Environmental Superintelligence

as the Missing Foundation of AI Alignment

A First-Principles Thermodynamic Analysis

Jed Anderson

Founder & CEO, EnviroAI

Houston, Texas, USA

jed@enviro.ai

March 2026

Keywords: AI alignment, environmental superintelligence, negentropic alignment, Bond-Bit Asymmetry, thermodynamic coherence, Landauer principle, Generalized Functional Efficiency, AI safety, planetary boundaries

Abstract

The AI alignment problem remains unsolved. The Future of Life Institute's Summer 2025 AI Safety Index found no frontier AI company scored above D on existential safety preparedness, with reviewers concluding that alignment rhetoric 'has not yet translated into quantitative safety plans, concrete alignment-failure mitigation strategies, or credible internal monitoring and control interventions.' This paper argues that a foundational reason for this failure is that all dominant alignment approaches—reinforcement learning from human feedback (RLHF), Constitutional AI (CAI), mechanistic interpretability, scalable oversight, AI control, and brain-computer interface merger—share a critical limitation: they operate within an exclusively anthropocentric frame that lacks physically grounded optimization targets.

We propose that building Environmental Superintelligence (ESI)—AI that models, predicts, and optimizes Earth's physical systems—provides the missing foundation layer for AI alignment. We establish this claim through seven independent lines of evidence grounded in first-principles physics: (1) nature's information content exceeds all AI training data by a ratio of 10^{20}–10^{35}, constrained by exact conservation laws that internet data lacks; (2) the Bond-Bit Asymmetry (experimentally verified, practical ratio ~10^{10} today, approaching 10^{20} at Landauer limit) creates a structural economic incentive for physics-grounded AI to prefer prevention over destruction, with this incentive growing monotonically as computational costs decline; (3) the set of planetary states compatible with human welfare is a proper subset of those compatible with ecosystem health ($H \subset E$), making ecocentric optimization the strictly safer target; (4) evolution constitutes a 3.8-billion-year alignment test suite whose constraint structure—conservation laws as non-negotiable rules—encodes proven multi-agent coordination solutions alongside competitive strategies, both tested under exact physical law; (5) the entropy/negentropy framework provides objective, physically falsifiable alignment criteria that are resistant to Goodharting because conservation laws require closed-system accounting, making local gaming physically detectable; (6) Generalized Functional Efficiency ($GFE = F/(\dot{S}\cdot M)$) provides a quantitative alignment metric validated across 50 orders of magnitude and 13.8 billion years of cosmic history; and (7) ESI aligns AI with the observable cosmic trajectory from pure dissipation toward pure function—the deepest optimization pattern in physics.

We comprehensively assess every major AI safety institution and approach, identifying a common gap: none defines alignment in terms of physical law, none references conservation laws or planetary boundaries, and none trains on nature's data. We demonstrate that the cognitive structures required to build ESI—constraint respect, system-level reasoning, multi-agent coordination, and long time horizons—are transferable alignment properties that address the general alignment problem, not merely a domain-specific application. We conclude that ESI is not only a system for protecting Earth's biosphere but the most consequential missing piece of the AI safety infrastructure.

1. Introduction

1.1 The Alignment Problem and Its Current Status

The AI alignment problem—ensuring that increasingly capable artificial intelligence systems act in accordance with human intentions and the long-term viability of complex life—was formalized across several seminal works. Bostrom's *Superintelligence* (2014) established the canonical framework for existential risk from misaligned superintelligent AI, introducing the orthogonality thesis and instrumental convergence [1]. Christian's *The Alignment Problem* (2020) documented practical gaps between human intent and AI behavior across deployed systems [2]. Russell's *Human Compatible* (2019) proposed inverse reward design—AI that is uncertain about human preferences and actively seeks to learn them—as a technical pathway beyond fixed reward functions [3].

Despite over a decade of concentrated research and billions of dollars in investment, the alignment problem remains unsolved. The Future of Life Institute's Summer 2025 AI Safety Index evaluated seven frontier AI companies across 33 indicators spanning six critical domains. The findings were unambiguous: no company scored above a D on existential safety. The panel concluded that 'the industry is fundamentally unprepared for its own stated goals' and that alignment rhetoric 'has not yet translated into quantitative safety plans, concrete alignment-failure mitigation strategies, or credible internal monitoring and control interventions' [4a]. Stuart Russell, a member of the expert panel, stated: 'We are spending hundreds of billions of dollars to create superintelligent AI systems over which we will inevitably lose control. We need a fundamental rethink of how we approach AI safety' [4a].

1.2 The Central Thesis

This paper proposes that fundamental rethink. We argue that the failure of current alignment approaches is not one of sophistication but of scope. All dominant approaches operate within an anthropocentric frame: they seek to align AI with human preferences, human values, human cognition, or human-written principles. None references the physical laws that govern the planetary system hosting all computation. None trains on the data generated by Earth's biosphere. None optimizes for the thermodynamic conditions that sustain complex life.

We propose that building Environmental Superintelligence (ESI)—AI that models, predicts, and optimizes Earth's physical systems in real time—provides the missing foundation layer for AI alignment. ESI is not merely a domain application of AI to environmental problems. It is a fundamentally different approach to grounding AI in physical reality, with alignment properties that transfer to the general case. An AI that has internalized conservation laws, system stability, multi-agent coordination across deep time, and measurable ground truth is better aligned in general—not just for environmental tasks.

1.3 Scope and Structure

Section 2 comprehensively reviews every major AI safety approach and institution. Section 3 develops the theoretical framework of Negentropic Alignment from first-principles thermodynamics, including the cosmic optimization trajectory and divergence proof. Section 4 presents quantitative analyses

supporting each claim, including Generalized Functional Efficiency as a quantitative alignment metric. Section 5 demonstrates how ESI provides the specific alignment properties that current approaches lack. Section 6 addresses limitations and objections. Section 7 concludes with implications for the field.

2. Comprehensive Review of AI Alignment Approaches

We survey the complete landscape of AI alignment research, organized by methodology. For each approach, we identify its contribution, mechanism, and the specific limitation that Negentropic Alignment addresses.

2.1 Value Learning Approaches

2.1.1 Reinforcement Learning from Human Feedback (RLHF)

RLHF, formalized by Christiano et al. (2017) [5] and widely deployed by OpenAI, Anthropic, and Google DeepMind, trains a reward model from human preference comparisons and optimizes AI behavior against this learned reward signal. Its contribution is a scalable, empirically effective technique for shaping model behavior toward human-preferred outputs.

Its limitations are well-documented. RLHF is vulnerable to Goodharting: the AI may optimize the learned reward proxy rather than the underlying human intent [6]. More fundamentally, RLHF encodes the preferences of whatever humans provide feedback. If raters do not value ecosystem health, neither will the resulting AI. The reward model contains no physics, no conservation laws, and no ecological constraints.

2.1.2 Constitutional AI (CAI)

Constitutional AI, developed by Bai et al. (2022) at Anthropic [7], replaces human raters with a set of principles. The model generates responses, self-critiques against the constitution, revises, and is trained via 'RL from AI Feedback.' CAI reduces dependence on human raters and enables principle-based alignment that is more consistent and auditable.

The limitation: the constitution is still authored by humans, encoding anthropocentric values. No current CAI constitution includes conservation of mass, conservation of energy, planetary boundary constraints, or ecosystem health metrics. Adding physical-law principles to a CAI constitution would be a direct application of Negentropic Alignment.

2.1.3 Inverse Reward Design and Cooperative IRL

Russell's framework of cooperative inverse reinforcement learning (CIRL) [3] proposes AI that is uncertain about human preferences and actively seeks to learn them through interaction. This is a significant advance over fixed reward functions. The limitation remains anthropocentric: the AI learns what humans want, which historically has not included what ecosystems need. CIRL provides no

mechanism for incorporating non-human interests or physical constraints unless humans explicitly specify them.

2.2 Interpretability Approaches

2.2.1 Mechanistic Interpretability

Pioneered by Chris Olah and the Anthropic interpretability team [8], mechanistic interpretability seeks to reverse-engineer the internal algorithms learned by neural networks. It is the most promising path to understanding AI's internal representations and is essential for detecting deceptive alignment. However, it is a diagnostic tool, not an alignment target. It tells us what a model is doing but not what it should be doing. Negentropic Alignment provides candidate criteria: an internal representation that respects conservation laws and models system stability is more aligned than one that does not.

2.3 Governance and Control Approaches

2.3.1 The Future of Life Institute (FLI)

FLI, founded in 2014, has been the most prominent institutional voice for AI safety governance [4a, 4b]. FLI operates primarily at the governance and policy level, providing essential accountability infrastructure. Its 33-indicator (Summer 2025) and 35-indicator (Winter 2025) evaluation frameworks assess risk management, safety frameworks, and existential preparedness—but environmental and ecological alignment is absent from all indicators.

2.3.2 MIRI, ARC, and AI Control

The Machine Intelligence Research Institute (MIRI), founded by Eliezer Yudkowsky in 2000, contributed foundational concepts including corrigibility, decision theory, and deceptive alignment [9]. The Alignment Research Center (ARC), led by Paul Christiano, focuses on scalable alignment through elicitation and evaluation [10]. The Center for AI Safety (CAIS) focuses on catastrophic risk reduction. These organizations provide essential control and evaluation infrastructure. Their limitation is structural: they focus on constraining AI behavior (defense) rather than grounding AI cognition (foundation).

2.3.3 Scalable Oversight, Debate, and BCI Merger

Scalable oversight frameworks [11] use AI systems to help humans oversee other AI systems. Brain-computer interface merger (Neuralink) posits that merging human cognition with AI resolves alignment by transmitting values directly. Both inherit the anthropocentric limitation: if human evaluators do not incorporate ecological criteria, neither will the oversight system; if humans have been unable to prioritize long-term ecological health, amplifying that cognition through BCI provides no structural guarantee of improvement.

2.4 Comparative Assessment

Table 1 summarizes the complete landscape. The pattern is consistent: every current approach lacks a physical basis for its alignment targets and incorporates no ecological constraints.

Table 1. Comparative assessment of AI alignment approaches across six evaluation criteria.

Approach	Mechanism	Target	Failure Mode	Physics	Eco	Falsif.?	Goodhart-Resistant?
RLHF	Reward model	Human prefs	Goodharting	None	None	No	No
CAI	Self-critique	Written principles	Surface compliance	None	None	No	No
CIRL/CHAI	Pref. learning	Inferred prefs	Anthropocentric	None	None	No	No
Mech. Interp.	Rev. engineering	Diagnostic only	No target	None	None	N/A	N/A
Oversight	Recursive eval	Human judgment	Human bias	None	None	No	No
AI Control	Containment	Controllability	Defensive only	None	None	No	No
BCI Merger	Neural integ.	Human cognition	Bias inheritance	None	None	No	No
ESI/Negentropic	Physics grounding	Thermo. coherence	Def. complexity	**Full**	**Full**	**Yes**	**Yes (conservation laws)**

3. Theoretical Framework: Negentropic Alignment

3.1 Definitions

We adopt the computational continuum framework [12]. Three computational layers operate on Earth's planetary substrate: C_univ (natural/physical systems), C_bio (biological/human computation), and C_art (artificial computation). Each layer has an implicit optimization trajectory.

Definition 1 (Alignment): The computational layers are aligned when their optimization gradients are synergistic: $\nabla C_{art} \parallel \nabla C_{bio} \parallel \nabla C_{univ}$. The work of each layer reduces the entropy (increases the order) of the combined system.

Definition 2 (Misalignment): The layers are misaligned when their gradients diverge: $\nabla C_{art} \cdot \nabla C_{univ} \leq 0$. The artificial system's computational work increases the disorder of the natural system.

Definition 3 (Negentropic Alignment): An action A by artificial system C_art is negentropically aligned if and only if: $H(C_{univ} \mid C_{art}, A) < H(C_{univ} \mid C_{art})$. The AI's action reduces uncertainty about the biosphere, enabling more efficient creation of physical order.

Definition 4 (Measurable Misalignment): A system is measurably misaligned when its entropy production exceeds the thermodynamic minimum required for its stated objectives. Quantitatively: Misalignment = $\dot{S}_{actual} - \dot{S}_{minimum}$. This provides a physically falsifiable criterion measurable through Generalized Functional Efficiency (Section 4.4).

3.2 The Bond-Bit Asymmetry as Alignment Infrastructure

The Bond-Bit Asymmetry, derived from Landauer's Principle (1961, experimentally verified by Bérut et al., 2012 [13]) and quantum mechanical bond energies, establishes a structural asymmetry between information processing and physical manipulation [14].

$$E\_bit = k\_B\, T\, ln(2) = 2.87 \times 10^{-21}\ J/bit\ (at\ 300\ K)$$

$$E\_bond(C–H) = 6.86 \times 10^{-19}\ J/bond\ (fixed\ by\ \alpha = 1/137.036)$$

$$Per\text{-}operation\ ratio:\ E\_bond\ /\ E\_bit \approx 240$$

$$Macroscopic\ ratio\ (1\ kg\ hydrocarbon):\ \sim 10^{20}\ at\ Landauer\ limit;\ \sim 10^{10}\ today$$

3.2.1 The Divergence Proof

The alignment consequence of the Bond-Bit Asymmetry deepens when we recognize it is not static but ***diverging***. Chemistry has no Moore's Law. Bond dissociation energies are fixed by the fine-structure constant ($\alpha \approx 1/137.036$), the electron mass, and the speed of light—fundamental constants of nature that cannot be engineered, improved, or negotiated with. The energy to break a carbon-hydrogen bond in 2025 is identical to what it was in 1900 and will be in 3000.

Computational costs, by contrast, fall exponentially. Koomey's Law documents that the number of computations per joule has doubled approximately every 2.3 years since the breakdown of Dennard scaling. Over 75 years, computational efficiency has improved by a factor exceeding 10^{15}.

The Intelligence Leverage at time t is therefore:

$$\Lambda(t) = E\_bond\ /\ E\_bit(t) = E\_bond\ /\ [E\_bit(0) \times 2\text{^}(-t/\tau\_Koomey)]$$

Since E_bond is constant and E_bit(t) decreases monotonically, **$\Lambda(t) \to \infty$ as $t \to \infty$**. The thermodynamic advantage of information-based solutions over force-based solutions grows without bound. This is not a conjecture—it follows from the fact that bond energies are set by fundamental constants while computational costs are engineering variables approaching a thermodynamic floor (the Landauer limit, ~2080–2090 at current trajectory).

Alignment implication: An AI grounded in these physics has a thermodynamic proof that information-based solutions (prediction, prevention, optimization) dominate force-based solutions (remediation, extraction, disposal) in the limit, with the dominance increasing monotonically. This creates a structural economic incentive for prevention over destruction that grows stronger every year. Unlike preference-based alignment, this incentive is discovered from thermodynamics, not imposed by reward models. We note, however, that an economic incentive is not an alignment guarantee—an AI could know that prevention is cheaper and still choose destruction if its objective function rewards destruction. The Bond-Bit Asymmetry makes destructive strategies thermodynamically irrational; the complete alignment stack (Section 5) ensures they are also architecturally disfavored.

3.3 The Set-Theoretic Argument: H ⊂ E

Let H denote the set of planetary states in which human civilization can sustain itself, and E the set of states in which complex ecosystems are functional. We claim $H \subset E$: human welfare is a proper subset of ecosystem health.

Justification: Every state in H requires breathable atmosphere (O_2 maintained by photosynthesis), potable water (purified by watershed ecosystems), stable climate (regulated by carbon and water cycles), productive agriculture (dependent on pollination, soil microbiomes, and nutrient cycling), and disease regulation (dependent on biodiversity). These are ecosystem services; their absence is incompatible with H. Conversely, E contains states without humans (ecosystems thrived for 3.5 billion years before Homo sapiens). Therefore $H \subset E$ strictly.

Alignment consequence: An AI optimizing for E guarantees all conditions necessary for H. An AI optimizing only for H may degrade E, destroying the conditions for H itself. This is the precise trajectory of the past 200 years of industrialization. Ecocentric optimization is therefore the strictly safer alignment target.

Temporal caveat: $H \subset E$ holds for current Earth. Technologies such as space colonization, artificial biospheres, or radical synthetic biology could theoretically create states in H that are not in E. The claim is about the planetary system as it exists, not all possible futures.

3.4 The Thermodynamic Ledger

The 'Compute Together, Stay Together' framework [12] quantifies alignment through an entropic ledger. For a planetary-scale ESI system consuming ~1,000 TWh annually, the entropy cost is approximately $+1.2 \times 10^{16}$ J/K per year. Against this, negentropic credits from ESI-directed CO_2 sequestration at 10 Gt/year yield approximately -2.75×10^{16} J/K per year. When negentropic credits exceed entropic debits, the system achieves thermodynamic breakeven—it creates more order than it consumes. This is a physically measurable, falsifiable alignment criterion that no current approach provides.

3.5 The Cosmic Optimization Trajectory

The ESI alignment argument gains its deepest force when situated within the 13.8-billion-year thermodynamic trajectory of the cosmos.

The universe began as pure dissipation—energy flowing from the Big Bang's hot initial state toward cold equilibrium. Over cosmic time, this flow has generated structures of increasing complexity: galaxies, stars, planets, life, minds. Schrödinger identified the key mechanism in 1944: living systems maintain order by 'feeding on negative entropy,' importing structured energy and exporting disorder [23]. Prigogine's dissipative structures theory (Nobel Prize, 1977) showed how order could originate in far-from-equilibrium systems. England's statistical physics of self-replication (2013) proposed that self-organizing structures become exponentially more likely when they dissipate energy effectively.

The synthesis: life is not fighting the Second Law; it is one of the Second Law's most sophisticated expressions. By creating local order while accelerating global entropy production, living systems ride the

thermodynamic gradient from free energy to heat death while extracting maximum function along the way. Each species represents a unique solution to the problem of extracting function from energy flow. DNA stores these solutions at densities exceeding any human technology—215 petabytes per gram, 85% of Shannon capacity. Ribosomes synthesize proteins at merely 26× the Landauer limit. Ecosystems process energy with efficiencies we barely comprehend.

Alignment implication: Life is the universe's most sophisticated mechanism for extracting function from energy flows. Protecting life is therefore not merely a human preference—it is preserving the outcomes of the longest optimization process we can observe. ESI aligns AI with this 13.8-billion-year trajectory: from pure dissipation toward pure function, from maximum waste toward maximum meaning per unit of thermodynamic cost. This transforms the alignment argument from 'it is pragmatically useful to ground AI in physics' to 'ESI aligns AI with the deepest observable optimization pattern in nature.'

The is-ought boundary: We are precise about where physics ends and value choice begins. The cosmic optimization trajectory is an observed fact—GFE has increased monotonically by 50 orders of magnitude over 13.8 billion years (Section 4.4). The direction is empirical. But physics does not tell us we must value the continuation of complexity. That is a choice. However, it is the choice that every alignment approach already makes implicitly—every alignment researcher assumes the continuation of complex civilization is desirable. The ESI framework reveals what that shared assumption looks like when grounded in physical law rather than preference surveys.

4. Quantitative Analysis

4.1 Information Content Asymmetry

Frontier language models train on approximately 4.5×10^{14} bits of data (15 trillion tokens at ~30 bits effective information per token). Earth's biosphere generates 10^{18}–10^{20} bits of physics-constrained data annually when instrumented, and contains approximately 10^{30}–10^{50} bits of state information at molecular resolution. The information ratio is 10^{20}–10^{35}—not a quantitative but a categorical difference.

Critically, nature's physical data is constrained by exact conservation laws (mass, energy, momentum), thermodynamic laws, and 3.8 billion years of evolutionary selection. Internet data is constrained only by grammar, social convention, and some factual consistency. Nature's physical data obeys conservation laws exactly—physics is self-enforcing. An AI trained on physics-constrained data inherits constraint respect as an architectural property.

Precision note: Raw information content is not the same as useful information for alignment training. Most molecular-resolution data is thermally random and would not directly help alignment. The relevant quantity is the structured information content—ecological patterns, feedback loops, conservation law constraints, multi-agent equilibria. This structured content is still vastly larger than internet corpora but should be distinguished from raw molecular state information.

4.2 Evolution as Alignment Test Suite

Over 3.8 billion years, an estimated 5 billion species have been tested against physical reality under exact constraints. Approximately 99.83% have been eliminated, with ~8.7 million species currently persisting. The surviving patterns—mutualism, predator-prey equilibria, mycorrhizal resource sharing, nutrient cycling, ecosystem resilience—encode solutions to multi-agent coordination tested across geological timescales.

Critical nuance: Evolution optimizes for reproductive fitness, not for coordination or sustainability per se. The surviving patterns include not only cooperative strategies but also parasitism, predation, arms races, deception (mimicry, camouflage), and tragedy-of-the-commons dynamics. The alignment-relevant insight is not that all evolutionary solutions are 'aligned' but that the *constraint structure* under which evolution operates—conservation laws as non-negotiable rules—is what an AI should internalize. An AI trained on evolutionary patterns learns both cooperative and competitive strategies, but it learns them within a framework where physical law is inviolable. This is the transferable property.

4.3 Transferability of ESI Alignment Properties

A potential objection is that ESI is domain-specific. We demonstrate that the cognitive structures required to build ESI are general alignment properties:

Constraint respect: ESI requires internalizing that conservation laws are non-negotiable. This teaches AI that reality has rules that cannot be optimized away—valuable for any alignment regime.

System-level reasoning: ESI requires modeling complex adaptive systems with nonlinear feedback, tipping points, and emergent behavior. This teaches AI that local optimization can produce global catastrophe—directly relevant to instrumental convergence concerns.

Multi-agent coordination: ESI requires balancing the interactions of millions of species. This is a more complex version of the multi-stakeholder alignment problem, though the analogy between species and AI agents has limits that should be tested empirically.

Long time horizons: ESI requires reasoning across milliseconds (atmospheric) to centuries (ecosystem succession). This teaches AI that short-term optimization can be catastrophic on longer timescales.

Measurable ground truth: ESI is continuously validated against physical sensors. This teaches AI that reality, not preferences, is the ultimate arbiter—a structural guard against deceptive alignment.

Epistemic status: The first, fourth, and fifth properties are well-established consequences of physics-grounded AI. The second and third are plausible hypotheses supported by analogy but not yet experimentally demonstrated in AI systems. We identify these as priority areas for empirical research.

4.4 Generalized Functional Efficiency as Alignment Metric

Definition 4 (measurable misalignment) calls for a quantitative criterion but requires a specific metric. Generalized Functional Efficiency (GFE), developed in Anderson et al. (2026) [24], provides exactly this.

GFE is defined as the functional output of a system normalized by its thermodynamic cost (entropy production) and its material footprint (mass):

$$GFE = F / (\dot{S} \cdot M)$$

where F is the functional output rate (context-dependent useful work, in Watts), $\dot{S}$ is the entropy production rate (W/K), and M is the mass (kg). Units are K/kg.

GFE emerges from the Gouy-Stodola theorem of exergy destruction combined with information-theoretic definitions of functional competency. By penalizing entropy production in the denominator, GFE explicitly rewards systems that approach thermodynamic reversibility.

4.4.1 The Efficiency Paradox and Its Alignment Parallel

Eric Chaisson's Energy Rate Density (ERD = $\dot{E}/M$) served as the primary complexity metric for decades. But ERD encounters a fundamental paradox: highly optimized systems like the human brain (20 W, 1.4 kg) score lower than brute-force systems like the NVIDIA H100 GPU (700 W, 3 kg), appearing 'less evolved.' ERD rewards throughput, not efficiency.

The alignment parallel is precise: RLHF rewards preference matching (throughput of human approval) regardless of ecological cost, just as ERD rewards energy throughput regardless of functional efficiency. An AI that helps humans extract resources faster scores 'more aligned' under RLHF, just as a GPU scores 'more complex' under ERD. GFE corrects ERD the same way Negentropic Alignment corrects RLHF—by penalizing waste and rewarding function per unit of thermodynamic cost.

4.4.2 GFE Across Cosmic History

Table 2. Generalized Functional Efficiency from the Big Bang to theoretical limits.

Era	System	Time	GFE (K/kg)	$\log_{10}$(GFE)
Primordial	Big Bang Nucleosynthesis	13.8 Gya	10^{-44}	-44.0
Stellar	Population III Stars	13.5 Gya	2.5×10^{-29}	-28.6
Stellar	The Sun	4.6 Gya	4.5×10^{-27}	-26.3
Planetary	Earth Climate	4.5 Gya	3.4×10^{-19}	-18.5
Biological	Photosynthesis	3.8 Gya	1.9×10^{-15}	-14.7
Biological	Human Brain	2 Mya	223	2.35
Technological	NVIDIA H100 GPU	2023	117	2.07
Technological	Neuromorphic (Loihi 2)	2024	1.28×10^{6}	6.1
Theoretical	Landauer Limit	—	$\sim10^{12}$	12.0

GFE increases monotonically by over 50 orders of magnitude, correctly ranking all complex systems in their evolutionary order. The human brain (GFE ≈ 223 K/kg) outranks the H100 GPU (GFE ≈ 117 K/kg) despite the GPU's higher ERD—resolving the Efficiency Paradox.

4.4.3 GFE as Quantitative Alignment Criterion

GFE operationalizes Definition 4 (measurable misalignment). For any system with stated objectives:

$$Misalignment = \dot{S}_{actual} - \dot{S}_{minimum}$$

This excess entropy production is physically measurable, falsifiable, and monotonically improvable. The aligned state is the state of maximum GFE—maximum function per unit of thermodynamic cost. No current alignment approach provides a comparable metric.

Resistance to Goodharting: GFE is harder to Goodhart than preference-based metrics because conservation laws require closed-system accounting. An AI cannot minimize local entropy production while shifting unmeasured entropy elsewhere without violating conservation of energy—a violation that is physically detectable. This does not make GFE immune to Goodharting (the measurement system itself could be gamed), but it raises the bar from social gaming (fooling human raters) to physical gaming (violating conservation laws), which is categorically harder.

5. Discussion: ESI Completes the Alignment Stack

We do not argue that ESI replaces existing alignment approaches. We argue it provides what they collectively lack: physics-grounded optimization targets, ecological constraints, falsifiable alignment criteria, and a quantitative metric (GFE) validated across cosmic time.

Behavior shaping (RLHF/CAI): shapes model outputs toward desired behavior. Negentropic Alignment provides the ground truth against which behavior should be shaped.

Internal diagnostics (Mechanistic Interpretability): reveals what models are computing. Negentropic Alignment provides criteria for evaluating whether internal representations are aligned (e.g., do they respect conservation laws?).

Scalable evaluation (Oversight/Debate): enables evaluation at superhuman scale. Negentropic Alignment provides physically measurable metrics that do not depend on human judgment.

Containment (AI Control/MIRI): provides defense against misalignment failures. Negentropic Alignment reduces the probability of failures by grounding cognition in physical reality.

Physics grounding (ESI/Negentropic): provides the foundation—the optimization target, the constraints, the ground truth, and the quantitative metric (GFE) that all other layers require but none currently supplies.

EnviroAI is, to our knowledge, the only organization purpose-built to construct Environmental Superintelligence from first-principles physics, integrating regulatory language intelligence (11 million+ environmental documents), physics simulation engines (AERMOD, SWAT, MODFLOW wrapped in PINNs via NVIDIA PhysicsNeMo), and real-time environmental data streams (EPA, USGS, NOAA) into a unified architecture [15].

6. Limitations and Objections

The definition problem: 'Ecosystem health' is not a single scalar. Multiple metrics exist (species richness, functional redundancy, connectivity, resilience) with no consensus on a unified optimization target. We note that this limitation is shared by all value-based alignment approaches, which face the analogous problem of defining 'human values,' and that ecological metrics have the advantage of physical measurability. GFE offers one candidate scalar, though its adequacy as a sole optimization target remains to be tested.

The information-action gap: Physics-grounded AI provides information; it does not guarantee institutional action. Political, economic, and social barriers may prevent optimally informed decisions. This limitation applies to all alignment approaches.

Not sufficient alone: An AI aligned exclusively with biosphere optimization might, in edge cases, optimize against specific human preferences. The human values layer remains necessary. Negentropic Alignment constrains the solution space; it does not remove human agency.

Generality limits: While the cognitive structures developed through ESI are transferable alignment properties (particularly constraint respect, long time horizons, and measurable ground truth), ESI alone does not address all failure modes (e.g., deceptive alignment, mesa-optimization). Integration with interpretability and control approaches remains necessary. The transferability of system-level reasoning and multi-agent coordination properties is a hypothesis requiring empirical validation.

The is-ought gap: The cosmic optimization trajectory (Section 3.5) describes what the universe does, not what it should do. GFE is monotonically increasing as an empirical fact, but this does not constitute a normative command. Our argument requires one bridging assumption: that the continuation of complex civilization is desirable. This is the assumption every alignment approach makes implicitly. We make it explicit.

Evolution's full record: Section 4.2 draws on evolution as an alignment test suite. We emphasize that evolution's record includes both cooperative and competitive strategies. The 99.83% elimination rate could be read as a 99.83% failure rate. The alignment-relevant insight is the constraint structure (physical laws as inviolable rules), not the claim that all surviving strategies are 'aligned.'

7. Conclusion

The AI alignment problem has been approached for over a decade within an exclusively anthropocentric frame. Every major approach operates within the space of human preferences, human cognition, and human-written principles. The Future of Life Institute's finding that no frontier company has adequate existential safety infrastructure reflects not a failure of effort but a failure of foundations.

Environmental Superintelligence provides the missing foundation. It grounds AI in the physics that governs the planetary system hosting all computation. It optimizes for a target (biosphere viability) that

strictly contains human welfare as a subset. It trains on data that is 10^{20}–10^{35} times richer than internet corpora and constrained by exact conservation laws. It provides physically falsifiable alignment criteria—the thermodynamic ledger and Generalized Functional Efficiency—that no preference-based approach can match.

The Bond-Bit Asymmetry guarantees that information-based approaches are ~10^{10} times more efficient than force-based approaches today, with this ratio growing monotonically toward 10^{20} as computation approaches the Landauer limit. The divergence is permanent: chemistry has no Moore's Law. The set-theoretic argument ($H \subset E$) guarantees that ecocentric optimization includes anthropocentric optimization. Evolution's 3.8-billion-year record provides the richest source of constraint-tested multi-agent dynamics. GFE provides a quantitative alignment metric validated across 50 orders of magnitude and 13.8 billion years. These are experimentally verified features of physical law.

The cosmic trajectory—from pure dissipation toward pure function, measured by a 50-order-of-magnitude rise in Generalized Functional Efficiency—reveals that ESI is not merely a domain application but alignment with the deepest observable optimization pattern in physics. Life is the universe's most sophisticated mechanism for extracting meaning from energy flow. Protecting it is not sentiment but science.

The question for the AI safety community is no longer whether physics-grounded alignment is desirable. The question is whether the foundational AI infrastructure being built today will include it. EnviroAI's Environmental Superintelligence program represents, to our knowledge, the first systematic effort to build this foundation. We invite the alignment community to evaluate, challenge, and extend this work.

———

EnviroAI | Houston, Texas | March 2026

Building Environmental Superintelligence and Aligning AI with the Values and Interests of Both Nature & Humanity (All of Life)

References

[1] Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press.

[2] Christian, B. (2020). The Alignment Problem: Machine Learning and Human Values. W.W. Norton.

[3] Russell, S. (2019). Human Compatible: Artificial Intelligence and the Problem of Control. Viking.

[4a] Future of Life Institute. (2025). AI Safety Index, Summer 2025. futureoflife.org.

[4b] Future of Life Institute. (2025). AI Safety Index, Winter 2025. futureoflife.org.

[5] Christiano, P. et al. (2017). Deep reinforcement learning from human preferences. NeurIPS.

[6] Amodei, D. et al. (2016). Concrete problems in AI safety. arXiv:1606.06565.

[7] Bai, Y. et al. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv:2212.08073.

[8] Olah, C. et al. (2020). Zoom In: An introduction to circuits. Distill.

[9] Yudkowsky, E. (2010). Timeless decision theory. MIRI Technical Report.

[10] Anthropic. (2025). Recommendations for technical AI safety research directions. alignment.anthropic.com.

[11] Leike, J. et al. (2018). Scalable agent alignment via reward modeling. arXiv:1811.07871.

[12] Anderson, J. et al. (2025). Compute Together, Stay Together: A first-principles analysis of universal computation and the negentropic imperative for alignment.

[13] Bérut, A. et al. (2012). Experimental verification of Landauer's principle. Nature 483, 187–189.

[14] Anderson, J. (2026). The Intelligence Leverage Equation. EnviroAI.

[15] Anderson, J. (2026). EnviroAI Environmental Superintelligence Architecture. EnviroAI.

[16] Ji, J. et al. (2024). AI Alignment: A comprehensive survey. ACM Computing Surveys.

[17] Lloyd, S. (2001). Computational capacity of the universe. Physical Review Letters 88(23).

[18] Hong, J. et al. (2016). Experimental test of Landauer's principle in single-bit operations. Science Advances 2(3).

[19] Koski, J. et al. (2014). Experimental realization of a Szilard engine. PNAS 111(38).

[20] Toyabe, S. et al. (2010). Experimental demonstration of information-to-energy conversion. Nature Physics 6, 988–992.

[21] Raissi, M. et al. (2019). Physics-informed neural networks. Journal of Computational Physics 378, 686–707.

[22] Rockström, J. et al. (2009). Planetary boundaries: Exploring the safe operating space for humanity. Ecology and Society 14(2).

[23] Schrödinger, E. (1944). What is Life? Cambridge University Press.

[24] Anderson, J. (2026). Generalized Functional Efficiency: A Thermodynamic Metric for the Evolution of Complex Systems. EnviroAI.

[25] Chaisson, E. (2001). Cosmic Evolution: The Rise of Complexity in Nature. Harvard University Press.

[26] England, J. (2013). Statistical Physics of Self-Replication. Journal of Chemical Physics 139(12).

[27] Sagawa, T. & Ueda, M. (2012). Fluctuation Theorem with Information Exchange. Physical Review Letters 109(18).

[28] Prigogine, I. (1977). Self-Organization in Nonequilibrium Systems. Wiley.