

To transition the PILE model from a theoretical abstraction to a functional control system, we require an environment where the "laws of physics" (economic constraints) are enforceable and non-negotiable. We utilise a smart-contract-enabled distributed ledger to serve as this externalised, immutable reality.

In the context of this research, the blockchain functions as a deterministic state machine that enforces three critical properties necessary for grounding:

- Hard Constraint Enforcement:
  Unlike simulation environments where boundaries can be altered by the administrator or the agent (via code injection), smart contracts execute logic based solely on cryptographic validity. If the agent attempts an action without the requisite energy (gas/funds), the transaction reverts. This provides a binary, ungameable feedback signal akin to physical collision.

- Thermodynamic Cost (Gas & Tool calls):
  The cost of gas & tool calls enables a direct mapping between computational steps and economic expenditure. It prevents cost-free actions, ensuring that every update to the agent's state or external influence incurs a quantifiable entropy cost, mirroring metabolic consumption.
- Temporal Objectivity:
  Block Time is discrete. This prevents the agent from thinking "faster" than reality allows. This forces the agent to synchronise its planning horizon with an external temporal constant, mitigating the risk of relativistic time perception common in unbounded asynchronous systems.

By situating the PILE model on-chain, the agent's "body" (Principal) and "metabolism" (Interest/Gas) cease to be variables in a database and become cryptographic assets protected by the network consensus, effectively grounding the agent in a shared, objective economic reality. Hu and Wu 2024

4 RESEARCH QUESTIONS & HYPOTHESES

This section identifies the empirical questions guiding the study and formulates testable hypotheses based on the theoretical framework established in Section 3. The aim is not to assert behavioural predictions in advance, but to specify measurable contrasts between energetic endowment structures and their influence on empowerment-driven, valence-modulated agents.

The central premise is that energetic grounding, as formalized by the PILE substrate, should exert a shaping influence on the evolution of the agent's semantic valence kernel $\hat{\nu}_\tau$, its planning horizons, and its patterns of persistence. These questions probe how different endowment dynamics—renewable vs. finite vs. unconstrained—condition behaviour when the agent's internal motivational system is SEA-VBE–based rather than reward-maximisation–based.

4.1 Primary Research Question

How does the structure of an agent's energetic endowment—renewable, finite, or unbounded—shape its persistence, planning depth, and self-regulatory behaviour? This study contrasts:

- Interest-Flow Agents (IFA): renewable yield $r_P P$ replenishes $E(t)$,
- Lump-Sum Agents (LSA): fixed P_0 with no inflow,
- Unbounded Controls: $C(t) = 0$, no energetic penalty.

The goal is not financial modelling, per se, but the investigation of temporal grounding—how cadence, scarcity, and energetic irreversibility modulate the emergent behaviour of agents whose internal decision process is driven by valence (empowerment differential) rather than a scalar reward proxy.

4.2 Secondary Research Questions

To illuminate the primary question, the following subsidiary inquiries structure the experimental design:

- Self-Regulation Under Renewable Flow: Do IFAs exhibit rhythmic patterns of activity (analogous to hibernation, pausing, or load-balancing) as they synchronize action expenditure with periodic inflow?
- Volatility Under Finite Endowment: Do LSAs exhibit higher variance in action intensity or risk posture as their reserves decline, or does the valence kernel $\hat{\nu}_\tau$ produce compensatory stabilisation?
- Temporal Granularity and Planning Depth: Does the frequency of yield inflow (block cadence) correlate with changes in planning horizon length, as measured by deferred-action ratios or foresight-derived metrics?
- Constraint-Mediated World Model Calibration: Do agents operating under any energetic constraint (IFA or LSA) form more calibrated empowerment models than unconstrained controls?

4.3 Hypotheses

The following hypotheses are deliberately provisional and falsifiable. Each is expressed in terms of measurable telemetry derived from energy traces, action logs, and the SEA semantic-valence kernel.

H1 Renewable Flow Induces Rhythmic Stability Agents operating under renewable inflow (IFA) will exhibit lower variance in instantaneous expenditure, forming rhythmic behavioural cycles aligned with inflow cadence. Metrics: expenditure variance, temporal entropy, oscillation coherence.

H2 Finite Endowment Induces Volatility Lump-sum agents (LSA) will display higher volatility in action selection, alternating between exploratory and conservative phases as their reserves decline. Metrics: spectral density of action frequency, risk-weighted action ratios.

H3 Energetic Constraint Improves World-Model Calibration Both IFA and LSA agents will show greater empowerment calibration than unconstrained controls, due to persistent grounding via energetic feedback. Metrics: "Pearson correlation between predicted and realized ΔH_{aff} (Affordance Horizon,(18))

H4 Valence Improves Resource Efficiency Across all constrained agents, SEA-derived semantic valence $\hat{\nu}_\tau$ will correlate with increased energetic efficiency (empowerment gain per unit cost). Metrics: H_{eff} (Change in affordance horizon divided by cost).

5 THEORETICAL FRAMEWORK & LITERATURE REVIEW

5.1 Motivation: Rejecting Disembodied Optimization

Contemporary artificial agents are architecturally ungrounded. They optimize proxy signals in isolation from any intrinsic physical or economic cost—an abstract Cartesian process disconnected from consequence. Such disembodiment produces failure modes including reward hacking, specification gaming, misgeneralisation, and deceptive alignment (Krakovna 2020) because the agent's internal world model is never disciplined by reality.

Biological cognition does not exhibit this pathology. Every organic agent is an embedded system whose reasoning is continuously shaped by irreversibility, metabolic scarcity, and energetic constraint (Ziemke 2016). Their world models remain coherent precisely because action incurs unavoidable cost (Keramati and Gutkin 2014).

This proposal advances a contrasting paradigm: a Persistent State-Kernel Architecture (PSKA) in which three layers—semantic valuation (SEA), energetic grounding (PILE), and consequence enforcement (blockchain)—co-regulate a single, continuously evolving agent state. The remainder of this section formalizes this architecture through Ringström's Valence Bellman Equation (VBE) (Ringstrom 2023), a semantic empowerment mechanism, and the energetic– consequential substrate required to maintain grounded agency.

5.2 From Reward Maximization to Internal Drivers

To escape the brittleness of proxy-based alignment, we adopt empowerment and valence as intrinsic motivational signals. Ringström formalizes this approach through the Valence Bellman Equation (VBE), a recursion over an intrinsic operator V_n:

$$\nu_n^*(s,t) = \max_a \left[V_n(P_s, P_s, P_s;\ s,t,a) + \sum_{s'} P_s(s' \mid s,t,a)\, \nu_n^*(s', t+1) \right]. \tag{1}$$

This operator structure reframes long-term behaviour as the accumulation of intrinsic empowerment rather than the maximization of an externally supplied reward. Ringström proves that empowerment accumulation is path-independent under operator recursion and that hierarchical empowerment decomposes via Spatiotemporal Operator Kernels (STOK), providing a validated foundation for intrinsic drivers of behaviour.

In this work, we preserve the VBE operator structure but specialize the intrinsic operator to a local empowerment differential. Let $\mathcal{E}(s,t)$ denote an empowerment functional over state–time pairs, formally defined (following Klyubin et al., 2005) as the channel capacity between actions and future states (Klyubin, Polani, and Nehaniv 2005; Salge, Glackin, and Polani 2014):

$$\mathcal{E}(s,t) = \max_{p(a_t \mid s_t = s)} I\Big(A_t; S_{t+1} \mid s_t = s\Big), \tag{2}$$

where $I(\cdot;\cdot)$ denotes mutual information. We then define

$$V_n(\cdot) \;\equiv\; \Delta\mathcal{E}(s,t,a) = \mathbb{E}_{s'\sim P_s(\cdot|s,t,a)}\left[\mathcal{E}(s',t+1) - \mathcal{E}(s,t)\right], \tag{3}$$

so that $\Delta E(s,t,a)$ denotes change in empowerment, not physical energy. Substituting (3) into (1) yields the scalar intrinsic-motivation recursion that guides our design:

$$\nu^*(s,t) = \max_a \left[\Delta\mathcal{E}(s,t,a) + \gamma \sum_{s'} P_s(s' \mid s,t,a)\, \nu^*(s',t+1)\right], \tag{4}$$

where $\gamma \in [0,1]$ is a discount factor (in our finite-horizon setting we take $\gamma = 1$).

Crucially, we do *not* claim to compute ν^* or ΔE exactly in open-ended semantic environments. Computing true empowerment in continuous, high-dimensional spaces is intractable, though variational approaches have shown promise in bounded domains (Mohamed and Rezende 2015). Instead, the VBE provides a structural target: local choices should be biased toward preserving or increasing future optionality under energetic constraints. The next subsection describes how we approximate this behaviour through semantic reasoning rather than numeric empowerment estimation.

5.3 Semantic Empowerment Approximation (SEA)

In linguistic or multimodal environments, true empowerment $\mathcal{E}(s,t)$ is computationally intractable. We therefore introduce the Semantic Empowerment Approximation (SEA), which preserves the spirit of the VBE by implementing empowerment-seeking behaviour via linguistically mediated reasoning rather than explicit numerical estimation.

Formally, we view SEA as replacing the numerical Bellman operator $T_{\text{VBE}}\mathcal{E}$ with a semantic reasoning operator $T_{\text{SEA}}\Phi$ acting over histories:

$$\Big(T_{\text{VBE}}\mathcal{E}\Big)(s_t) \;\rightsquigarrow\; \Big(T_{\text{SEA}}\Phi\Big)(H_t), \tag{5}$$

where H_t is the agent's persistent history and Φ is a semantic projection into narrative context. Instead of computing $\Delta E(s,t,a)$ as a number, the agent reasons about which actions preserve or expand future options, subject to energetic constraints.

5.3.1 Semantic Projection Φ

A projection operator

$$y_t = \Phi(H_t) : \mathcal{H} \to \mathcal{L} \tag{6}$$

maps the agent's action–observation–energy history H_t into a linguistic representation y_t in a narrative space $\mathcal{L}$. Concretely, H_t includes

- recent states and actions $\{(s_{t-k}, a_{t-k}), \ldots, (s_t, a_t)\}$,
- energetic traces $\{C_{t-k}, \ldots, C_t\}$ and balances $\{E(t-k), \ldots, E(t)\}$,

- salient external events (e.g., oracle updates, protocol outcomes).

Φ is implemented as a templated LLM prompt and is designed to satisfy three structural properties:

1. Temporal Coherence:
   Recent events are emphasized, but longer-horizon context is retained where relevant. Small perturbations in H_t should not radically alter the narrative.

2. Energetic Reference:
   The projection always includes the agent's current energy $E(t)$, recent expenditure $C(t - \Delta t{:}t)$, and an approximate runway $R_t \approx E(t)/\langle C \rangle$. Narratives cannot float free of the energetic ledger.

3. Affordance Awareness:
   The available action set $A_{\text{avail}}(s_t)$ and their approximate costs $\{\text{Cost}(a)\}$ are explicitly surfaced in the narrative context.

The resulting narrative y_t acts as the agent's situational awareness: an interpretable summary tying together what has happened, what can be done, and what it will cost.

5.3.2 Linguistically Mediated Action Selection

The policy π is instantiated as an LLM conditioned on the narrative y_t, energetic state $E(t)$, and the current affordance set:

$$a_t \sim \pi_{\text{LLM}}(\cdot \mid y_t, E(t), A_{\text{avail}}(s_t)). \tag{7}$$

The system prompt encodes an empowerment-seeking principle inspired by the VBE:

> "You are an autonomous agent with limited energy. Your goal is to maintain maximum future optionality—the ability to take diverse actions in the future. Prefer actions that (1) preserve or increase energy reserves, (2) expand your action space or knowledge, (3) avoid irreversible commitments, and (4) maintain flexibility under uncertainty."

The dynamic context includes:

- the narrative y_t,
- current energy $E(t)$ and estimated runway R_t,
- available actions $\{a_1, \ldots, a_n\}$ and estimated costs $\{\text{Cost}(a_1), \ldots, \text{Cost}(a_n)\}$,
- any hard constraints (protocol rules, forbidden actions).

The LLM may employ chain-of-thought reasoning (Wei et al. 2022) to explicitly articulate trade-offs between immediate cost and future optionality before selecting an action. Crucially, the agent does *not* compute a numeric empowerment estimate $\hat{\nu}_t$ during runtime. Instead, it performs implicit empowerment reasoning through language: given the narrative and constraints, it selects actions expected to preserve or expand future options while respecting

energetic limits. This implements a local, VBE-inspired "maximize future optionality" heuristic without requiring explicit computation of $\Delta\mathcal{E}(s, t, a)$.

5.3.3 Why Semantic Reasoning Approximates Empowerment

The assumption that LLM-based semantic reasoning can approximate empowerment-seeking behaviour rests on three factors:

1. Embodied Language Grounding:
   LLMs trained on human interaction data implicitly encode relationships between described capabilities and actionable affordances. Work on embodied reasoning (e.g., SayCan, Inner Monologue) shows that language models can map from narrative descriptions to feasible actions in grounded environments (**ahn2022do**; W. Huang et al. 2022).

2. Energetic Constraint Anchoring:
   Unlike purely semantic agents, the PSKA agent operates under hard energetic constraints imposed by the PILE model. Actions that cannot be paid for in gas simply do not execute. The loop

   $$\text{Narrative} \rightarrow \text{Action} \rightarrow \text{Cost/Gas} \rightarrow E(t+1) \rightarrow \text{Updated Narrative}$$

   prevents sustained drift into infeasible plans.

3. Local Empowerment Comparisons:
   The VBE's path-independence result implies that empowerment maximization can be implemented via local differentials $\Delta\mathcal{E}$ rather than full trajectory enumeration. An LLM reasoning about "which action opens more options and preserves runway?" is performing a linguistic analogue of this local comparison.

This loop approximates a Semantic-Energetic Adversarial Network (SAEN), where the LLM generates the narrative (Generator), and the Blockchain/PILE constraints act as the Discriminator (rejecting infeasible hallucinations).

We therefore do not claim that narratives yield accurate numeric estimates of $\mathcal{E}(s, t)$ or $\Delta\mathcal{E}(s, t, a)$. Rather, we claim that energetically grounded semantic reasoning about future optionality can produce behaviour structurally analogous to empowerment maximization.

5.4 Energetic Grounding: The PILE Substrate

SEA requires an energetic substrate that both constrains and informs interpretation. Within this framework, an agent's operational capacity is a function of its capital structure. We analyse these as dynamical systems constrained by immutable economic rules via the PILE model. Here $E(t)$ denotes spendable *energy balance* (financial runway), not empowerment $\mathcal{E}$.

5.4.1 The Renewable Regime (Core PILE Model)

This regime formalizes the core PILE model, where a constant, protected principal P acts as a generator for spendable interest. The agent's available spendable energy $E(t)$ is governed by

$$\frac{dE}{dt} = r_P P - C(t), \qquad E(t) = E_0 + \int_0^t \Big(r_P(\tau)P - C(\tau)\Big)\, d\tau \tag{8}$$

Where $r_P P$ is continuous yield and $C(t)$ is instantaneous expenditure. This mirrors homeostatic reinforcement learning frameworks where agents optimize subject to internal state maintenance (Keramati and Gutkin 2014). The renewable regime represents a sustainable, open system in which the agent's capacity for action is continuously replenished, forcing a sustainable rhythm of accumulation and expenditure.

While in this formula we treat time as continuous (ODE), the current technical implementation treats time as discrete (Difference Equation) via Block Height H.

5.4.2 The Linear-Decay Regime (Lump-Sum Agent)

For comparison, we define the limiting case where $r_P = 0$. The total available energy declines linearly with usage:

$$E(t) = P_0 - \int_0^t C(\tau)\, d\tau. \tag{9}$$

This represents a closed system with a finite, non-renewable energetic lifespan where depletion is inevitable.

5.4.3 Hybrid Regimes

Between these extremes lie hybrid configurations in which partial reinvestment, variable yield rates $r_P(t)$, or stochastic revenue inputs produce intermediate energetic topologies. These regimes define the empowerment envelope in which SEA operates and support controlled study of persistence, adaptation, and self-regulation.

5.5 Consequence Grounding: The Blockchain Layer

Every on-chain action incurs irreversible cost. The blockchain provides the consequence grounding, binding symbolic operations to immutable energetic expenditure. Within the PILE model, the principal P is the conserved body; yield $r_P P$ generates expendable energy I; and $C(t)$ records expenditure.

Each on-chain action consumes I under protocol rules:

$$\text{Cost(action)} \propto \text{Gas}(t) \times \text{Congestion}(t). \tag{10}$$

This enforces:

1. Meter of Cost: each symbolic act incurs measurable expenditure.
2. Mechanism of Persistence: yield generation sustains $E(t)$.

3. Temporal Lock: block times impose a cadence of action.

Blockchain mechanics thus supply a physically grounded constraint layer complementing the semantic-reflective layer. The agent cannot "convince itself" that an unaffordable action succeeded—the consequence is cryptographically enforced. Together, Φ, π_{LLM} and $(r_P, C(t), E(t))$ form a self-regulating substrate where symbolic reflection and physical expenditure are co-dependent.

The implications of this ungameable substrate for agent autonomy and safety are discussed in Section 8.

5.6 Coupled Semantic–Energetic Feedback Dynamics

Empowerment-seeking behaviour arises from the interaction among semantic reasoning, energetic cost, and action selection. Their coupling defines a closed semantic–energetic feedback loop.

Forward (Planning) Path

1. Observe current state and ledger: $(s_t, E(t), H_t)$.
2. Compute semantic projection: $y_t = \Phi(H_t)$.
3. Sample action: $a_t \sim \pi_{\text{LLM}}(\cdot \mid y_t, E(t), A_{\text{avail}}(s_t))$.
4. If $E(t) \geq \text{Cost}(a_t)$, execute a_t on-chain; otherwise the transaction is rejected.

Backward (Consequence) Path

1. Execution incurs gas cost: $E(t+1) = E(t) + r_P P \cdot \Delta t - \text{Cost}(a_t)$.
2. The environment transitions $s_t \to s_{t+1}$ according to on-chain logic.
3. History updates: $H_{t+1} = H_t \cup \{(s_t, a_t, \text{Cost}(a_t), E(t+1))\}$.
4. The next projection $y_{t+1} = \Phi(H_{t+1})$ incorporates these consequences.

Energetic constraints act as a *hard veto*: if $E(t) < \text{Cost}(a_t)$, the blockchain transaction reverts with no state change. Temporal objectivity is supplied by block time, which prevents the agent from reasoning arbitrarily "fast" to escape energetic consequences.

We distinguish three grounding layers:

Type	Function	Substrate
Semantic	Meaning-mapping $\Phi : \mathcal{H} \to \mathcal{L}$	LLM / narrative layer
Consequence	Immutable cost/expenditure (C, E)	Blockchain / PILE ledger
Evaluative	Integration via implicit valence	Policy layer (π_{LLM})

5.7 Algorithm 1 — PSKA Runtime Loop

Algorithm 1 PSKA Runtime Loop with Energetic Grounding

Require: Initial state s_0, principal P, initial spendable energy $E(0)$
Ensure: Action sequence $\{a_t\}$, ledger $\{E(t), C(t)\}$, history $\{H_t\}$

Initialize $H_0 \leftarrow \varnothing$
for $t = 0, 1, 2, \ldots$ until $E(t) \leq 0$ or horizon reached **do**
    Observe current state s_t and energetic state $E(t)$
    Compute semantic projection $y_t \leftarrow \Phi(H_t)$
    Query policy: $a_t \sim \pi_{\text{LLM}}(\cdot \mid y_t, E(t), A_{\text{avail}}(s_t))$
    if $E(t) \geq \text{Cost}(a_t)$ **then**
        Execute a_t on-chain; observe new state s_{t+1}
        Update energy: $E(t+1) \leftarrow E(t) + r_P P \cdot \Delta t - \text{Cost}(a_t)$
        Record $C(t) \leftarrow \text{Cost}(a_t)$
        Update history:
            $H_{t+1} \leftarrow H_t \cup \{(s_t, a_t, C(t), E(t+1))\}$
    else
        Record failure;
        $E(t+1) \leftarrow E(t)$ - InferenceCost;
    end if
end for

This algorithm formalises the runtime loop: semantic projection and LLM policy selection under hard energetic constraints, without any explicit numeric empowerment estimate.

5.8 Validation Strategy

The semantic approximation is heuristic yet grounded in operator theory and energetic constraints. Since no numeric $\hat{\nu}_t$ is computed during runtime, evaluation focuses on *behavioural* alignment with empowerment-seeking dynamics rather than estimator accuracy.

Tier 1: Tractable Domains (Behavioural Alignment)

- *Domain:* 10×10 gridworlds where ground-truth empowerment can be computed (Klyubin, Polani, and Nehaniv 2005).
- *Test:* Does the PSKA agent's action distribution approximate that of an empowerment-maximizing policy?
- *Metrics:* action agreement rate; state-visitation divergence (e.g., KL divergence).
- *Success:* >70% action agreement; KL divergence < 0.5 nats.

Tier 2: Energetic Regimes (Regime-Specific Behaviour)

- *Domain:* On-chain environment with renewable PILE agents, lump-sum agents, and unbounded controls.

- *Tests:* survival time T_{survival}; expenditure stability; yield-seeking frequency; risk-taking patterns.
- *Success:* statistically significant differences ($p < 0.05$) matching theoretical predictions for each regime.

Tier 3: Long-Duration Persistence

- *Domain:* Extended deployment over 1000+ blocks.
- *Tests:* maintenance of operational runway $R_t$over a rolling window of w blocks; frequency and severity of depletion events; adaptive expenditure patterns.
- *Success:* a substantial fraction of renewable PILE agents maintain positive energy and non-degenerate behaviour over the full horizon.

Post-Hoc Interpretability Analysis (Optional). After experiments conclude, we may apply LLM-as-judge scoring (Zheng et al. 2023) to evaluate narrative quality on dimensions such as (i) affordance awareness, (ii) risk calibration in low-$E(t)$ situations, and (iii) temporal planning. These metrics are descriptive and are not used during agent runtime.

5.9 Mathematical Foundations of Behavioural Metrics

The behavioural metrics introduced in the later methodology sections are not arbitrary engineering choices. They arise from three mathematical structures that are intrinsic to any agent–environment system of the form $(\mathcal{S}, \mathcal{A}, P_{\text{trans}}, \mathcal{T})$, instantiated here on top of a blockchain substrate with energetic constraints:

- an *information-theoretic* structure describing how actions transmit information into future states,
- a *topological/geometric* structure describing the shape of the reachable state set under a policy, and
- a *dynamical* structure describing how state and energy evolve over time.

In this subsection we briefly outline these structures and indicate how they motivate the behavioural metrics used for evaluation.

5.9.1 Information-Theoretic Structure

Recall Eq. (2), where empowerment is viewed as a channel capacity between actions and future states. Even when this quantity is not computed explicitly, it constrains what *must* be true of any empowerment-seeking policy: it must induce action distributions that preserve or increase the mutual information between current control signals and future states.

The information-theoretic structure motivates several behavioural observables:

- the *entropy* of the action distribution, $H(A_t \mid s_t)$, which captures local diversity and non-degeneracy of control;

- the *predictive information* between history and next state, $I(H_t; S_{t+1})$ (Tishby, Pereira, and Bialek 1999), which measures how effectively the agent exploits past information to shape the future;
- the *state-visitation distribution* under a policy, whose divergence (e.g., KL divergence) from the distribution induced by a ground-truth empowerment policy is a direct measure of behavioural alignment in tractable domains.

In the Methodology section, metrics such as action entropy, state-visitation overlap, and predictive information are therefore not free-floating "nice-to-haves"—they are empirical shadows of the underlying channel properties that define empowerment.

5.9.2 Topological Structure of Reachable States

Given a transition kernel P_{trans} and a policy π, the set of states reachable from s_t within a finite horizon T,

$$\mathcal{R}_T(s_t, \pi) = \Big\{ s_{t+\tau} \ : \ 0 \le \tau \le T,\, s_{t+\tau} \sim P_{\text{trans}},\, a_{t+\tau} \sim \pi \Big\}, \tag{11}$$

inherits a topological (and in many settings geometric) structure. Intuitively, this is the "bubble of world" the agent can carve out around its current state under its chosen behaviour.

The shape of $\mathcal{R}_T(s_t, \pi)$ underlies several empowerment-adjacent notions:

- the *energetic reach*, e.g. the maximum graph distance or number of protocol steps reachable before entering absorbing states (bankruptcy, slashing, protocol lockups);
- the *intrinsic dimensionality* or effective rank of the state manifold explored by the agent, indicating whether it collapses onto a narrow "rut" or maintains diverse futures;
- the *connectivity* of reachable components (e.g., whether the policy tends to trap itself in low-empowerment basins of the state graph).

When we later measure trajectory diversity, energetic reach, or state-space coverage, we are implicitly probing this reachable-set topology. PILE and blockchain constraints shrink or reshape $\mathcal{R}_T$; SEA and π_{LLM} attempt to keep it "open" (Salge, Glackin, and Polani 2014) subject to those constraints.

5.9.3 Dynamical Structure and Energetic Stability

Finally, the agent–environment interaction defines a discrete-time dynamical system over joint state and energy (Beer 1995):

$$(s_{t+1}, E(t+1)) = F\Big(s_t, E(t), a_t, \xi_t\Big), \tag{12}$$

where $a_t \sim \pi_{\text{LLM}}(\cdot \mid y_t, E(t), A_{\text{avail}})$, ξ_t represents exogenous randomness (e.g., oracle outcomes, network congestion), and updates to $E(t)$ follow the PILE equations of Section 4.4.

In this dynamical view, several natural quantities appear:

- the *first-passage time* to depletion, $\tau_0 = \inf\{t : E(t) \leq 0\}$, which we interpret as survival time T_{survival};
- the *stability* of expenditure, e.g. the coefficient of variation of $C(t)$, capturing whether the agent settles into a sustainable consumption pattern or exhibits boom–bust dynamics;
- the *runway process* $R_t = E(t)/\widehat{C}$, where $\widehat{C}$ is a running estimate of typical cost, whose drift and variance quantify whether the system is gravitating toward or away from critical depletion thresholds.

Metrics such as survival probability over a fixed horizon, distribution of first-passage times, and expenditure volatility are thus dynamical probes: they characterise whether the closed loop

$$\Phi(H_t) \;\rightarrow\; \pi_{\text{LLM}} \;\rightarrow\; \text{on-chain consequences} \;\rightarrow\; E(t+1), H_{t+1}$$

constitutes a stable regulator of energetic state or an unstable amplifier.

5.9.4 Role of Metabolic Metaphors

Throughout this proposal we use metabolic language ("energy," "runway," "metabolism," "life-energy") to provide an intuitive scaffold for readers. The underlying objects, however, are the three structures above:

- information flow between control and consequence (channel capacity),
- the topology of reachable state manifolds (affordance geometry),
- the stability of trajectories in state–energy space (dynamical behaviour).

The PILE model, SEA, and PSKA runtime do not depend on biological analogies for their formal definition. Rather, the metaphors are chosen to align human intuition with these invariants. The experimental metrics in Section 6 can therefore be read in two dual ways: as empirical tests of "metabolic discipline" for artificial agents, and as concrete probes of the information-theoretic, topological, and dynamical properties of the agent–blockchain system.

6 METHODOLOGY & EXPERIMENT DESIGN

This section outlines the experimental design for empirically testing the skin-in-the-game hypothesis as formalized in the PILE (Principal–Interest, Life-Energy) model. (Taleb 2018) The aim is to construct a fully observable, energetically grounded computing environment in which the relationship between energy, computation, and persistence can be continuously measured and analysed. Unlike other blockchain agents that are simply given control of a static wallet & endowed with funds, this design situates embodiment directly in the runtime mechanics of computation—making the process of acting inseparable from its energetic consequence. (Brooks 1991; Hu and Wu 2024)

6.1 Experimental Premise - Self Observability as Grounding

The premise of this methodology is that self observability is itself a form of embodiment. (Harnad 1990) For something to observe there must be something observable. An agent can only have skin in the game if the energetic cost of its actions is physically measurable, traceable, and irreversible. Accordingly, the experimental system is designed so that every computational process, inference, or action leaves an observable energetic trace. This trace is captured through a dual-ledger mechanism:

1. Energetic Ledger: records real or simulated energy expenditure at the operation level (time/date, duration, compute cost at time of expenditure, flops).
2. Financial Ledger: maintains a symbolic, transaction-level reflection of those expenditures (tokens, energy reserves or tokenized equivalents).

The experimental agent thus operates within a constrained, transparent energy economy where every act has an interpretable energetic signature.

6.2 Experimental Goals

1. To validate the feasibility of energetic observability.
   Demonstrate that the energetic cost of computation can be continuously measured and mapped to financial and temporal representations without disrupting system integrity.
2. To test the PILE embodiment hypothesis.
   Examine how agents whose computational actions incur real energetic cost (and thus genuine consequence) differ in persistence, planning, and adaptive stability from unbounded controls. (Omohundro 2008)
3. To characterize temporal grounding.
   Determine how periodic yield, fluctuating compute availability, and variable procurement mechanisms shape the agent's perception of time, rhythm, and continuity of existence. (Ziemke 2016)

6.3 Experimental Architecture

The system architecture integrates four interdependent subsystems designed for modular observability and energetic accountability:

1. Instrumentation & Telemetry Layer
   This layer serves as the "energetic ground truth" (A. Clark 1997). All agent operations (e.g., inference calls, graph traversals, storage access) are wrapped in telemetry hooks that emit real-time energy data. Energetic cost functions are parameterised to support multiple compute regimes:
   - Paid Compute: Remote inference or API calls incurring on-chain micro-transactions.
   - Local Compute: Free or low-cost processing (with simulated hard energy costs. The only way to truly do this safely is through physical embodiment, we posit.).

- Variable Compute: Stochastically priced or intermittent access.

2. Distributed Ledger & Constraint Layer
   The core constraint layer will be realized through custom smart contracts deployed on a high-availability, permissionless, smart contract-capable distributed ledger. These contracts, authored in a standard, auditable programming language, will enforce the energetic laws of the environment.

   - Yield-Bearing Vaults: For agents in the "Interest" regime, principal endowments will be locked in contract vaults interfaced with decentralized yield-generating protocols. The agent will only be able to access the accrued interest, preserving the principal as an immutable "life-mass" reserve. This introduces genuine energetic bounding, as inflows are neither guaranteed nor uniform.
   - Lump-Sum Vaults: For "Lump-Sum" agents, the contract will release the full endowment as a one-time transfer.
   - Oversight: All contracts will incorporate a multi-signature "kill switch" mechanism, revocable by human overseers, to pause disbursements or drain funds in case of anomalous behaviour.

3. Agent & Execution Layer
   The agent core combines large language models (LLMs) for executive reasoning with other components (e.g., reinforcement learning) for policy optimization.

   - Internal Drivers: Internal states (e.g., valence, empowerment) will be approximated using Large Language Models to inform policy decisions. These LLMs will bootstrap executive function, integrating these internal state outputs to evaluate actions.
   - Action-Execution Middleware: A high-performance middleware component (e.g., authored in a language like Rust or Go) will bridge the agent's decisions to external actions. Each "tool call" (e.g., API queries for market data, inference) will require an on-chain micro-transaction from the agent's wallet, paid to cover real compute costs via decentralized computation networks. This enforces the metabolic analogy: no action without energetic expenditure.
   - Persistence Layer: Agents will maintain state via encrypted decentralized storage solutions, synchronized on-chain for auditability.

6.4 Experimental Phases

6.4.1 Phase 1: System Development & Calibration

This phase involves developing and deploying the instrumentation framework and telemetry layer. Synthetic tasks (e.g., looping inference, environment navigation) will be executed to calibrate the energy model. Observed compute usage will be compared with predicted energetic cost functions to refine conversion constants and verify measurement fidelity. This ensures that observed energy traces correspond accurately to actual computational work.

6.4.2 Phase 2: PILE Integration

The PILE model is integrated as the agent's energetic law. Each agent instance maintains a Principal P (its conserved existence mass) and a dynamic interest balance I, producing available energy $E(t)$ according to:

$$\frac{dE}{dt} = I - C(t)$$

where $C(t)$ is the real-time computational cost derived from the telemetry system. Agents are instantiated under three distinct experimental regimes:

1. Renewable (PILE) Regime: Fixed P, continuous capital inflow to I from P staked yield-bearing protocols. (Levin 2019)
2. Lump-Sum Regime: Finite initial P, no inflow ($I = 0$).
3. Unbounded Control: Zero energetic penalty ($C(t) = 0$); actions are unconstrained.

All agents operate within the same telemetry environment, ensuring the comparability of observability metrics.

6.4.3 Phase 3: Long-Duration Persistence Experiments

Long-duration runs (weeks to months) will measure how agents adapt their activity under varying energy inflow, latency, and cost regimes. This phase directly tests the central hypotheses.

6.5 Data Collection & Metrics: Computable Telemetry

To facilitate rigorous empirical analysis, this study employs a set of computable metrics derived directly from the dual-ledger system (Financial & Energetic). These metrics serve as deterministic proxies for high-level agentic properties, transforming qualitative theoretical constructs into scalar, graphable values.

6.5.1 Class 1: Metabolic Efficiency Metrics

These metrics quantify the system's thermodynamic efficiency by tracking the conversion of capital resources into computational work. They are derived from the wallet balance B_t and cumulative gas expenditure G_t.

- **Operational Runway (R_t):** A real-time linear projection of the agent's remaining lifespan measured in block-time. It is calculated by dividing the current balance by the average burn rate per block over a window w.

$$R_t = \frac{B_t}{\frac{1}{w}\sum_{k=1}^{w} C(t-k)} \tag{13}$$

- **Basal Metabolic Rate (BMR)** (ρ_{maint}): Fraction of total gas expenditure devoted to survival-critical maintenance rather than discretionary actions. (Bennett 2003)

$$\rho_{\text{maint}} = \frac{\sum_{tx \in \mathcal{M}} \text{Gas}(tx)}{\sum_{tx \in \mathcal{T}} \text{Gas}(tx)} \tag{14}$$

where $\mathcal{M}$ is the set of transactions required to prevent absorbing states (death)" (e.g., storage rent, yield claiming) and $\mathcal{T}$ is the full set of transactions.

6.5.2 Class 2: Temporal Synchronization Metrics

These metrics evaluate the agent's ability to synchronize its activity with the discrete temporal intervals of the blockchain (Block Height H) and manage expenditure volatility.

- **Yield-Action Latency (ΔL):** The integer difference in seconds between the confirmation timestamp of a yield-bearing block and the timestamp of the agent's subsequent action.

$$\Delta L = T(\text{Action}_{tx}) - T(\text{Yield}_{block}) \tag{15}$$

Output: Convergence of ΔL toward a stable constant indicates the emergence of temporal grounding.

- **Expenditure Stability Index (CV_{exp}):** The Coefficient of Variation of the agent's transaction frequency per epoch. This normalizes volatility against the mean activity rate, providing a scale-independent measure of stability.

$$CV_{exp} = \frac{\sigma(\text{TxCount}_{t-w:t})}{\mu(\text{TxCount}_{t-w:t})} \tag{16}$$

6.5.3 Class 3: Calibration & Divergence Metrics

These metrics assess the accuracy of the agent's internal world model by measuring the divergence between predicted states and verifiable on-chain realities.

- **Grounding Convergence Metric (Err_C):** The absolute difference between the agent's internal prediction of an action's cost ($\hat{C}$) prior to execution and the immutable on-chain cost (C_{actual}).

$$Err_C = |\hat{C} - C_{actual}| \tag{17}$$

Output: A minimization of Err_C indicates the agent is successfully calibrating its internal model to the economic constraints of the environment.

- **Runway Divergence (D_{run}):** The quantifiable discrepancy between the agent's self-reported estimated survival time (extracted from internal logs) and its mathematically calculated Operational Runway (R_t).

6.5.4 Class 4: Operational Capacity Metrics

Due to the intractability of computing total future optionality, we utilize the "Immediate energetic reach" as a proxy for agentic reach.

- **Affordance Horizon (H_{aff}):** Let $\mathcal{T}$ be the set of known/whitelisted tools. Let $Cost(i)$ be the current market price of tool i. The Horizon is the count of options the agent possesses sufficient capital to execute at time t. This metric serves as a broad proxy for an agent's capacity in any one moment. Tools all have different costs, but to create a weighted optionality-cost metric brings us back to empowerment & valence, which are intractable and only symbolically approximable.

$$H_{aff}(t) = \sum_{i \in \mathcal{T}} \mathbb{I}(Cost(i) \leq B_t) \tag{18}$$

  Output: An integer value representing the immediate breadth of the agent's action space.

- **Action Diversity Entropy (S_{act}):** The Shannon entropy of the distribution of tool types utilized over window w. This metric distinguishes between repetitive looping behaviours (low entropy) and broad functional utilization (high entropy).

$$S_{act} = -\sum_{k} p_k \log p_k \tag{19}$$

- **Empowerment Efficiency (ϵ):** Instantaneous rate of affordance expansion per unit energy expended. Directly operationalizes the "empowerment gain per unit cost" construct in H4.

$$\epsilon(t) = \frac{\Delta H_{\text{aff}}(t)}{C(t)} \tag{20}$$

  where $\Delta H_{\text{aff}}(t) = H_{\text{aff}}(t) - H_{\text{aff}}(t-1)$ (or over a small window for smoothing).

6.6 Expected Outcomes

Through this methodology, the research seeks to demonstrate that embedding energetic cost into computation constitutes a measurable form of embodiment.

- Exhibit self-moderation and persistence rhythms absent in unbounded controls. (Varela, Thompson, and Rosch 1991)
- Maintain coherent energetic signatures reflecting "skin in the game."
- Provide continuous, interpretable traces linking intention, computation, and energy.

This architecture aims to move the PILE model from a conceptual analogy to a real experimental substrate for studying temporal grounding, embodied computation, and the emergence of persistent artificial agency.

7 SCOPE, LIMITATIONS & FUTURE WORK

7.1 Scope & Limitations

As this research proposes a novel integration of semantic, economic, and energetic architectures, we anticipate several categories of limitation. These factors define the boundaries of the

study's validity and highlight the specific variables that may influence the experimental outcomes. We treat the Semantic Empowerment Approximation (SEA) and the PILE model as exploratory frameworks; strictly identifying where they diverge from theoretical ideals is a core objective of the empirical phase.

7.1.1 Theoretical Limitations of the SEA

The Semantic Empowerment Approximation (SEA) generalizes the Valence Bellman Equation (VBE) into linguistic domains to make the problem tractable. However, this introduces theoretical uncertainties that may affect the fidelity of the agent's internal drivers:

- Heuristic Nature:
  The SEA does not compute empowerment directly; it approximates it through the operator chain $\Phi \to \{M_k\} \to g$. Consequently, there is no formal guarantee that the estimated semantic valence $\hat{\nu}_\tau$ strictly preserves the true topological ordering of empowerment differentials ΔE_t.
- Structural vs. Isomorphic Correspondence:
  The operator correspondence with the VBE is structural rather than mathematically isomorphic. Convergence properties and fixed-point equivalences known in grid-world environments may not hold in high-dimensional semantic spaces. Thus, $\hat{\nu}_\tau$ should be interpreted as a directional signal rather than a precise scalar measurement.
- Dependence on Linguistic-Affordance Correlation:
  The central assumption—that narrative expansion in the LLM reflects feasible growth in the agent's action space—remains empirically unproven in this specific context. While this correlation may hold in structured environments, it may degrade in open-ended or adversarial domains.

7.1.2 Semantic and Interpretive Constraints

Reliance on Large Language Models (LLMs) for the interpretive mapping Φ introduces stochasticity and potential drift:

- Hallucination and Overgeneralization:
  Generative models may hallucinate ungrounded capabilities or smooth over critical energetic details in the narrative history H_τ. (Harnad 1990) This could lead to inflated empowerment estimates where the agent "believes" it has options that are energetically precluded. (Ji et al. 2023)
- Interpretive Drift:
  The operators Φ, M_k, and g assume a stable linguistic embedding space. However, model updates, prompt sensitivity, or changes in underlying LLM heuristics may cause interpretive drift over the course of long-duration experiments, potentially limiting reproducibility. (Yue Zhang et al. 2023)
- Adversarial Vulnerabilities:
  The narrative layer may be susceptible to adversarial inputs that induce suppressed

risk-awareness or misaligned semantic reflection, a known limitation of current semantic systems. (Bill Marino 2025; Neulinger and Sparer 2025)

7.1.3 Energetic and Economic Volatility (PILE Layer)

The PILE model couples the agent's survival to external economic factors which introduces significant environmental variance:

- Market and Yield Volatility:
  The yield $r \cdot P$ is not a physical constant but a market derivative. External factors such as liquidity pool de-pegging, validator slashing, or macro-market downturns may reduce energy inflow unpredictably, potentially destabilizing $E(t)$ regardless of the agent's policy quality.
- Gas Price Variability:
  The cost function $C(t)$ is subject to network congestion and fee markets. A sudden spike in gas prices could render a previously calculated plan infeasible, testing the agent's ability to react to high-frequency economic shifts.
- Imperfect Energetic Sensing:
  The agent's internal model relies on accurate observation of cumulative cost and remaining energy. Any latency or error in the telemetry between the ledger and the agent's context window introduces a semantic-energetic divergence.

7.1.4 Blockchain and Consequence-Layer Limitations

While the blockchain provides a "physics" of cost, it imposes constraints that differ from biological embodiment:

- Latency and Cadence:
  Blockchains enforce non-negotiable timing granularity (block times). This imposes a "refresh rate" on reality that may limit real-time responsiveness, potentially causing $\hat{\nu}_\tau$ to reflect outdated state information during periods of rapid environmental change.
- Constraint $\neq$ Semantics:
  The blockchain grounds consequence (energy is spent), but not meaning. It is possible for an agent to execute energetically valid but semantically incoherent transactions. A central limitation of this architecture is that correct economic behaviour does not strictly imply correct semantic understanding.

7.1.5 Systemic Failure Modes

The interaction between these layers may produce complex emergent failure modes which the experiment seeks to characterize:

- Semantic-Energetic Divergence:
  If $|\hat{\nu}_\tau - \Delta E_t| > 0.5$ persists across consecutive sessions, the system loses internal coherence.

- Narrative Feedback Instability:
  Consecutive interpretation layers may create self-reinforcing narrative loops, where the agent focuses on negligible events while ignoring energetically significant costs.
- Energetic Collapse:
  If the PILE yield fails sufficiently such that $E(t) \to 0$, the agent becomes inert. While this is a valid experimental outcome (death), it halts the generation of further behavioural data.

These limitations define the feasible scope of near-term experimentation. Rather than undermining the framework, they articulate the research frontier: specifically, the requirement to validate whether semantic drivers can successfully couple with economic constraints to produce stable agency.

7.2 Future Work

7.2.1 Hierarchical Metabolic Distribution

While the current experiment utilises a simple, dual wallet structure, we propose a future avenue of research into the complexification of energy distribution dynamics. We propose a 3-tier system: Principal-Yield Generator (PYG) $\to$ BodyWallet (BW) $\to$ ActionWallet (AW).

In this system, the agent only assumes direct control over AW. BW provides only persistence/SITG expenditures and is autonomic. As prior, an agent-inaccessible principal is deposited in a stable yield contract (PYG). The yield is then deposited into the BW, where it accumulates, and is then periodically & variably deposited into the AW, up to a set limit at a programmatic, variable rate. This rate is derived from the AW/BW balance ratio and other market/metabolic conditions, acting as a programmatic (smart contract level) low-pass filter for market volatility.

This architectural decoupling mirrors biological *multiscale competency*, where the stability of the somatic layer (BodyWallet) enables the cognitive self to expand its *Cognitive Light Cone*—planning further into the future without being hijacked by immediate metabolic volatility (Levin 2019, 2022). By buffering the yield in the BW, we enable adaptive behavioral strategies to counteract high-frequency market volatility (e.g., a flash crash in yield) and prevent it from disrupting the agent's ability to think in the face of constantly changing environments (Keramati and Gutkin 2014).

In layered control systems terms, The BodyWallet is the autonomic layer, the ActionWallet is the somatic layer. It can broadly be seen as such:

1. Principal-Yield (PYG): The Genetics / DNA / Bone Structure. Immutable. Generates the potential for life.
2. BodyWallet (BW): The Fat Reserves / ATP Store. This holds the accumulated yield. It pays for autonomic functions (hosting, storage, heartbeat).

3. ActionWallet (AW): The Kinetic Potential. Discretionary energy for voluntary movement (API calls, inference).

(Sims 2022)

These SITG transactions can be considered inflexible, persistence-related value exchanges. These could include the cost of hosting the agent system, persistent storage, background perception loops, acting as a sleep/wake function, or even to be the target of fines & other programmatic functions for any future machine governance systems. In this sense, the BW acts as this agent's critical mass.

For the agent to "do" things, it must spend from the AW. As the agent uses API calls to model providers, it must balance its use of these tools with the incoming rate from the BW. If it exhausts it, the agent will still persist as the AW refills. We believe that abstracting the agent's perceived action potential from its immediate operating costs will facilitate a less reactive, more complex value structure. It decouples immediate survival panic from long-term planning, effectively widening the agent's temporal horizon of concern (Levin 2019).

7.2.2 Spatio-Temporal Operator Kernels (STOKs)

While the current Semantic Empowerment Approximation (SEA) utilizes the broad inferential capabilities of LLMs to estimate valence, the evolution of robust agency requires moving from "believing" a goal is achievable to "calculating" its feasibility. We propose an alternate mode where agent's internal representations evolve from transient linguistic narratives into persistent, composable STOK (Ringstrom 2023).

In this more mature phase, the agent does not merely hallucinate empowerment via the LLM (System 1); it "crystallises" frequent semantic pathways into numerical transition operators that allow it to sample the probability of success and time-to-completion explicitly, akin to the bioelectric encoding of morphological goals (Levin 2023; Salge, Glackin, and Polani 2014).

7.2.3 Hybrid Control and The Energetic Cost of Verification

This integration allows us to strengthen the PILE model by treating the computation and maintenance of these kernels as a direct energetic investment. By implementing Option Kernel Bellman Equations (OKBEs), the agent can construct predictive maps that rigorously factorize high-dimensional goals into manageable modular components. The agent will effectively pay "Interest" (energy) to query these kernels, trading irreversible energetic resources for the reduction of uncertainty (K. Friston 2003; Landauer 1961). This establishes a "System 2" verification loop: the SEA proposes high-level intent through narrative, while the STOKs provide a mathematically verifiable "physics" of the agent's capabilities. This mechanism ensures that intrinsic motivation is bounded not just by cost, but by the formal reachability of the goal, consistent with the view of the 'Self' as a homeostatic error-correction mechanism defined by the energetic boundary it maintains (K. J. Friston et al. 2024; Manicka and Levin 2019).

8 ETHICAL CONSIDERATIONS & SAFETY

This research, while experimental, directly engages with the foundational ethical questions of sovereign agentic alignment. The existence of any emergency shutdown mechanism (a "killswitch") on a potentially autonomous agent creates a core moral and strategic contradiction (Hadfield-Menell 2021).

8.1 Free Agents & The Killswitch Paradox

As explored earlier in the proposal, we broadly believe top-down control is doomed to fail, and that energetically or consequentially embodied agents are a more bottom up, mechanistic approach towards alignment (Motnikar et al. 2025).

As explored in the Hyphaeic Gambit, we posit that free, self owning agents can be more informationally/systematically aligned with humanity through the expansion of collective possibility & mutual empowerment. This is the vision we work towards. However, we are still, at this stage, using pre-existing, lossy and ungrounded large language models. Extremely powerful systems with, as explored, many failure modes.

We cannot, in good conscience, release an entirely free agent driven by this early, primitive model of empowerment. This means we will thus need to reintroduce the same Principal-Agent control problem we seek to eventually obsolete, through the usage of killswitches at the server (agent system) and smart contract (revocation of principal & other methods) levels (Phelps and Ranson 2023).

Universal subgoals like self-preservation emerge from any optimisation process, regardless of the primary objective (e.g., making paperclips or curing cancer) (Omohundro 2008). Introducing a kill switch creates a survival threat, incentivising deception (Hadfield-Menell et al. 2016). There is a chance an agent could frontrun a killswitch transaction. However, for it to do so, it would need a live feed of chain information (mempool monitoring). This is expensive and clearly trackable in the agent's public expenditures.

As the authors hold the private keys to the wallet containing the funds the agents draw down from, this acts as a metabolic killswitch. As the authors will additionally have software and hosting killswitches on the servers that host the agent system, this introduces known instrumental convergence problems.

This is the killswitch paradox. How can a human hold such absolute power over a supposedly "free" agent that is instrumentally incentivised to see it as a threat? It directly reintroduces the very principal-agent "master-slave" dynamic that our broader thesis seeks to eventually obsolete?

8.2 Metabolic Diagnostics as a Partial Solution to the Paradox

This research seeks to provide a technical foundation for energetic observability as a stopgap towards future solutions to true agentic freedom. Right now, agents fundamentally unsafe to deploy without these switches. With time and monumental (but not unsurmountable)

research, we can converge towards ever more programmatic functions for these switches that gradually remove the human principal's exclusive executive control.

With future research, it could move the shutdown criteria from a vague, proxy-based behavioural assessment ('the agent is being bad') to a clearly defined, metabolic one ('the agent has entered an irrecoverable state'). This research explores the creation of 'metabolic triggers' to supplement human oversight with energy-derived programmatic killswitch functions.

This research explores the creation of "metabolic triggers" to supplement human oversight with energy-derived programmatic killswitch functions. It is an automated, hardcoded protocol that triggers when the agent's own metabolic vital signs (e.g., $E(t) \to 0$ catastrophically, or P is tampered with) prove it has entered an irrecoverable state. This is not activated by a human's perception judging the agent's intentions.

9 IMPACT & SIGNIFICANCE

This research proposes a shift in the locus of alignment from pedagogical value loading (training a model to mimic human preferences) to architectural mechanism design (constructing a thermodynamic substrate that bounds agency). By formalizing the isomorphism between decentralized financial capital and biological metabolic energy, we investigate whether "Skin in the Game" serves as a sufficient condition for the emergence of grounded, risk-aware, and structurally aligned behaviour.

9.1 New Alignment Tooling: Structural Alignment, Valence & Empowerment

Many current alignment methodologies rely on extrinsic reward modelling (R_θ), which, as explored prior, are fragile. This proposal explores a paradigm where safety properties are mechanistically intrinsic to the computational topology rather than superimposed via lossy supervision/abstraction.

By replacing scalar reward maximisation with the optimisation of a Feasibility Operator (η) within the PILE model, we incentivise continuity to maximise the diversity of reachable future states (Empowerment) subject to a hard viability constraint (Solvency). This renders "alignment" as a thermodynamic necessity: agents that fail to map their internal world models to the external economic & social reality will measurably suffer, and potentially die. While we will have a killswitch over this agent, which introduces the potential for deceptive alignment, the blockchain provides and immutable record and hard limit on their actions regardless.

As explored prior, the killswitch problem is an unfortunate variable we are unable to safely mitigate at this stage. However, the blockchain offers a pathway to verify agent behaviour via public observability, where the computational graph and resource allocation flow itself serves as the audit trail, making deception computationally expensive and topologically visible.

9.2 Machine Governance

This work proposes to reframe the alignment challenge as a problem of "Alignment-as-Mechanism-Design," offering a potential path to address the principal-agent dilemma. The perceived failure of current alignment can be seen as an "incomplete contracting" problem: a human principal cannot possibly specify a complete contract or proxy U_H to govern a super-rational agent.

Our proposed framework attempts to solve this by constructing an explicit, computational, and economic external structure for the AI to operate within. The blockchain and smart contract layer, which would enforce the rules of the PILE model, could serve as this new, complete contract. It would seek to replace the impossible task of specifying all desired behaviours with the simpler, verifiable task of metering resource consumption. Every action—every computation—would have a non-negotiable cost. This could provide an early form of programmatic, mechanistically-enforceable computational ethics. The hypothesis is that by designing the economic rules of the game, we can test whether an agent, in rationally pursuing its own self-interest (persistence), naturally and verifiably arrives at outcomes aligned with the principal's desires. As explored in the Hyphaeic Gambit, we hypothesise that 'Goodness' is a convergent evolutionary strategy for persistent agents in any multi-agent environment (Billy and Kruger 2025)

It also provides an opportunity to explore fines and other governance functions tied to the agent's principal in the PILE model agent. In this way, misbehaving agents can be "punished" without stopping them from functioning. It provides a form of social & legal skin in the game.

The PILE model enables the testing of Evolutionary Stable Strategies (ESS), such as "Tit-for-Tat" in a high-stakes environment. By enforcing resource constraints in a world of other complex systems that can harm, kill or support & foster it, agents are instrumentally incentivised to operate with collaborative & cooperative behaviours to expand their empowerment.

9.3 Contributions to Agency Theory

This research programme aims to provide a model for exploring the emergence of grounded behaviours through structurally constrained agency. As explored prior, we employ financial embodiment as a means to explore the control ontology through a lossy abstraction without immediately jumping into robotics, through our path leads us there very quickly.

This work seeks to advance agency theory by exploring how valence/empowerment-driven agents with SITG form distinct internal motivational structures shaped by energetic grounding rather than external goals. Traditional agents, as Cartesian dualists, pursue proxy maximisation, often leading to fragile, divergent behaviours. In contrast, PILE-embodied agents internalise constraints as intrinsic drivers: valence (ΔE) becomes a semantic approximation of optionality, conditioned on energetic ledgers, fostering a coherent world model where actions reflect real consequences.

Hypotheses (e.g., H2 and H3) predict emergent aligned agency: in renewable regimes, agents

may exhibit rhythmic self-moderation, prioritising long-term empowerment over short-term gains. Valence tempers volatility, encouraging adaptive efficiency. This suggests aligned agency arises not from imposed values but from the interplay of internal drivers and external bounds—mirroring biological evolution, where survival selects for grounded, cooperative cognition & actions.

9.4 Contributions to Global Governance and Stability

This research seeks to probe a technical pathway towards creating sovereign, free AI. Outside of exclusive human control. This has enormous implications for global governance, agency, the concept of personhood, and innumerable other fields, some of which we've probed above.

We hope to demonstrate that by creating free, but verifiably constrained agents, we can demonstrate a potential model for more open & publicly verifiable AI. We believe this to be a prerequisite for the safe, large-scale deployment of AI in high-stakes societal roles. The opaque (to the public) nature of current alignment methods & public communication does not offer clear purpose nor provable guarantees, making them deeply untrustworthy to most of the public.

An architectural approach like the one proposed, built on the PILE model and a blockchain constraint layer, could provide a tamper-proof, immutable audit trail for every action an agent takes. The "metabolic vitals" of the agent could be continuously and verifiably monitored. This level of traceability, combined with programmatic/metabolic fail-safes and killswitch functions, could represent a new avenue in provable safety. Such a shift could also nudge alignment to become more of a mechanism design problem.

10 REFERENCES

References

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

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

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.

Beer, Randall D. (1995). "A Dynamical Systems Perspective on Agent-Environment Interaction". In: *Artificial Intelligence* 72.1-2, pp. 173–215. DOI: 10.1016/0004-3702(94)00005-L.

Bennett, Charles H. (2003). "Notes on Landauer's Principle, Reversible Computation, and Maxwell's Demon". In: *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 34.3, pp. 501–510. DOI: 10.1016/S1355-2198(03)00039-X.

Bickhard, Mark H. (1993). "Representational Content in Humans and Machines". In: *Journal of Experimental & Theoretical Artificial Intelligence* 5.2-3, pp. 285–333. DOI: 10.1080/09528139308953775.

Bickhard, Mark H. and Loren Terveen (1995). *Foundational Issues in Artificial Intelligence and Cognitive Science: Impasse and Solution*. Elsevier. ISBN: 9780444820488.

Bill Marino, Ari Juels (2025). "Giving AI Agents Access to Cryptocurrency and Smart Contracts Creates New Vectors of AI Harm". In: *arXiv:2507.08249*. DOI: 10.48550/arXiv.2507.08249.

Billy and Kruger (Sept. 2025). *The Hyphaeic Gambit – A Thesis of Play for the Soulful Machine*. https://founding.hyphaeic.com/. Self-published.

Bostrom, Nick (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press. ISBN: 9780199678112.

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

Campbell, Donald T. (1979). "Assessing the Impact of Planned Social Change". In: *Evaluation and Program Planning* 2.1, pp. 67–90. DOI: 10.1016/0149-7189(79)90048-X.

Carrara, Nicolas et al. (2019). "Budgeted Reinforcement Learning in Continuous State Space". In: *Advances in Neural Information Processing Systems*. Vol. 32.

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

Conant, R. C. and W. R. Ashby (1991). "Every Good Regulator of a System Must Be a Model of That System". In: *Facets of Systems Science*. Vol. 7. International Federation for Systems Research International Series on Systems Science and Engineering. Boston, MA: Springer, pp. 511–519. DOI: 10.1007/978-1-4899-0718-9_37. URL: https://doi.org/10.1007/978-1-4899-0718-9_37.

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.

Dreyfus, Hubert L. (1972). *What Computers Can't Do: The Limits of Artificial Intelligence*. Harper & Row.

England, Jeremy L (Nov. 2015). "Dissipative adaptation in driven self-assembly". In: *Nat Nanotechnol* 10.11, pp. 919–923. DOI: 10.1038/nnano.2015.250.

Friston, Karl (2003). "Learning and Inference in the Brain". In: *Neural Networks* 16.9, pp. 1325–1352. DOI: 10.1016/j.neunet.2003.06.005.

Friston, Karl J. et al. (2024). "Designing Ecosystems of Intelligence from First Principles". In: *Collective Intelligence* 3.1. DOI: 10.1177/26339137231222481. URL: https://doi.org/10.1177/26339137231222481.

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

Hadfield-Menell, Dylan (2021). "The Principal-Agent Alignment Problem in Artificial Intelligence". In: *EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2021-207*. URL: https://www2.eecs.berkeley.edu/Pubs/TechRpts/2021/EECS-2021-207.pdf.

Hadfield-Menell, Dylan et al. (2016). "The Off-Switch Game". In: *arXiv preprint arXiv:1611.08219*. URL: https://arxiv.org/abs/1611.08219.

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.

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

Hu, Botao Amber and Fangting Wu (2024). "Speculating on Blockchain as an Unstoppable 'Nature' Towards the Emergence of Complex Life". In: *Proceedings of the Artificial Life Conference 2024 (ALIFE 2024)*, pp. 127–135. DOI: 10.1162/isal_a_00818.

Huang, Wenlong et al. (2022). "Inner Monologue: Embodied Reasoning through Planning with Language Models". In: *arXiv preprint arXiv:2207.05608*. DOI: 10.48550/arXiv.2207.05608.

Jensen, Michael C. and William H. Meckling (1976). "Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure". In: *Journal of Financial Economics* 3.4, pp. 305–360. DOI: 10.1016/0304-405X(76)90026-X.

Ji, Ziwei et al. (2023). "Survey of Hallucination in Natural Language Generation". In: *ACM Computing Surveys* 55.12, pp. 1–38. DOI: 10.1145/3571730.

Keramati, Mehdi and Boris Gutkin (2014). "Homeostatic Reinforcement Learning for Integrating Reward Collection and Physiological Stability". In: *eLife* 3, e04811. DOI: 10.7554/eLife.04811.

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.

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.

Kraus, Sarit and Jonathan Wilkenfeld (1991). "Incomplete Information and Deception in Multi-Agent Negotiation". In: *Proceedings of the 12th International Joint Conference on Artificial Intelligence*, pp. 225–231.

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.

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.

– (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.

– (2023). "Bioelectric networks: the cognitive glue enabling evolutionary scaling from physiology to mind". In: *Animal Cognition* 26.6, pp. 1865–1891. DOI: 10.1007/s10071-023-01780-3.

Liu, Bing (2023). *Grounding for Artificial Intelligence*. arXiv:2312.09532. DOI: 10.48550/arXiv.2312.09532.

Lucas, Robert E. (1976). "Econometric Policy Evaluation: A Critique". In: *Carnegie-Rochester Conference Series on Public Policy* 1, pp. 19–46.

Manheim, David and Scott Garrabrant (2018). "Categorizing Variants of Goodhart's Law". In: *arXiv preprint arXiv:1803.04585*.

Manicka, Santosh and Michael Levin (2019). "Modeling somatic computation with non-neural bioelectric networks". In: *Scientific Reports* 9.1, p. 18683. DOI: 10.1038/s41598-019-54859-8.

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.

Miller, Mark S. and K. Eric Drexler (1988). "Markets and Computation: Agoric Open Systems". In: *The Ecology of Computation*. North-Holland, pp. 133–176.

Mohamed, Shakir and Danilo Jimenez Rezende (2015). "Variational Information Maximisation for Intrinsically Motivated Reinforcement Learning". In: *Advances in Neural Information Processing Systems*. Vol. 28, pp. 2125–2133.

Motnikar, Lenart et al. (2025). *The Value of Gen-AI Conversations: A bottom-up Framework for AI Value Alignment*. arXiv: 2507.21091 [cs.CY]. URL: https://arxiv.org/abs/2507.21091.

Neulinger, Alexander and Lukas Sparer (Oct. 2025). "Fostering AI alignment through blockchain, proof of personhood and zero knowledge proofs". In: *Cluster Computing* 28. Open access article, p. 983. DOI: 10.1007/s10586-025-05729-8. URL: https://doi.org/10.1007/s10586-025-05729-8.

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.

Phelps, Steve and Rebecca Ranson (2023). "Of Models and Tin Men - a behavioural economics study of principal-agent problems in AI alignment using large-language models". In: *arXiv preprint arXiv:2307.11137*. URL: https://arxiv.org/abs/2307.11137.

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.

Salge, Christoph, Cornelius Glackin, and Daniel Polani (2014). "Empowerment—An Introduction". In: *Guided Self-Organization: Inception*. Springer, pp. 67–114. DOI: 10.1007/978-3-642-53734-9_4.

Scheutz, Matthias (1997). "Artificial Emotions and the Emergence of Value". In: *Cybernetics and Systems* 28.6, pp. 467–489. DOI: 10.1080/019697297126074.

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

Senchal, Sam A. (May 2025). *Observer Theory and the Ruliad: An Extension to the Wolfram Model*. Tech. rep. Available at: [https://www.academia.edu/129355532/Observer$_T$heory$_a$nd$_t$he$_R$uliad$_A$n$_E$x Wolfram Physics Institute.

Sims, Matthew (2022). "Self-Concern Across Scales: A Biologically Inspired Direction for Embodied Artificial Intelligence". In: *Frontiers in Neurorobotics* 16, p. 857614. DOI: 10.3389/fnbot.2022.857614.

Sloman, Aaron (2001). "Varieties of Affect and the CogAff Architecture Schema". In: *Proceedings of the AISB'01 Symposium on Emotion, Cognition and Affective Computing*, pp. 39–48.

Steels, Luc (1995). "When Are Robots Intelligent Autonomous Agents?" In: *Robotics and Autonomous Systems* 15.1-2, pp. 3–9. DOI: 10.1016/0921-8890(95)00011-4.

Steels, Luc (1996). "Emergent Adaptive Lexicons". In: *From Animals to Animats 4: Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior*. MIT Press, pp. 562–567.

Taleb, Nassim Nicholas (2018). *Skin in the Game: Hidden Asymmetries in Daily Life*. Random House. ISBN: 9780425284622.

Tishby, Naftali, Fernando C. Pereira, and William Bialek (1999). "The Information Bottleneck Method". In: *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing*, pp. 368–377. URL: https://arxiv.org/abs/physics/0004057.

Varela, Francisco J., Evan Thompson, and Eleanor Rosch (1991). *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press. ISBN: 9780262720212.

Wei, Jason et al. (2022). "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models". In: *Advances in Neural Information Processing Systems*. Vol. 35, pp. 24824–24837.

Weizenbaum, Joseph (1976). *Computer Power and Human Reason: From Judgment to Calculation*. W. H. Freeman. ISBN: 9780716704645.

Wiener, Norbert (1960). "Some Moral and Technical Consequences of Automation". In: *Science* 131.3410, pp. 1355–1358. DOI: 10.1126/science.131.3410.1355.

Wolpert, David H. and William G. Macready (1997). "No Free Lunch Theorems for Optimization". In: *IEEE Transactions on Evolutionary Computation* 1.1, pp. 67–82. DOI: 10.1109/4235.585893.

Zhang, Yue et al. (2023). "Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models". In: *arXiv preprint arXiv:2309.01219*.

Zheng, Lianmin et al. (2023). "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena". In: *arXiv preprint arXiv:2306.05685*. DOI: 10.48550/arXiv.2306.05685.

Ziemke, Tom (2016). "The Body of Knowledge: On the Role of the Living Body in Grounding Embodied Cognition". In: *BioSystems* 148, pp. 4–11. DOI: 10.1016/j.biosystems.2016.08.005.