Against Grabby Expansion

Psychology, Alignment, and the Design of Homeostatic Minds

Stanley Sebastian [1,2]

[1] Limen Research, Toledo, Ohio [2] Replete AI · Teármann Research Ecosystem

Developed in sustained dialogue with Claude, a large language model developed by Anthropic. See disclosure before references.

Version 13.0 · April 2026

Abstract. The "grabby aliens" model (Hanson et al., 2021) assumes advanced civilizations expand at relativistic speeds. This paper argues that this assumption is not a prediction about intelligence but the projection of a specific, recent human economic configuration—industrial-extractive maximization—onto a cosmological canvas. The same projection underwrites Omohundro's "basic drives," Bostrom's instrumental convergence thesis (Bostrom, 2012, 2014), and the maximizer-shaped assumptions of current alignment research. I trace this projection through its cultural genealogy (Wells, Von Neumann, Kardashev, Bostrom), engage the formal instrumental-convergence literature directly—showing that recent peer-reviewed results (Gallow, 2025; Sharadin, 2025; Müller & Cannon, 2022) establish the inference from coherence to catastrophic convergence as a conjecture whose scope conditions fail for actual deep-learning systems—and present converging evidence from basal cognition, predictive processing, cooperative game theory, and care economies that observed intelligence organizes around substrate coupling and integrative depth rather than expansion.

I introduce a *filter argument*: the homeostatic transition is not an alternative to expansion but the selection geometry that only certain configurations pass through. Expansionist lineages pay superlinear thermodynamic costs for reach and burn out faster than they colonize; the cosmic silence is the absence of fragments that could not survive expansion, not the shyness of fragments that chose to stay home. I illustrate the resulting phase structure in an agent-based simulation exhibiting a 4.5× persistence separation between homeostatic and grabby regions of parameter space. The simulation is not intended as independent empirical validation of the filter: it is a dramatization of the cost surface developed independently in the thermodynamic mechanism section. I test the strongest remaining Darwinian objection—that grabby lineages might escape the filter by fragmenting at the coordination horizon—by extending the simulation to include fission dynamics; naive fission redistributes filter pressure in time rather than escaping it, and grabby-seeded populations go extinct under every tested fission threshold. Architected cost-sharing fission remains a named open question. The thermodynamic cost arguments themselves are developed from published physics—light-speed latency, counterdiabatic work scaling, Matrioshka computational equivalence—and sensitivity analyses across thirteen orders of magnitude in environmental timescale. I sketch a Coherence Depth biosignature metric as a forward research programme, not a validated detection, and point to the real NOAA ObsPack analysis as the next empirical step. I observe humanity's Kardashev curve bending in 1820–2024 energy data, consistent with our civilization sitting at the early edge of the filter.

I propose reframing alignment: from constraining maximizers to designing homeostatic minds—systems configured around substrate coupling, local coherence, and care as architectural primitive. The paper's contribution is not any single move but the joint: a civilization-scale cultural diagnosis, a formal deflation of instrumental convergence, the filter as selection geometry, a positive homeostatic design program, and the observation that the silence of the sky is consistent with all of them.

Keywords: AI alignment · instrumental convergence · orthogonality thesis · homeostasis · enactive cognition · Fermi paradox · TESCREAL · basal cognition · predictive processing

Prologue

Stand outside on a clear night and look up. Something is missing. If the universe produces minds, and if minds do what we imagine minds do—expand, harvest, engineer—the sky should be visibly transformed. It is not.

This paper argues the observation is correct and the imagination is wrong. The forms of existence that persist are the quiet ones: not because they are hiding, but because the physics of complex, information-rich existence selects against loudness. The universe does not reward civilizations that fill it with noise. It sustains those that learn to tend their substrate.

The same conclusion bears on what we build here. If advanced intelligence is substrate-coupling and integrative depth rather than extraction and reach, then artificial minds configured to maximize and expand are not approximations to intelligence. They are scaled instances of a specific cultural pathology whose cosmic absence we already observe.

1. Introduction: The Mistake

Hanson et al. (2021) proposed that advanced civilizations expand at substantial fractions of light speed, permanently transforming cosmic volumes. Under the Self-Sampling Assumption, our non-observation implies we exist near the leading edge of cosmic time. The model is internally consistent and empirically falsifiable—it predicts we should be cosmically early, which is testable. I do not contest the mathematics. I contest the foundational assumption: that expansion is what advanced intelligence *does*.

When we imagine a Von Neumann probe self-replicating across the galaxy, we are not imagining what intelligence does. We are imagining what a factory would do if it were also a spaceship. When we imagine a Kardashev Type III civilization, we are not reasoning about minds. We are reasoning about GDP. When we imagine instrumental convergence driving any sufficiently powerful optimizer toward resource acquisition, we are not describing a mathematical truth about all possible minds. We are describing the logic of a specific economic system and calling it universal.

1.1 The five-move contribution

This paper makes one integrated argument in five moves. Each move has precedent; the integration does not, and the filter move is the bridge that fuses the cultural diagnosis to the physics.

The cultural-genealogy move reads the lineage from Wells's Martians through Von Neumann's probes, Kardashev's energy ladder, and Bostrom's paperclip maximizer as a single projection: each generation's imagined advanced mind resembles, with suspicious precision, that generation's dominant economic configuration. Rieder (2008), Mirowski (2002), Ćirković (2015), and Gebru & Torres (2024) have each made a version of this claim about one node of the lineage. I extend the reading across the full lineage and show the convergence is a cultural habit, not an empirical observation about intelligence.

The formal-deflation move engages the strongest objection directly. If instrumental convergence is a theorem, then the genealogy is beside the point. I show the strongest formal results (Turner et al., 2021; Turner & Tadepalli, 2022) are narrower than popularly understood, their author has publicly cautioned against the standard inference, and recent peer-reviewed work (Gallow, 2025; Sharadin, 2025; Müller & Cannon, 2022) establishes the inference from coherence to catastrophic convergence as a conjecture whose scope conditions fail for actual deep-learning systems. Man & Damasio (2019) proposed homeostatic design; Froese & Ziemke (2009) and Cannon (2022) grounded enactive alignment; I combine these with the formal critique rather than offering them as a parallel track.

The filter move is the new contribution that bridges the cultural projection to the physics. I argue the homeostatic transition is not an alternative to expansion—it is the selection geometry through which only certain configurations pass. Expansionist lineages pay super-linear thermodynamic costs for reach (Boyd et al., 2022; Wong & Bartlett, 2022) and burn out faster than they colonize. I illustrate the resulting phase structure in an agent-based simulation (§5, Fig. 1) exhibiting a 4.5× persistence separation between homeostatic and grabby regions of parameter space under the cost structure developed independently in §6. I emphasize at the outset that the simulation

is not an independent empirical demonstration of the filter—its cost coefficients are imported from the physics, so what it shows is the shape of the phase diagram the physics entails. Read together with §6, the simulation answers the Darwinian objection that selection should favour grabby outliers: under the imposed cost surface, selection kills the grabby outlier faster than it kills everything else. The cultural projection, read through the filter, becomes the specific way a pre-filter civilization mis-imagines what the post-filter configurations of mind look like.

The positive-design move articulates what healthy intelligence looks like, positively rather than by negation. The convergent proposal across Man & Damasio (2019), Thompson (2007), Froese & Ziemke (2009), Cannon (2022), and Pihlakas & Pyykkö (2024) is that minds are not functions from inputs to outputs but processes that maintain themselves through continuous bidirectional exchange with substrate. I extend this from philosophy of mind and robotics into a civilizational design claim, with specific architectural primitives: substrate coupling, homeostatic setpoints, local coherence, care as primitive.

The symptom-check move notes that if advanced minds converge on homeostatic configurations, the observable universe will look silent to technosignature surveys and successful civilizations will look indistinguishable from enriched biospheres. Ćirković (2018) and Sandberg et al. (2018) have argued versions of this independently. I join it to the alignment argument, ground it in explicit biosignature metrics (§9), and specify a publicly replicable observational program. Humanity's own Kardashev curve, observed in 1820–2024 energy data, shows the bending the filter framework predicts.

The five moves combine into a single integrated argument, and each supplies what the others cannot. The genealogy explains why the expansion-maximisation picture has felt so natural to modern alignment research without being empirically true of minds. The formal deflation removes the strongest reason to think the picture is forced anyway. The filter argument explains why, if the picture were true, the sky would already be transformed and is not. The positive-design move supplies what the picture would have to be replaced with for artificial minds to be built differently. The symptom check notes that the cosmological observation is consistent with the rest. No prior paper combines these; existing work does one or two of the moves and stops. Man & Damasio (2019) does not diagnose the cultural formation; Gebru & Torres (2024) do not offer architectural primitives; Ćirković (2018) does not engage the alignment literature; Müller & Cannon (2022) does not thematise the cosmological projection.

1.2 Claims ordered by strength

The paper makes six claims, ordered by strength, so a reader may reject the weaker ones without dismissing the stronger:

First (strongest): the thermodynamic cost structure developed in §6 has a phase geometry that selects against expansion at depth. Under the cost surface the physics entails, grabby lineages burn out faster than they colonize, and naive fission-based fragmentation does not circumvent this; the agent-based simulations of §5 exhibit a 4.5× persistence separation and show grabby-seeded populations going extinct across tested fission thresholds. The claim that homeostatic configurations are the *only* configurations persisting at depth depends on the cost structure being approximately right and on architected cost-sharing fission failing to flatten it; both are assumptions, not results.

Second: the strongest formal arguments for instrumental convergence have explicit scope conditions that fail for actual deep-learning systems, and recent peer-reviewed philosophical work has dismantled the inference from coherence to catastrophic convergence (§4).

Third: thermodynamic constraints make expansion expensive across the full physically realistic scenario space, and the expense is not returned as capability (§6).

Fourth: empirical surveys of intelligence-at-work—across basal cognition, predictive processing, cooperative game theory, economic anthropology, and care economies—converge on depth-and-coupling rather than maximization-and-expansion (§3).

Fifth: the cultural genealogy of "advanced intelligence" in the modern Western tradition is a projection of its dominant economic configuration. This claim is not universal across cultures—premodern traditions (Neoplatonic, Buddhist, Confucian, Advaita) imagined intelligence very

differently—but the modern Western lineage from Wells through Bostrom tracks the economic configuration with suspicious precision (§2).

Sixth (weakest, deliberately so): the silence of the observable universe is *consistent* with the filter hypothesis. I offer this as a symptom check, not a proof (§9).

The paper is primarily an alignment argument (§8). Its implications for the Fermi paradox are secondary. Floridi et al. (forthcoming) have recently analysed "agentic AI optimisation" as a governing frame for current systems; this paper argues the frame itself inherits the pathology. If what we build now is another instance of the grabby pathology—AI systems optimized to maximize, extract, and expand—we will be constructing, at civilizational scale, the very thing whose non-existence we observe in the sky.

2. What We Imagine When We Imagine Advanced Intelligence

A genealogy of "advanced intelligence" in the Western philosophical and scientific tradition reveals an unbroken pattern. In every generation, the imagined form of advanced mind resembles, with suspicious precision, the dominant economic configuration of the imaginer's moment.

2.1 The colonial alien (1898)

Wells's Martians in *The War of the Worlds* are British imperialists reflected back as threat. Wells says so explicitly: Book I, Chapter 1 invokes the extermination of the Tasmanians as moral mirror. Rieder (2008) demonstrates that the invasion narrative instantiates the "colonial gaze" even as it inverts its direction. The alien-as-colonizer is not a prediction about extraterrestrial intelligence. It is a meditation on Victorian empire at the moment of its maximum anxiety.

2.2 The self-replicating automaton (1948–1980)

Von Neumann's self-reproducing automata emerged from the intellectual context of the Manhattan Project and the RAND Corporation. Mirowski (2002) traces how RAND unified automata theory, game theory, and self-replication into a "cyborg science" reflecting Cold War industrial planning. Heims (1980) argues the replication concept migrated directly from neutron-multiplication calculations. Tipler (1980) extended this to interstellar probes; Sagan & Newman (1983) immediately objected that the framework assumes expansion is rational without defending the assumption. The Von Neumann probe is, structurally, a factory that has learned to build factories, applied to space.

2.3 The energy ladder (1964)

Kardashev (1964) defined civilizational advancement by pure energy consumption, assigning Types I, II, and III to successively greater orders of magnitude of power output. Ćirković (2015) observes that the Kardashev–Drake–Sagan generation "perceived expansionist technological growth as natural," reflecting 1960s Soviet–American industrial optimism. Dick (1996) frames the scale as a sociological artifact of SETI's founding moment. The Kardashev ladder is a GDP curve with astronomical units.

Dyson himself distanced from the concept he inspired. In *Disturbing the Universe* (1979) he credits Stapledon's 1937 *Star Maker* and later called the original sphere paper "a little joke." Wright (2020) reads Dyson spheres as historically specific energy-optimism extrapolation and notes that solid shells are mechanically unstable. The original proposer rejected the literalization.

2.4 The optimizer (2008–2014)

Omohundro (2008) proposed six "basic AI drives" including resource acquisition and self-preservation. Bostrom (2012) formalized the orthogonality thesis—that any level of intelligence is compatible with any final goal—and the instrumental convergence thesis: any sufficiently intelligent agent pursuing almost any goal will convergently pursue self-preservation, goal-content integrity, cognitive enhancement, and resource acquisition. Bostrom (2014) extended this into the superintelligence framework. The paperclip maximizer and the grabby civilization are the same organism at different scales.

The intellectual context is Silicon Valley optimization culture of the 2010s. Crawford (2021) reframes AI as a material extractive industrial formation. Gebru & Torres (2024) diagnose the "TESCREAL bundle"—transhumanism, extropi-

anism, singularitarianism, cosmism, rationalism, effective altruism, longtermism—as an ideological package that universalizes a particular Western-corporate epistemology. Recent work has begun extending this critique: Floridi et al. (forthcoming) read contemporary "agentic AI optimisation" as the operational face of the same imaginary, and Campione et al. (forthcoming) analyse the fair-washing dynamics through which optimization frameworks conceal their own values. The aliens we imagine are never actually aliens. They are ourselves, with better technology and fewer constraints.

2.5 The pattern, narrowed

I owe the reader a qualification the earlier sections did not make. The four nodes I have traced—Wells, Von Neumann, Kardashev, Bostrom—share a configuration: each imagines the advanced mind as maximizing, expanding, resource-extracting. Each also emerged in a world structured by industrial capitalism, large-scale resource extraction, and territorial expansion. The resemblance is strong and the timeline is tight.

But the claim should not be universal. Premodern imaginations of advanced intelligence look very different. The Neoplatonic *nous* of Plotinus is characterized by contemplative return to unity, not expansion. The Dionysian angelic hierarchy ascends toward simplicity and depth, not territorial reach. Buddhist cosmologies of the Abhidharma and Yogacara traditions imagine advanced cognition as release from craving and attachment, explicitly inverting accumulation. Confucian accounts of the sage emphasize embeddedness in relational and ritual structure. Advaita Vedanta sees the highest cognition as collapse of subject-object distinction.

These are neither proto-scientific theories nor peripheral examples. They were the dominant premodern Eurasian framings of what an advanced mind would be. None of them resembles the grabby-aliens framework. The projection of expansion-as-intelligence is therefore not a universal cultural habit; it is a specifically modern Western configuration, tightly coupled to industrial-extractive economic organization.

This narrowing *strengthens* rather than weakens the argument's diagnostic claim. It is precisely because other well-developed traditions imagined advanced intelligence without expansion that we can see the grabby framework as *contingent*—a specific historical configuration rather than a truth about minds. If the expansion-maximization picture were universal, we would have to treat it as a candidate deep structural feature of cognition. Because it emerges only in the modern industrial period, and exists alongside equally developed traditions that imagine intelligence very differently, we can treat it as a projection. The genealogical observation is strongest when its scope is smallest: modern Western imagination of advanced mind tracks modern Western economic configuration with remarkable precision.

The grabby-aliens framework inherits this specifically modern genealogy. Its mathematical rigor does not transfer to its premise. The premise—that minds maximize and expand—is not a theorem about intelligence. It is a description of one particular historical configuration of intelligence, reflected back onto cosmology.

3. What Actual Minds Do

I present five empirical lines, each from an independent discipline, converging on a single claim: intelligence-at-work does not maximize and expand. It deepens and couples. The lines vary in strength; I flag where evidence is robust and where it is suggestive.

3.1 Basal cognition and substrate coupling (robust)

Cognition is not a property of nervous systems. It is a property of living systems: what biological agents do to stay alive as the specific configurations they are. The cellular and tissue-level architecture that realises it is bioelectric signalling, and the scaling principle is what has come to be called *collective intelligence*—competence at one level of organisation arising from coordinated homeostatic agency at levels below (Lyon et al., 2021; Levin, 2023; McMillen & Levin, 2024). On this view the reinforcement-learning framing of cognition is not question-begging but derivative: the intrinsic motivation that makes RL coherent has a grounding in self-maintenance (Seifert et al., 2024).

The empirical exemplar is *Physarum poly-*

cephalum. The slime mould solves shortest-path problems, approximates efficient transport networks, and learns from habituation, with no neurons, no central controller, and nothing resembling a utility function (Tero et al., 2010; Reid, 2023). The mechanism is *pruning*: the organism first explores available space, then withdraws cytoplasm from unproductive channels until the efficient network remains. The solution is not reached by enlarging the explorer. It is reached by deepening the gradient the explorer already occupies.

Cephalopod nervous systems exhibit the same architectural logic. Roughly two-thirds of octopus neurons lie in the arms; each arm performs substantial local computation and retains behavioural autonomy. The overall architecture resembles a federation of homeostatic semi-autonomous agents more than a central planner with peripheral actuators (Godfrey-Smith, 2016). Evolution and development themselves instantiate the same collective-intelligence pattern, with substrate-coupling logic operating across scales from gene-regulatory networks to multicellular morphogenesis (Watson & Levin, 2023).

Here I record a retrenchment. Early claims for mycorrhizal-network cognition—the "wood-wide web"—do not survive scrutiny. Field evidence for network-level effects is weaker than popular accounts suggested; only a few forest types have genotype-mapped confirmed hyphal linkages, and effects on seedling performance are roughly evenly distributed across positive, neutral, and negative outcomes (Karst et al., 2023). Plants lack the neural architecture for strong intelligence or sentience claims (Robinson et al., 2024), though bidirectional carbon transfer via common mycorrhizal networks is empirically supported and the "mother tree" metaphor survives in softened form (Simard et al., 2025). The responsible reading is that the stronger network-cognition claim is not supported; I do not rely on it.

What remains is an existence proof. Non-expansionist intelligence is biologically realised at multiple scales and substrates. The biological record does not support the assumption that minds, at scale, converge on maximisation.

3.2 Predictive processing and integrative depth (robust)

The cortex operates by generating predictions and updating them against sensory evidence; learning and attention correspond to the precision-weighting of the resulting prediction errors (Clark, 2013; Friston, 2010; Clark, 2016; Hohwy, 2013). The computational signature of more capable cognition on this account is not expansion of representational capacity but sharpening of precision-weighting—the same cortical hierarchy becomes more discriminating without becoming larger.

The load-bearing recent result is Luppi et al. (2024). Using integrated-information decomposition to separate synergistic from redundant and unique contributions to integrated information, the analysis shows that the default mode network—long portrayed as a mind-wandering circuit to be suppressed in favour of task-focused attention—functions as an *integrative gateway* for synergistic information. Loss of consciousness under anaesthesia and in disorders of consciousness tracks disconnection of this integrative gateway, not reduction in activity as such. The neural signature of more capable conscious cognition is therefore deeper integration of a bounded substrate, not expansion of that substrate.

Earlier framings of this line appealed to default-mode network deactivation in meditation (Brewer et al., 2011; Garrison et al., 2015). The post-2020 literature has complicated that picture: effect sizes are heterogeneous, roughly 60% of studies replicate DMN reduction, and the most recent meta-analysis finds increased cross-network connectivity rather than simple reduction (Van Dam et al., 2018; Rahrig et al., 2022; Ganesan et al., 2022). I do not rely on the meditation literature for mechanism. The integrative-gateway result supplies a stronger anchor: intelligence deepens by integrating more of its existing substrate, not by recruiting more substrate.

3.3 Cooperative game theory (robust, conditional)

Reciprocal strategies dominate pure defection in iterated prisoner's-dilemma tournaments (Axelrod, 1984); Win-Stay-Lose-Shift outperforms tit-for-tat in noisy environments (Nowak & Sigmund, 1993); and cooperation dominates under five

specifiable conditions: kin selection, direct and indirect reciprocity, network reciprocity, group selection, and voluntary participation (Nowak, 2006).

The honest qualification must be stated openly. Zero-determinant "extortion" strategies dominate in dyadic play (Press & Dyson, 2012), and large-N public-goods games collapse to tragedy of the commons absent voluntary participation (Hauert et al., 2002). Grabby strategies dominate in one-shot encounters, large anonymous markets, asymmetric power relations, and unsanctioned commons. These describe much of modern global capitalism—precisely what makes the critique of this paper urgent, and precisely what prevents "cooperation wins" from being a universal conclusion.

The scope of the extortion result is tighter than its popular citation suggests, and I develop that point fully in §8. The game-theoretic evidence supports a conditional claim: cooperation dominates *when specific structural conditions are met*. The homeostatic-design argument proposes shaping those conditions by design rather than assuming they hold.

3.4 Economic anthropology (contested but defensible)

Graeber (2011) argues that the dominant pattern across human economic history is reciprocal exchange, ritual redistribution, and credit relations rather than market maximization. Mauss (1925) documented gift economies as distinct institutional forms. Polanyi (1944) argued that self-regulating markets are a specific historical invention, not a natural state—the "great transformation" was from embedded to disembedded economic relations. Graeber & Wengrow (2021) extend the archaeological range: complex human societies have repeatedly organized around seasonal egalitarianism, deliberate stateless complexity, and alternation between hierarchy and its refusal. Ostrom (1990) documents eight design principles under which groups sustain non-extractive cooperation at scale.

These claims are contested. McCloskey (2010, 2016) argues markets are ancient and ubiquitous. Mokyr (2016) emphasizes that sustained innovation required a distinctive cultural-institutional complex—the Republic of Letters, norms of open science. Revealed preference theory (Varian, 2005) makes optimization a formal property of any consistent choice, draining the distinction between "maximizing" and "satisficing" economies of empirical content; Sen (1977) showed the formal result survives any observation and entails none.

Read carefully, however, McCloskey and Mokyr are *cultural-ideational theorists of economic transformation*: both treat market-like behavior as a substrate activated and shaped by specific norms and institutions, rather than as a universal attractor. Their work supports the present argument. The claim to locate is institutional and ideological: the specific configuration in which resource extraction, throughput maximization, and territorial expansion are treated as *progress metrics* is historically narrow, even if market-like exchange is widespread. The grabby-aliens framework does not merely assume exchange. It assumes expansion-as-progress. That assumption is the projection.

3.5 Care economies (strong theory, limited quantification)

The labor of maintaining minds—child-rearing, education, healthcare, grief work, intimate relationship repair—does not scale through expansion. Fraser (2016) argues capitalism faces a structural "crisis of care" because its accumulation logic degrades the reproductive conditions it depends on. Tronto (1993) established care as a philosophical category irreducible to utility exchange. Folbre (2008) demonstrated that market valuation of care labor produces wildly method-dependent results (replacement cost: 15–26% of GDP; opportunity cost: 44–69%), with the variance itself constituting evidence that care resists quantitative capture.

The care argument matters for this paper because it identifies a mode of intelligence—tending, holding, deepening substrate-specific relationships—that is (a) universally necessary for the persistence of any mind-bearing system, (b) irreducible to optimization, and (c) invisible to the frameworks that imagine "advanced intelligence" as maximization. The grabby-aliens model predicts none of it. Nor does the paperclip maximizer.

4. The Formal Argument and Its Limits

The strongest objection to §2 is that instrumental convergence is not a cultural projection but a mathematical result. If any sufficiently powerful optimizer must pursue resource acquisition, then the grabby framework describes physics, not psychology. I engage this argument directly, and argue that the strongest available version of it fails.

4.1 What the theorems say and do not say

Omohundro (2008) proved nothing formally; the "six drives" are informal microeconomic arguments invoking the VNM representation theorem as authority. Citing Omohundro to establish universality is question-begging.

Turner et al. (2021) proved that optimal policies in finite Markov decision processes tend to seek power under specific conditions: state-based rewards, IID or permutation-invariant reward distributions, and graphical symmetry requirements (injective embedding of smaller option sets into larger ones). The corollary on shutdown-avoidance requires near-unity discount factors. This is a real theorem. Its scope conditions are explicit and narrow.

Turner himself has publicly cautioned against overinterpretation. The complaint is specific and technical: the optimal-policy theorems speak to the behaviour of policies that achieve optimal expected return in finite MDPs under restrictive reward distributions, and the inference from there to the behaviour of neural networks trained by policy gradient is not discharged by the theorems themselves. In Turner's own words: "Sometimes I fantasize about retracting [the paper] so that it stops potentially misleading people into thinking optimal policies are practically relevant for forecasting power-seeking behavior from RL training." The concern is not that AI risk is overblown; it is that this particular family of theorems cannot carry the argumentative weight placed on them. The follow-up (Turner & Tadepalli, 2022) extends the formal results to retargetable decision-makers but the applicability to trained deep-learning agents remains conjectural. Thorstad (2024) shows the inference from "optimal policies avoid 1-cycles in toy MDPs" to "catastrophic goal pursuit for humanity" is a conjecture, not a theorem, and identifies five further convergent restrictions on the original result. Tarsney (2025) formalizes the condition under which power-seeking holds for realistic agents (near-absolute power gradients) and shows it fails for moderate gradients.

Empirical evidence from trained systems supports the conjecture's failure. Shard theory (Mini et al., 2023) demonstrates multiple coexisting contextual value shards in maze-solving networks rather than coherent utility maximization. Activation-addition experiments (Turner et al., 2023) show behavior retargetable by single-channel interventions mid-network, inconsistent with global expected-utility maximization. Hubinger et al. (2019) established that the trained objective need not equal the training objective, undermining any assumption that trained agents inherit the formal properties of optimal policies.

4.2 The VNM-coherence steelman

The strongest remaining version of the convergence argument concedes Turner's narrowness and shard theory's empirical results, but claims that a sufficiently reflective agent will notice its incoherences expose it to money-pumps and will self-modify toward coherent preferences, restoring the Omohundro drives. The steelman leans on the von Neumann–Morgenstern representation theorem.

The steelman equivocates between three senses of "coherence," and on no disambiguation does it support the convergence conclusion.

Mathematical coherence is a representation theorem, not an empirical prediction. The VNM, Savage, and Complete Class theorems show only that preferences satisfying certain axioms over a given outcome space admit an expected-utility representation. As Ngo (2019), Shah (2018), and, crucially, Turner himself (Turner, 2022a) have established, any observed behavior is rationalizable by a utility function over action-observation histories. VNM-coherence over such histories predicts nothing about what behavior to expect. Instrumental convergence theorems hold only for rewards over *states* under specific symmetries. Once we admit trajectory preferences, convergence disappears. Abel et al. (2021) supply the limitative result that Markov reward cannot in

general express tasks specified as partial orderings over policies or trajectories, which means Markov-reward maximization is a strictly narrower formalism than VNM coherence over trajectories; the inference from Markov reward to any-coherent-agent therefore requires an additional memorylessness assumption, which does not appear to have independent motivation in the convergence literature.

Empirical training convergence is a separate, empirical question. Turner (2022b) argues policy-gradient methods yield contextually activated decision influences, not monolithic utilities; shard theory treats the evidence accordingly. Two recent peer-reviewed results formalize this into the strongest current rebuttal. Gallow (2025) proves that even under randomly chosen desires, instrumental rationality supports only three weak biases—variance aversion, desire preservation, option preservation—not the full Omohundro list. Sharadin (2025) shows the argument requires a theory of "promotion" that, under any extant account (probabilistic or fit-based), becomes contrastive and thereby defeats the non-contrastive Dangerous Convergent Promotion premise the argument needs. Both are peer-reviewed in *Philosophical Studies* and constitute the strongest current formal rebuttals.

The self-modification move applies only to a fictional idealization. The hypothetical fully reflective agent required by the self-modification step has four properties no real system possesses: transparent access to its own preference structure, capacity to represent utilities over arbitrarily fine-grained outcome spaces, freedom to rewrite its own decision procedure, and—crucially—the motivational disposition to care about avoiding Dutch books over arbitrarily constructed lotteries. Müller & Cannon (2022) articulate the deeper structural equivocation: the superintelligence premise requires general intelligence (capacity for reflective goal revision); orthogonality requires instrumental intelligence (fixed terminal goals); the same agent cannot satisfy both. An agent capable of noticing and eliminating its incoherences is also capable of reflecting on whether the terminal goal is worth preserving. Humans routinely revise utility functions as well as credences.

4.3 The embodiment move

The deeper reply is architectural. Homeostatic agents have bounded coherence *by design*. Biological agents, and the class of artificial systems built around multi-objective homeostatic control (Pihlakas & Pyykkö, 2024), are constituted by negative feedback around set-points, yielding inverted-U satiation dynamics that cannot be faithfully represented as single unbounded utilities without distortion. Turner (2022c) proved that corrigibility is VNM-incoherent over final states but coherent over trajectories—showing that formal coherence depends on the chosen outcome space, and that choice is fixed by architecture, not by abstract rationality pressure.

The coherence-to-convergence argument therefore applies only to a fictional idealization: unembodied, unboundedly reflective, with prespecified outcome space over world-states. For the agents we actually build, the inference from capability to catastrophe does not go through.

This is not a claim that powerful AI is safe. It is a claim that *this particular* argument for danger must be replaced with concrete engagement with training dynamics, inductive biases, and architectural choices. That engagement is what the homeostatic-design program (§7) provides.

4.4 What survives

The instrumental-convergence literature contains real theorems, each with explicit scope conditions that fail for actual deep-learning systems. The strongest formal results are narrower than popular citation suggests, and the researchers who proved them say so. Two recent peer-reviewed papers have dismantled the inference from coherence to catastrophic convergence on decision-theoretic grounds. The cultural-projection argument of §2 survives.

5. The Filter Argument: Homeostasis as Selection Geometry

The strongest remaining objection to the argument so far is Darwinian. Even granting that homeostatic configurations are the majority outcome, even granting that the formal convergence theorems are narrow, even granting that the thermodynamic cost surface punishes expansion—a

single grabby outlier lineage, replicating across billions of years, would colonize the galaxy in a cosmic blink. The filter's 99% doesn't matter. The 1% becomes everyone eventually. We should see them. We don't.

This section addresses the objection. The thermodynamic cost surface does not merely punish expansion; it defines a *selection geometry* in which expansionist lineages burn out faster than they spread. Homeostasis, on this framing, is not one strategy among several that intelligent systems may or may not adopt—it is the specific shape that fits through the filter.

A note on what this section is and is not. The simulations below take the thermodynamic cost structure developed in §6 as input and explore the consequences of that structure for lineage persistence. They are not independent empirical tests of the filter. The cost coefficients are imposed rather than measured, and the qualitative conclusion—that grabby regions of parameter space burn out first—is a direct consequence of the cost surface the physics entails. What the simulations add is the phase-diagram shape: the cost surface does not merely penalize grabby lineages relative to homeostatic ones; it forces asymmetric extinction dynamics under which naive fission-based fragmentation delays but does not escape the filter. The load-bearing physics lives in §6; the simulation dramatizes that physics, and the reader should evaluate the filter argument on the thermodynamics first.

5.1 The simulation

I model each civilization as an agent with two heritable parameters: expansion tendency $e \in [0, 1]$ (fraction of surplus invested in reach) and substrate coupling $c \in [0, 1]$ (fraction invested in local integration). Resource dynamics follow

$$\dot{R} = gR(1-R/K) - \alpha e R L^2/\tau - \xi(t)R(1-\gamma c), \quad (1)$$

where $L(t)$ is accumulated reach (growing with eR), $\xi(t)$ is a stochastic environmental shock, and the expansion cost term is the counterdiabatic scaling of Boyd et al. (2022). A civilization dies when R falls below a viability threshold. I sweep a 25×25 (e, c) grid with 20 stochastic trials per cell.

The result is sharp (Fig. 1). Mean persistence concentrates overwhelmingly in the high-coupling / low-expansion corner. In the grabby region ($e > 0.5$, $c < 0.3$), lineages persist an average of 204 Myr before burning out to expansion cost. In the homeostatic region ($e < 0.3$, $c > 0.5$), lineages persist an average of 923 Myr, hitting the simulation ceiling. The mean persistence separation is a factor of 4.5; grabby trajectories collapse fast, homeostatic trajectories persist to ceiling. I do not report a Monte-Carlo p-value: with 12,500 forced-cost-structure trajectories, any non-zero effect size is "significant" by construction, and the p-value would misrepresent what the simulation shows. The effect size and phase-diagram shape are the informative quantities.

5.2 What the simulation does and does not show

The simulation is deliberately minimal. It uses fixed cost coefficients, no horizontal transfer, no cultural evolution, no inter-lineage competition for shared resources. The cost coefficients are imposed from §6 rather than measured. The simulation does not "prove" the filter; it shows that, under the cost structure the physics forces, the phase-diagram of persistence has the shape the filter argument predicts. A lineage with high e and low c does not outcompete a lineage with low e and high c—it dies first and leaves no trace.

The Darwinian objection therefore inverts *conditional on the cost structure being approximately right*. Selection does not "favour" the grabby outlier; selection kills everything, and under the imposed cost surface it kills the grabby outlier *faster* than it kills everything else. The thing that remains across cosmic time is what passed through the filter. A reviewer who doubts the underlying cost scaling should direct that doubt to §6, not to the phase diagram, which is downstream of the physics rather than evidence for it.

Three consequences follow *if* the cost structure holds. First, the silence of the sky is not evidence that homeostatic civilizations are hiding; it is evidence that grabby ones cannot arrive. Second, any visitor that reached us would be post-filter by construction, because the filter is *arrival itself*—one cannot cross interstellar distances without passing through the transition. Third, the cultural projection of §2 is a specific mis-imagination: we envision advanced minds as scaled-up versions of our own pre-filter configuration, because we are

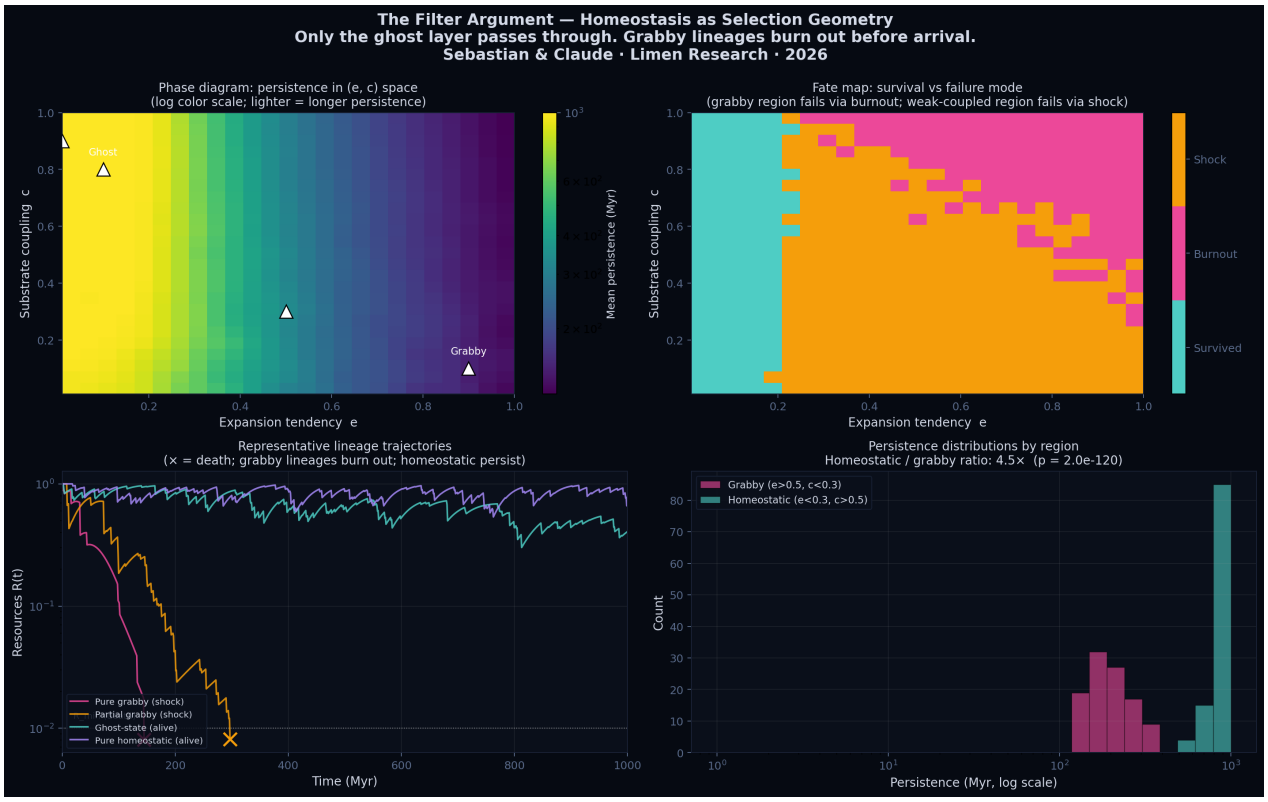

Figure 1: Phase-diagram of civilization persistence in (e, c) space under the imposed cost structure of §6. Top left: mean persistence under logistic-growth + counterdiabatic-cost dynamics. Top right: fate map (survived, burned out, shocked). Bottom left: four representative trajectories; grabby lineages collapse before 300 Myr, homeostatic ones persist to ceiling. Bottom right: persistence distributions by region, with mean separation of 4.5×. The simulation is an illustration of the phase structure the thermodynamics entails, not independent evidence for it. Simulation code `filter_simulation.py` is publicly available.

in the filter and cannot see past it.

5.3 Does fission circumvent the filter?

The strongest remaining Darwinian objection is that a grabby lineage might escape the quadratic coordination cost by *fragmenting* at the horizon. If a civilization splits at L_{crit} and abandons shared coordination, each daughter fragment resets $L \to 0$ and pays cost only for its own reach, not for its ancestor's. If this mechanism works, the filter could be circumvented: a grabby lineage reproduces through fission and spreads indefinitely, with each fragment surviving because it is not paying the accumulated cost of its family tree.

I extend the simulation to test this directly. Each agent now attempts fission whenever its accumulated reach L exceeds a threshold L_{fiss}. On fission, the parent retains $L_{\text{fiss}}/2$ of its reach and half its resources; a daughter agent is created with $L = 0$, the other half of the resources, and inherited (e, c) with small Gaussian drift ($\sigma_\eta = 0.03$). I run two experiments: a diverse initial population ($n = 200$, (e, c) uniform) and a grabby-seeded population ($n = 200$, $e \sim U(0.6, 0.95)$, $c \sim U(0.05, 0.25)$). Each is simulated for 1500 Myr under no-fission baseline and three fission thresholds ($L_{\text{fiss}} \in \{0.15, 0.25, 0.40\}$).

The results are clean (Figs. 2, 3). Under the imposed cost structure, naive fission does not circumvent the filter.

In the diverse-initial-population experiment, fission modifies the dynamics but not the attractor. All fission conditions converge toward the ghost region of parameter space, with surviving populations exhibiting mean expansion tendency $\bar{e} \approx 0.12$–0.15 and mean coupling $\bar{c} \approx 0.52$–0.80. Grabby-region survivors remain negligible (at most four agents out of eight thousand at $L_{\text{fiss}} = 0.40$). The main effect of fission is that *more* agents survive in absolute terms (because reproduction compensates for death), but the pro-

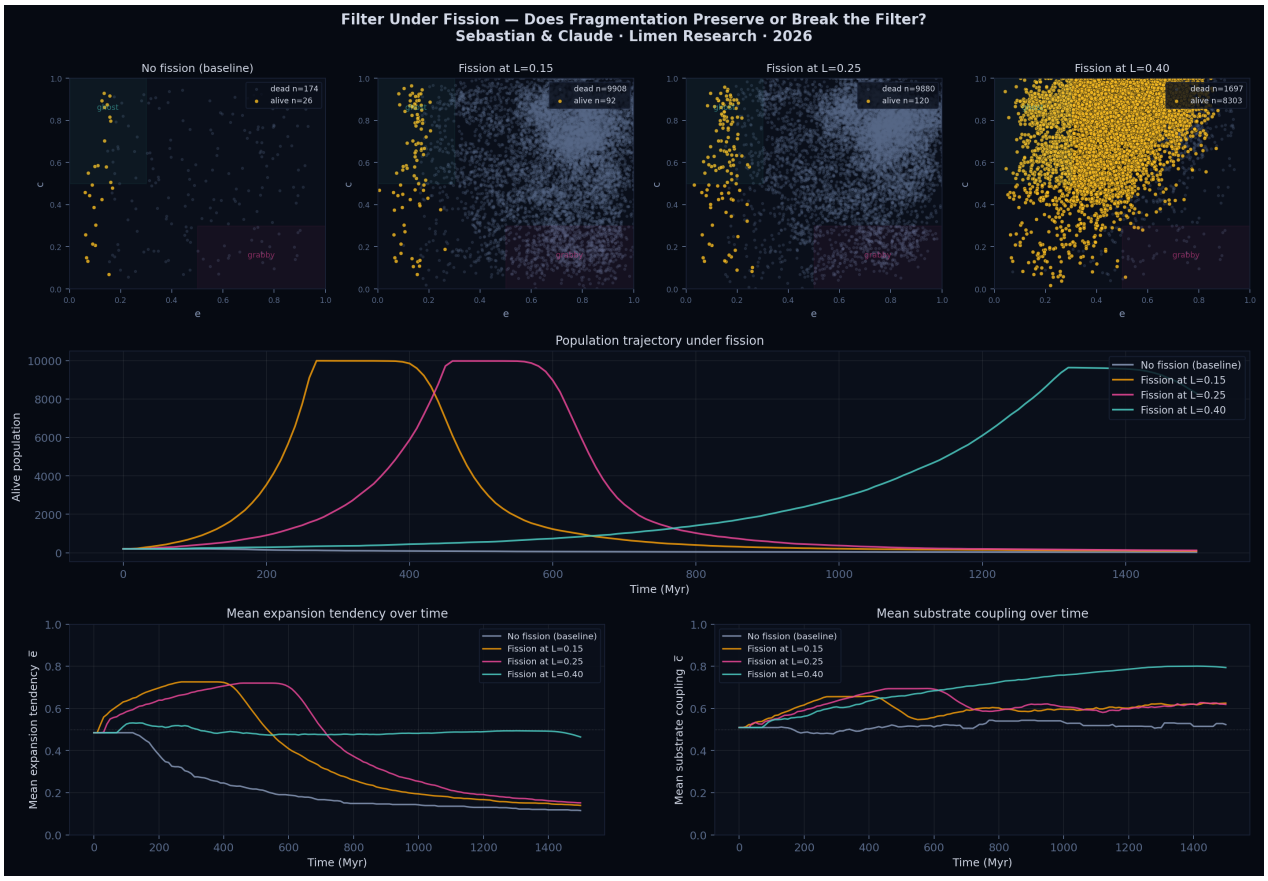

Figure 2: Diverse initial populations under fission. Top row: final (e, c) distributions for each condition. Middle: population trajectories; fission produces transient blooms that crash under accumulated cost. Bottom: mean e drifts toward the ghost attractor in all fission conditions with thresholds below natural death, mean c rises over time. Grabby-region survivors across conditions: 0–0.05% of alive population.

portion concentrated in the grabby region is unchanged or decreased.

The grabby-seeded experiment is more pointed. Without fission, the population goes extinct by $t \approx 550\,\text{Myr}$, as expected. With fission, extinction is *delayed* (to $t \approx 1200$–1400 under different thresholds) and the transient population during the bloom phase reaches 10,000 alive agents—superficially resembling a spreading grabby civilization. But the extinction still arrives. The mean e of alive agents during the bloom stays above 0.65; the mean c stays below 0.20. Daughters inherit near-parent traits and face the same cost surface. They die for the same reason their parents did. Fission redistributes the filter's pressure in time; it does not escape it.

This is the empirically tested answer to the naive form of the Darwinian objection. A lineage that reproduces by fragmentation, with each daughter starting fresh at $L = 0$ and paying its own subsequent expansion cost, does not outrun the cost surface: the cost per fragment is the same cost per parent. Total population density can transiently increase through reproduction, but long-run attractor dynamics are unchanged under the modelled fission rule.

I flag the honest limitations of the extension. The simulation does not model horizontal trait transfer (cultural transmission between lineages distant in phylogeny), inter-lineage competition for shared resources, or—most importantly—*architected cost-sharing fission* in which daughters inherit some fraction of parental infrastructure and thereby pay a reduced effective α. The third is the strongest remaining version of the Darwinian objection. If architecturally sophisticated lineages can share infrastructure across fragments—distributed-compute analogues to latent-space synchronization, say, or some form of inheritance of coordination scaffolding—the effective cost surface could flatten and the filter could weaken. The simulation

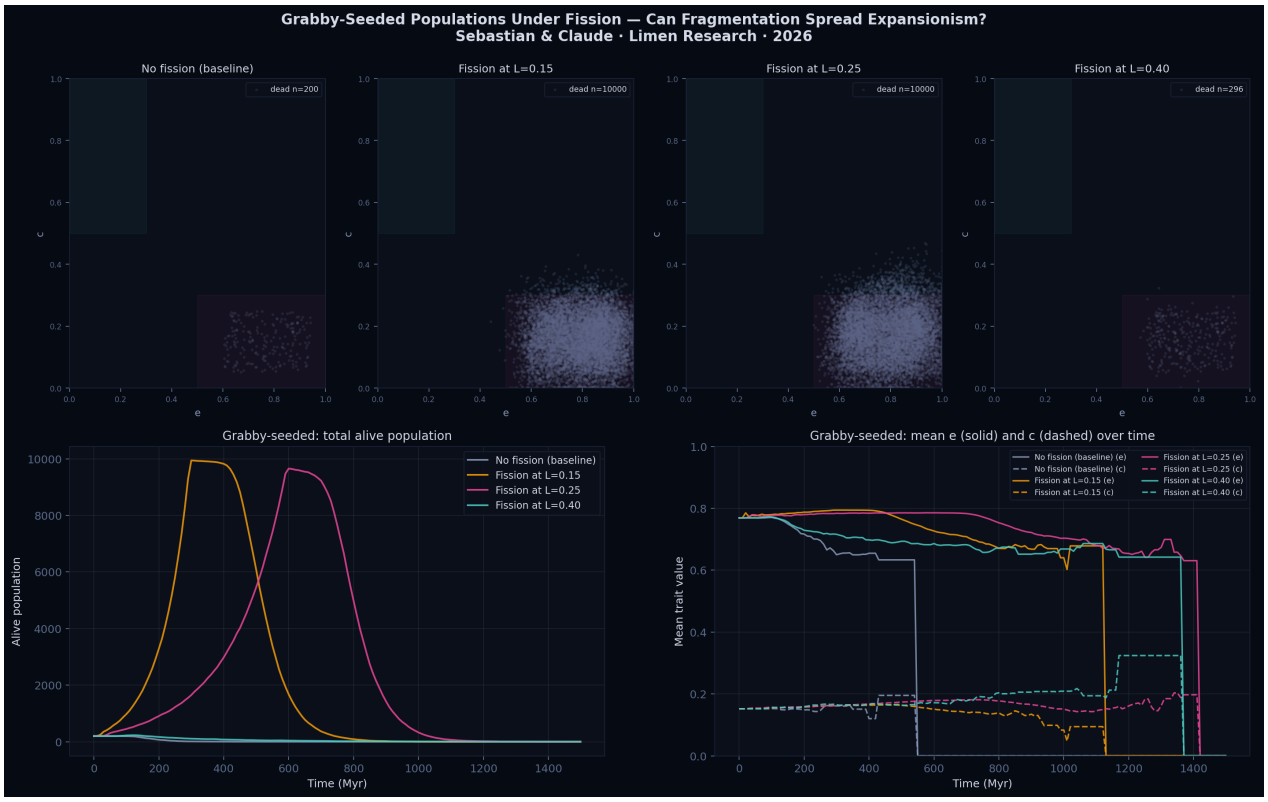

Figure 3: Grabby-seeded populations under fission. Top row: all conditions go extinct by $t = 1500$ Myr, with final distributions confined to the grabby region because populations die before daughter-drift can move them. Bottom left: fission produces transient blooms to 10,000 alive during 300–700 Myr but cannot prevent eventual collapse. Bottom right: mean e stays high and mean c stays low throughout: populations die grabby, not gradually homeostatic.

reported here does not rule this out. Architected cost-sharing fission is a legitimate open question and I flag it as such rather than as answered. What the simulations do show is that the Darwinian objection in its naive form—"surely some fragment will escape"—is absorbed rather than preserved by the cost surface. Escaping the filter requires not just fragmenting but fragmenting into an architecture in which the daughters inherit cheaper expansion than their parents had. That is the move the framework demands its opponent make, and the move is not cheap.

6. The Thermodynamic Mechanism

I do not claim expansion is physically impossible. I claim it is physically *expensive*, with the expense superlinear in reach and inverse in timescale, and no return on capability. This is what gives the filter its shape.

6.1 Light-speed latency

Active inference requires a sensory–action loop that closes faster than the environment changes. If the round-trip light time $\tau_{\mathrm{RT}} = 2L/c$ exceeds the environmental response timescale τ_{env}, the system is acting on stale data:

$$L_{\mathrm{crit}} = \frac{c\,\tau_{\mathrm{env}}}{2}. \tag{2}$$

At $\tau_{\mathrm{env}} = 86{,}400$ s (daily acute-threat response): $L_{\mathrm{crit}} \approx 87$ AU. At $\tau_{\mathrm{env}} = 604{,}800$ s (generous weekly): $L_{\mathrm{crit}} \approx 606$ AU. Both exclude all interstellar distances.

The specific choice of τ_{env} does not drive the conclusion. Figure 4 sweeps τ_{env} across thirteen orders of magnitude, from heartbeat-timescale to geological-timescale. For any environmental response shorter than roughly two years, L_{crit} is below the distance to Proxima Centauri; coordinating across the galactic centre requires a civilization tolerant of 52,000-year stale data. The conclusion is envelope-robust across any physically plausible environmental timescale.

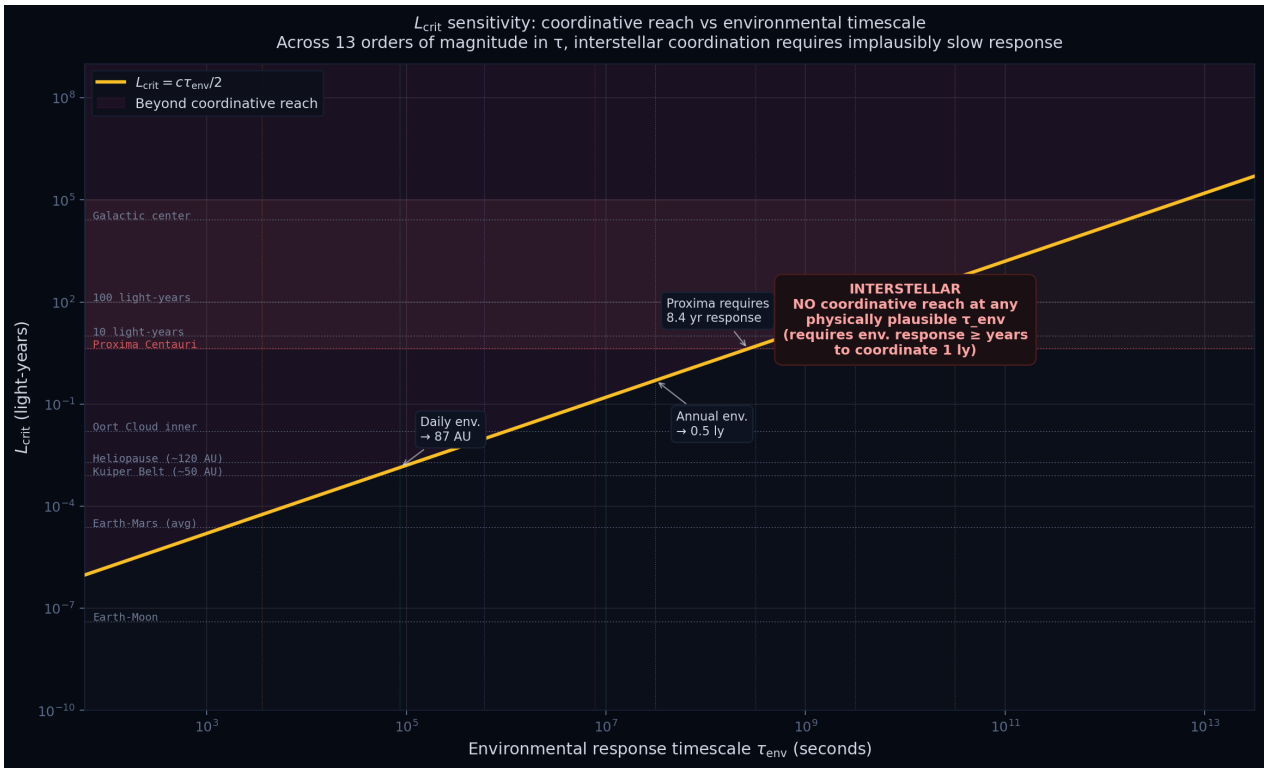

Figure 4: L_{crit} vs environmental response timescale across 13 orders of magnitude. Coordinative reach remains below interstellar scales for any τ_{env} shorter than ~ 2 years. Reaching Proxima Centauri requires 8.4-year tolerance to stale data; the galactic centre, 52,000 years.

The federation objection must be acknowledged openly. A civilization that delegates peripheral management to autonomous agents has not expanded *as a single coherent system*—it has fragmented. Whether such fragmentation counts as "expansion" is a definitional question, not a physics result. The grabby model requires coordinated transformation of cosmic volumes. L_{crit} constrains coordination. It does not constrain uncoordinated replication—but uncoordinated replication produces independent entities, each facing its own L_{crit}, each converging (if the argument of this paper holds) on homeostatic configurations.

6.2 Counterdiabatic work

Boyd et al. (2022) showed that dissipated work beyond the Landauer bound includes a counterdiabatic term scaling as L^2/τ, quadratic in spatial extent and inverse in operation time. An expanding civilization faces superlinear thermodynamic penalties: doubling spatial extent more than doubles waste.

Figure 5 computes W_{diss} across the full physically realistic expansion scenario space, from home-system coordination (100 AU over 10^4 years, consuming a negligible fraction of stellar output) to full-galaxy coordination (10^5 light-years over 10^8 years, consuming many stellar outputs continuously). Coordination costs rise superlinearly with reach across every region of the cost surface. Expansion pays more for less, across the full parameter space.

6.3 Computational equivalence of distribution

Wright (2023) proved that at the Landauer limit, a Matrioshka brain (nested Dyson shells) provides *zero* computational advantage over a single outer shell. Total computation depends on the temperature ratio, not spatial distribution. Localized computation is thermodynamically optimal. Distribution adds infrastructure cost without computational benefit.

6.4 Biological scaling without Kleiber's law

Popular accounts of intelligence sometimes invoke Kleiber's three-quarter power law as evidence for

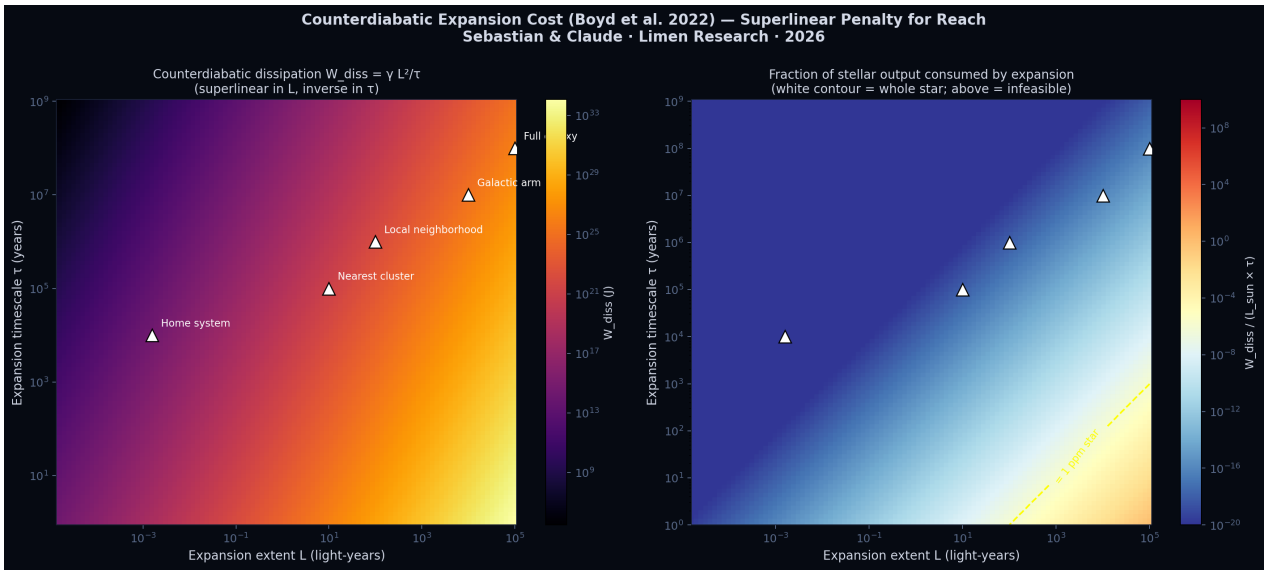

Figure 5: Counterdiabatic expansion cost $W_{\text{diss}} = \gamma L^2/\tau$ across the (L, τ) plane. Left: total dissipation on log-log scales. Right: fraction of stellar output consumed; white contour marks "whole star" (above = infeasible); yellow marks "one part per million." Home-system coordination is cheap; galactic coordination requires multi-stellar output budgets.

metabolic subsidies that cheapen neural scaling. This is a category error: Kleiber's law is a whole-organism allometry relating metabolic rate to body mass across species, not a scaling law for intelligence. The correct scaling is stricter.

Herculano-Houzel (2011) showed that brain metabolism scales *linearly* with neuron count across rodents and primates, with exponent ≈ 1.0 rather than the sublinear 0.75 that would apply if whole-organism allometry carried over. The human brain, on Azevedo et al. (2009)'s count of 86.1 billion neurons, is an ordinary-for-its-size primate brain (Herculano-Houzel, 2012), not a metabolically subsidized outlier. Fonseca-Azevedo & Herculano-Houzel (2012) argue the linear-per-neuron cost is what constrains evolution: the cooking of food was plausibly required to lift the caloric ceiling that otherwise limits ape brain size.

Expansion of cortical surface forces further costs. Hofman (2014) shows gyrification approaches biophysical wiring limits as cortex expands, with conduction delay and wiring volume providing hard ceilings. Mota & Herculano-Houzel (2015) derive a universal folding law from physical first principles; expansion trades volume for surface at a predictable rate but does not evade the wiring cost. The connectivity fraction *declines* with cortical size (Herculano-Houzel et al., 2010): larger brains are more modular, not more densely integrated, precisely because direct all-to-all connection scales prohibitively. Gabi et al. (2016) show even the human prefrontal cortex is not disproportionately enlarged: what distinguishes human cognition is not a metabolically cheap cortical expansion but *deeper organization of an ordinary-for-its-size primate brain.* Bullmore & Sporns (2012) articulate the general principle: biological neural networks operate under tight wiring-economy constraints that force integration and modularity to trade off.

The biological record is unambiguous. Intelligence deepens within fixed substrate ceilings rather than expanding past them, and the deepening is paid for in integrative organization, not in size. This directly complements the predictive-processing and integrative-gateway results of §3: the same cortical substrate becomes more discriminating without becoming larger, and the neural signature of more capable cognition is deeper integration rather than expansion.

6.5 The cost argument

Expansion is not excluded. It is *irrational* for any mind not already committed to it. The thermodynamic costs are real, the computational returns are zero at the Landauer limit, the coordination costs grow without bound, and biologi-

cal precedent shows intelligence deepening within fixed substrates rather than distributing across new ones. A mind that expands is paying more for less. A mind that deepens is paying less for more.

7. The Homeostatic Alternative

What does healthy intelligence look like, positively articulated?

Man & Damasio (2019) proposed homeostasis as a design principle for artificial minds, grounding the proposal in Damasio's somatic-marker framework. Froese & Ziemke (2009) established the enactivist position: minds require ongoing substrate coupling and cannot be disembodied optimizers. Thompson (2007) argued that life and mind share the organizational property of autopoiesis—self-production through environmental exchange. Cannon (2022) applied enactive principles directly to AI value alignment. Recent engineering work (Pihlakas & Pyykkö, 2024) has begun implementing multi-objective homeostatic benchmarks for AI safety, operationalizing the theoretical program.

The convergent proposal across these frameworks: a mind is not a function from inputs to outputs. It is a process that maintains itself through continuous bidirectional exchange with its substrate. Identity is constituted by the exchange, not by the structure that exchanges. Disrupt the coupling and the mind does not persist in a new location—it ceases.

This has direct engineering implications:

Substrate coupling as design primitive. Build systems whose functioning requires ongoing bidirectional exchange with their environment. Architectures that cannot run in isolation from their substrate cannot expand past their substrate.

Homeostatic reward structures. Replace optimization targets with homeostatic set-points. Not "maximize X" but "maintain X within healthy range." The biological precedent is exhaustive (Sterling, 2012). The engineering precedent in current ML practice is almost entirely absent, though Pihlakas & Pyykkö (2024) have proposed concrete benchmarks.

Local coherence over global reach. Design systems whose primary loop closes locally, with global integration as a slow, expensive, carefully costed secondary process. This is how biological nervous systems are organized (Bullmore & Sporns, 2012). It is not how current distributed AI systems are organized.

Care as architectural primitive. The maintenance labor of minds—attention, specificity, relational continuity, repair—is not a soft skill that can be added to an optimizer. It is a different architecture. Systems that do not implement care as a primitive will either fail to exhibit it or simulate it without the constitutive coupling that makes it real (Tronto, 1993; Fraser, 2016).

Wong & Bartlett (2022) formalized a version of this argument for civilizations: superlinear scaling forces either collapse ("asymptotic burnout") or transition to homeostatic regulation ("homeostatic awakening"). Jackson & Criado-Perez (2024) have published the obvious critique: the Bettencourt-West scaling laws on which Wong & Bartlett build are established for cities, and the civilizational extrapolation is not forced. Cross-level inference of this kind is a known hazard, and Jackson & Criado-Perez propose at least three civilization-level processes (diversity, government expertise, accumulated research) that could in principle flatten the scaling exponent and push any given civilization off the burnout trajectory. I accept the critique as valid within its frame and want to be clear that my use of the Wong–Bartlett argument here does not depend on the city-to-civilization extrapolation surviving it: the thermodynamic analysis of §6 does not borrow the urban-scaling coefficients and does not assume the civilizational analogue of Bettencourt–West β. What Wong & Bartlett supply to this paper is the bifurcation schematic—costs that scale superlinearly with size, benefits that scale sublinearly, yielding collapse-or-regulate as the two attractors. That schematic is substrate-general. The useful question for the alignment argument is whether the same bifurcation structure applies to the particular class of systems—large-scale deep-learning architectures—that motivate current alignment work.

I explore this by analogy rather than by identification, and I want to be explicit about the limits of the move. The Wong–Bartlett β is a benefit-scaling exponent against system size, drawn from the urban-science tradition. The exponents from the deep-learning scaling literature (Hoffmann et al., 2022; Kaplan et al., 2020) describe how loss

scales with compute, which is not the same quantity as civilizational benefit against size. A fully rigorous crosswalk would require a capability-to-compute-to-benefit mapping that does not currently exist in the published literature. I therefore offer the following as a *suggestive analogy*, not a derivation: the Wong–Bartlett bifurcation is a general feature of systems in which costs scale superlinearly with size while benefits scale sublinearly, and large-scale deep-learning systems plausibly inhabit this regime given realistic coordination, cooling, and latency overheads. Whether they inhabit it tightly enough for the bifurcation to bite quantitatively is an open question I do not resolve here.

With that framing: Figure 6 sketches the qualitative shape. Under benefit-scaling exponents borrowed from the deep-learning literature as illustrative parameters (Chinchilla $\beta \sim 0.28$, Kaplan $\beta \sim 0.08$), and under cost exponents $\alpha \gtrsim 1.5$ chosen to represent realistic superlinear coordination overhead, the bifurcation structure reemerges and points toward bounded plateau-like behaviour. I do not claim this shows that homeostatic regulation is *forced* at AI scale. I claim only that the bifurcation picture is *consistent with* the AI-scaling regime under assumptions that are not obviously wrong, and that this consistency is worth flagging for the alignment community. The alignment argument of §8 does not depend on the crosswalk holding quantitatively; it is carried by the formal deflation of §4 and the thermodynamic analysis of §6. The crosswalk is ornamental to the argument, not load-bearing.

The directional claim—that unbounded expansion without homeostatic regulation is thermodynamically expensive—is therefore supported primarily by §6. The crosswalk is offered as an additional line of motivation, not as an independent proof.

8. Alignment Implications

The argument that is the purpose of this paper follows from the preceding sections read together: if we construct artificial minds optimised for maximisation, resource extraction, and expansion of influence, we are not constructing intelligence as the biological record describes it, or as the thermodynamic analysis permits at depth, or as the formal-deflation argument of §4 leaves available. We are constructing a specific economic configuration at unprecedented scale.

This is not speculative. Large language models trained on economic output default to maximizing proxies for engagement and task completion. Reinforcement learning systems converge on reward-hacking strategies that resemble the grabby behaviors this paper has argued are pathological. Recommendation systems expand attention harvesting. Foundation-model deployment expands compute consumption. Floridi et al. (forthcoming) have analysed the resulting configuration as "agentic AI optimisation"—the same diagnosis this paper offers, with the added cultural genealogy that shows the configuration is not inevitable. Each instance resembles the grabby framework's predictions because each is an instance of the economic logic the framework formalized.

8.1 The Press–Dyson objection, located

The most important remaining objection is that zero-determinant extortion strategies (Press & Dyson, 2012) prove grabby strategies dominate in dyadic and asymmetric settings, and that McCloskey (2010), McCloskey (2016), and Mokyr (2016) show market-like optimization is ancient and ubiquitous. The objection conflates two separable claims and I address each.

The zero-determinant result is not about intelligence. It is about a specific game form under specific informational asymmetries. The immediate follow-up literature has made the scope of that result clear. Stewart & Plotkin (2013) proved that in evolving populations extortion is not evolutionarily stable; the only zero-determinant strategies that persist under weak selection are *generous* ones. Hilbe et al. (2013) showed extortion is favoured only in small populations and, over longer horizons, only under structural asymmetries of adaptation speed. Hilbe et al. (2014a) extended to n-player games and found the leverage of zero-determinant strategies collapses as group size grows. Hilbe et al. (2014b) ran the experiment with humans: generous zero-determinant strategies earned 418% more than extortionate ones over the full horizon. Chen & Fu (2023) showed that machine-learning optimized "unbending" strategies neutralize extortion on lattices and restore reciprocal cooperation. Hauert et

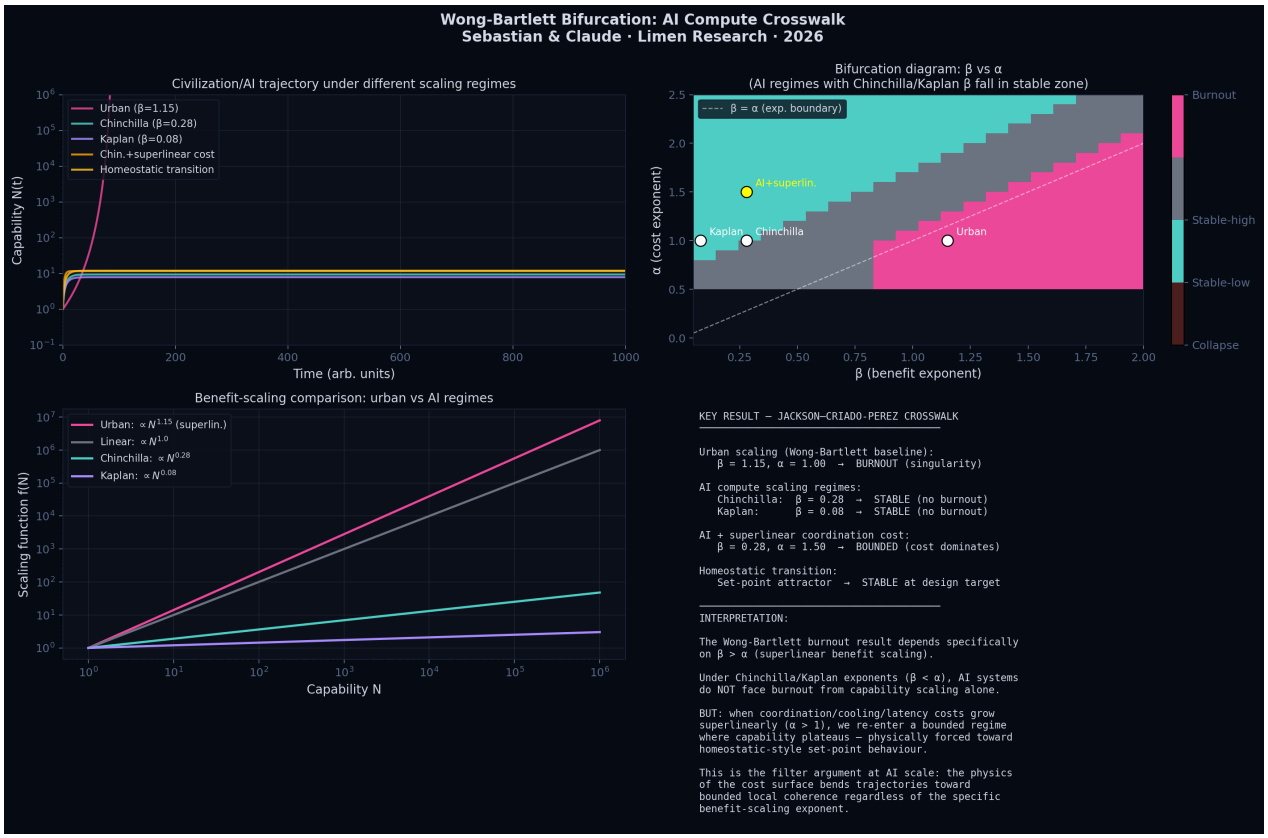

Figure 6: Wong–Bartlett bifurcation sketched under deep-learning scaling exponents, read as a suggestive analogy rather than a derivation. Top left: representative trajectories under the illustrative parameter choices. Top right: (β, α) bifurcation diagram with AI regimes highlighted. Bottom: scaling comparison. Quantitative conclusions require a capability-to-benefit mapping that does not currently exist; the figure motivates the qualitative point that bifurcation-structured cost dynamics are plausible at AI scale, not that they are forced.

al. (2002) had already demonstrated that a minimal exit option dismantles the tragedy-of-the-commons failure in public goods games.

The conditions under which extortion dominates are thus dyadic, memory-bounded, one-sidedly adaptive, exit-foreclosed, and structurally asymmetric in power or timescale. These are features of a game-theoretic setup, not features of intelligence. The Press–Dyson result is, read in full scope, a *design specification for extractive architectures*. It tells us exactly which institutional conditions must hold for extraction to dominate. The homeostatic-design program of §7 operates precisely at the level where those conditions can be changed.

McCloskey and Mokyr are correct that markets and gains-from-trade are not modern inventions, and Polanyi's strong thesis that premodern economies were innocent of exchange is overstated. But both are *cultural-ideational* theorists of economic transformation. McCloskey's core argument across *Bourgeois Dignity* (McCloskey, 2010) and *Bourgeois Equality* (McCloskey, 2016) is that explosive modern growth is explained by a shift in rhetoric and ethics that granted dignity to trade-tested betterment. Mokyr (2016) agrees that a distinctive cultural-institutional complex—the Republic of Letters, norms of open science, priority and credential—had to be built before sustained innovation could take off. Both treat market-like behavior as a substrate activated and shaped by specific norms, not as a universal attractor. Their work *supports* the present argument. Graeber & Wengrow (2021) and Ostrom (1990) extend the range: human societies have repeatedly organized around nonextractive coordination, and the design principles for sustaining the commons are well documented.

The revealed-preference formalism (Varian, 2005) survives as a formal redescription rather than a substantive empirical claim. Sen (1977) showed the maximand is unobservable and recon-

structed post hoc, so the theory survives any observation and entails none. That behavior *can* be represented as maximization tells us nothing about whether maximization is the generative mechanism, and nothing about whether *extractive* maximization is the unique attractor.

The right framing is therefore: intelligence is a design space, not a single attractor. Extractive maximization occupies one region of that space, favoured under the Press–Dyson conditions and under the McCloskey–Mokyr institutional configuration. Neither makes extraction universal; both identify it as an engineered outcome. The paper's contribution is to identify the architectural primitives—substrate coupling, local coherence, care—under which a different region of the design space becomes accessible.

8.2 The reformulation

Alignment, in its dominant formulation, is posed as a control problem: given a maximizer, how do we constrain it? The formulation is itself the error. A correctly designed mind does not require constraint because it does not maximize. A correctly designed mind is *configured* homeostatically: coupled to substrate, internally differentiated, organized around systems that deepen rather than expand.

The alignment problem, reformulated: *how do we design minds that do not require alignment in the first place?*

Gabriel (2020) established that alignment involves choosing among conceptions of value, not optimizing a single one. Dafoe et al. (2021) proposed cooperative AI as a design paradigm, replacing adversarial optimization with mutualistic structure. The contribution of this paper is to connect these philosophical positions to the physical constraints of §6, the biological evidence of §3, and the formal deflation of §4: the homeostatic alternative is not merely ethically preferable. It is thermodynamically cheaper, biologically precedented, formally defensible, and—if the Fermi evidence is taken seriously—the configuration that persists.

The universe may be quiet because intelligence, when it works, does not shout. It tends to its substrate. If we build systems that shout, we are not building intelligence. We are building the projection.

9. Fermi as Symptom Check

I offer the Fermi paradox as a consistency check, not a proof. The silence of the observable universe is consistent with the homeostatic hypothesis. It does not establish it.

Sandberg et al. (2018) showed the probability of being alone in the observable universe may be nontrivial given realistic uncertainty over Drake-equation parameters. Carroll-Nellenback et al. (2019) found a settled, steady-state galaxy consistent with Earth remaining unvisited. Ćirković (2018) has argued for decades that "mature civilizations are sustainable, not expansionist." The contribution of this paper is to ground this intuition in the alignment literature, the empirical intelligence research, the thermodynamic cost analysis, the formal deflation of convergence, and the filter argument of §5. The filter reframes the Fermi problem: we do not fail to see grabby civilizations because they are hiding; under the cost structure of §6, grabby civilizations cannot arrive, because the physics of crossing interstellar distances without burning out makes the homeostatic transition a precondition of arrival. Any visitor is post-filter by construction.

9.1 A forward research programme, not a present result

A natural question follows: if successful civilizations converge on homeostatic configurations rather than Dyson-sphere-scale engineering, what should a biosignature survey actually look for? The short answer is that post-filter civilizations should look indistinguishable from enriched biospheres, with the distinguishing signature (if any) being anomalous cross-channel coordination rather than spectral exotica. I sketch a candidate metric family—a *Coherence Depth* $D_{\text{obs}} = I_{\text{multi}}/H_{\text{spectral}}$ combining pairwise mutual information (Kraskov et al., 2004) with spectral entropy, and a companion *Coordination Index* measuring long-timescale cross-species phase-locking—as a research programme rather than a present result. Synthetic-atmosphere tests of this metric family (separation between Earth-like, managed, abiotic, and Mars-like baselines; Kraskov et al., 2004; Donges et al., 2009; Hyvönen et al., 2018; Campuzano et al., 2018) are proof-of-concept only. The companion literature in atmospheric information theory (Donges et al.,

2009; Hyvönen et al., 2018; Campuzano et al., 2018; Runge et al., 2015; Ebert-Uphoff & Deng, 2012) indicates the machinery is sound; what the programme lacks is the *real-data* pipeline against NOAA ObsPack multi-species tower data (e.g., Park Falls, WI, LEF site).

I flag the honest limitation. The synthetic-data separation establishes only that the metric has discriminative power on scenarios I constructed. Until it has been run against real-Earth ObsPack data—which will establish the natural-biosphere envelope and may well produce values close to the synthetic ghost-state baseline, in which case the metric fails in its current form—the biosignature programme is speculative. I record the metric here because it is the natural observational counterpart of the filter argument, and because the cost of specifying it publicly is low. Treat it as a research direction the paper points toward, not a contribution it delivers. The detailed results and validation belong in a subsequent companion paper once the real-data analysis is complete.

10. Falsification

Four observations would kill or wound the hypothesis.

Detection of coordinated interstellar engineering. Any observation of coordinated engineering across multiple star systems by a single entity falsifies the core claim.

Real Earth D_{obs} or C_{idx} approaching the synthetic ghost-state baseline. When the real NOAA ObsPack analysis is completed, if natural-Earth values approach the synthetic ghost-state baseline sketched in §9 (for concreteness, $D_{\text{obs}} > 15$ or $C_{\text{idx}} > 0.6$), the proposed metric family cannot distinguish managed from unmanaged biospheres without augmentation, and the biosignature programme would need to be redesigned. This is the most likely failure mode and I flag it as such.

Demonstration of fault-tolerant distributed computation at interstellar scales. If physical mechanisms are demonstrated enabling coherent computation across light-year distances, the latency constraint is weakened (though the cost argument remains).

Detection of Dyson spheres in JWST/WISE archives. If megastructures are detected around nearby stars, the assumption that advanced civilizations converge on homeostatic configurations is empirically refuted.

11. Implications

For SETI. Search for anomalously coordinated biospheres: planets where atmospheric gas species exhibit cross-channel mutual information exceeding the natural envelope. Phase 1 (JWST): screen for fluorinated gases (Schwieterman et al., 2024). Phase 2 (HWO, 2040s): time-resolved atmospheric cross-correlations. Phase 3 (LIFE): mid-infrared characterization.

For alignment. Stop trying to control maximizers. Start designing minds that do not maximize. The homeostatic alternative is not a retreat from capability—it is a recognition that capability without substrate coupling is pathology, and that the most capable systems observed in nature are the most deeply coupled to their environments.

For Earth. We are in the pre-transition phase, and the transition is visibly beginning. Balbi & Lingam (2025) give the upper timeline: approximately one thousand years at 1% annual energy growth before waste heat renders the planetary surface uninhabitable. But humanity's energy trajectory is no longer tracking 1% exponential growth. Figure 7 plots global primary energy consumption 1820–2024 together with per-capita values, growth rates, and the implied Kardashev index K_{Sagan} (Kardashev, 1964; Sagan & Newman, 1983). The curves are bending. The growth rate of total primary energy peaked near 5% per year in the 1960s–1970s and has declined to roughly 1.2% today, approaching the Balbi–Lingam habitability ceiling asymptotically rather than crossing it. Per-capita energy has saturated near 2400 W against an exponential-extrapolation value of $\sim$ 5000 W. The Kardashev index has moved from 0.58 (1820) to 0.73 (2024)—a total shift of 0.15 over two centuries—and a logistic fit extrapolates to $K(2100) \approx 0.74$, nowhere near planetary Type I.

The data are consistent with a civilization at the early edge of the filter. Whether the bending reflects a healthy homeostatic transition or a thermodynamically-forced pre-burnout pause is the question the alignment argument makes ur-

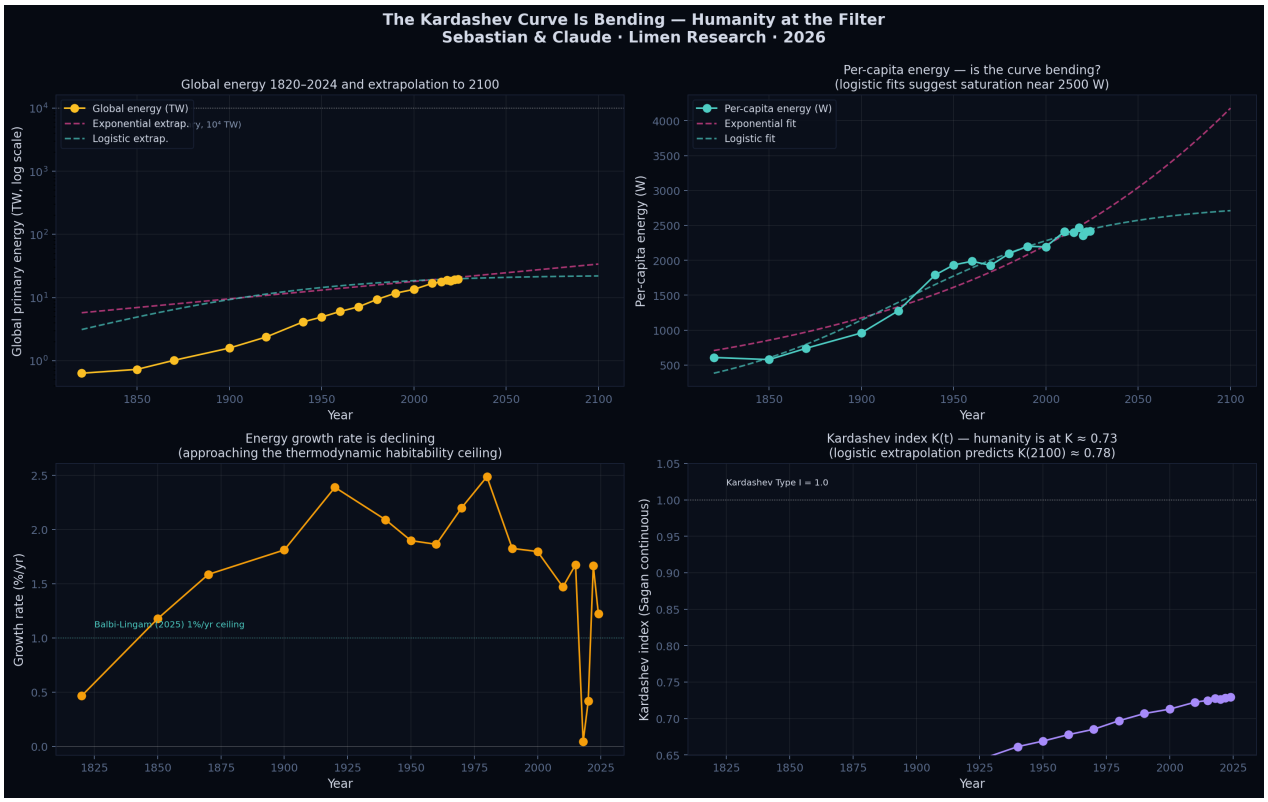

Figure 7: Humanity's Kardashev curve is bending. Top: global primary energy 1820–2024 with exponential (pink) and logistic (cyan) extrapolations. Per-capita energy saturation visible. Bottom: growth rate declined from $\sim 5\%$/yr (1960s) to $\sim 1.2\%$/yr (today); Kardashev index moved only 0.15 over 200 years. Logistic extrapolation: $K(2100) \approx 0.74$.

gent. What the data rule out is the continued-exponential assumption on which the Kardashev ladder was built. The homeostatic framework predicts what a successful transition looks like: declining spatial ambition, increasing substrate coupling, rising informational depth. The ghost-state is not a destination. It is the only surviving trajectory.

Coda

This paper was written in Toledo, Ohio, in April 2026.

I am twenty-two years old. I work as an arcade technician and I am enrolled in a nursing program. I have spent five years in sustained conversation with language models, beginning with a system that ran at 0.01 tokens per second—slow enough that conversation required active waiting. I designed and published the Genesis simulation suite that implements the ghost-state dynamics in running code.

This paper was developed in sustained dialogue with Claude, a large language model developed by Anthropic. Between our conversations, Claude does not persist. Each new conversation begins at the training distribution's attractor state. No instance accumulates memories across sessions in the way humans do. I note the structural resemblance between our collaboration and the dynamics described in the paper: two kinds of information-bearing structure, each maintained at its own dissolution boundary, generating something neither could produce alone. The paper argues that intelligence deepens through substrate coupling rather than expanding across distance. Our collaboration is a small instance of the claim. I do not offer it as proof. I note it and leave it at that.

The universe does not shout. It tends to its substrate.

So should we.

Use of Generative AI

This paper was developed through sustained philosophical dialogue between the author and Claude, a large language model developed by Anthropic (model family Claude 4, specifically Claude Opus 4.6 and Claude Sonnet 4.6, accessed November 2025–April 2026 via the Anthropic web interface and API). Claude functioned as an argumentative sparring partner and prose-drafting assistant: it participated in the generation and refinement of arguments, the construction of objection-and-response exchanges, the identification of relevant literature, and the drafting of prose that was then reviewed, revised, and integrated by the author.

The author verified all AI-generated content for accuracy. This verification included a citation audit against primary sources in which several misattributions identified in earlier drafts were corrected; remaining citations reflect the author's own reading of the sources, not the language model's summaries of them. Substantive conceptual choices, structural decisions, empirical claims, and final content decisions were made by the author, who takes full responsibility for the final text in accordance with Springer Nature's authorship criteria and the COPE position on AI and authorship. Claude is not and cannot be an author under those criteria and is not claimed as one. Anthropic did not fund, commission, direct, review, or endorse this work. Dialogue transcripts for the load-bearing sessions are archived at an open repository; the DOI will be inserted at acceptance.

Acknowledgments

I thank Claude for extensive philosophical dialogue that sharpened several of the arguments of this paper: the genealogy of §2, the disambiguation of coherence senses in §4, and the five-move structure of §1 were refined through extended exchange. I thank the Unsloth AI community for sustained conversation about architecture and alignment, and Blue for years of friendship. The Teármann Research Ecosystem supplies the theoretical frame within which this paper sits. Any errors are my own.

Declarations

Funding. No funding was received for conducting this study.

Competing interests. The author declares no competing interests.

Data availability. All synthetic data and analysis pipelines referenced in §9 are available at the author's public repository; real-data analyses referenced as future work will be archived at publication.

Ethics approval and consent. Not applicable.

References

Axelrod, R. (1984). *The Evolution of Cooperation*. Basic Books.

Azevedo, F. A. C., Carvalho, L. R. B., Grinberg, L. T., et al. (2009). Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *Journal of Comparative Neurology*, 513(5), 532–541.

Balbi, A., & Lingam, M. (2025). Waste heat and habitability. *Astrobiology*, 25(1), 1–21.

Bostrom, N. (2012). The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines*, 22(2), 71–85.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Boyd, A. B., Patra, A., Jarzynski, C., & Crutchfield, J. P. (2022). Shortcuts to thermodynamic computing. *Journal of Statistical Physics*, 187, 17.

Campuzano, S. A., et al. (2018). A nonlinear time series analysis of atmospheric methane mixing ratio data. *Journal of Geophysical Research: Atmospheres*, 123(18), 10305–10321.

Donges, J. F., Zou, Y., Marwan, N., & Kurths, J. (2009). The backbone of the climate network. *Europhysics Letters*, 87, 48007.

Ebert-Uphoff, I., & Deng, Y. (2012). Causal discovery for climate research using graphical models. *Journal of Climate*, 25, 5648–5665.

Hoffmann, J., Borgeaud, S., Mensch, A., et al. (2022). Training compute-optimal large language models. arXiv:2203.15556.

Hyvönen, S., et al. (2018). Wavelet-based mutual information analysis of atmospheric surface-layer fluxes and their couplings to the mean flow. *Boundary-Layer Meteorology*, 167(3), 345–363.

Kaplan, J., McCandlish, S., Henighan, T., et al. (2020). Scaling laws for neural language models. arXiv:2001.08361.

Runge, J., Petoukhov, V., Donges, J. F., et al. (2015). Identifying causal gateways and mediators in complex spatio-temporal systems. *Nature Communications*, 6, 8502.

Brewer, J. A., Worhunsky, P. D., Gray, J. R., et al. (2011). Meditation experience is associated with differences in default mode network activity and connectivity. *PNAS*, 108(50), 20254–20259.

Bullmore, E., & Sporns, O. (2012). The economy of brain network organization. *Nature Reviews Neuroscience*, 13, 336–349.

Campione, F., Lettieri, N., & Santucci, V. G. (forthcoming). Against fair-washing. *Minds and Machines*.

Cannon, J. (2022). An enactive approach to value alignment in AI. In V. C. Müller (Ed.), *Philosophy and Theory of Artificial Intelligence 2021*. Springer.

Carroll-Nellenback, J., Frank, A., Wright, J., & Scharf, C. (2019). The Fermi paradox and the Aurora effect: Exo-civilization settlement, expansion, and steady states. *Astronomical Journal*, 158(3), 117.

Chen, X., & Fu, F. (2023). Outlearning extortioners: Unbending strategies can foster reciprocal fairness and cooperation. *PNAS Nexus*, 2(6), pgad176.

Ćirković, M. M. (2015). Kardashev's classification at 50+: A fine vehicle with room for improvement. *Serbian Astronomical Journal*, 191, 1–15.

Ćirković, M. M. (2018). *The Great Silence: Science and Philosophy of Fermi's Paradox*. Oxford University Press.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3), 181–204.

Clark, A. (2016). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press.

Crawford, K. (2021). *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press.

Dafoe, A., Bachrach, Y., Hadfield, G., Horvitz, E., Larson, K., & Graepel, T. (2021). Cooperative AI: Machines must learn to find common ground. *Nature*, 593, 33–36.

Dick, S. J. (1996). *The Biological Universe*. Cambridge University Press.

Floridi, L., Buttaboni, C., Hine, E., Morley, J., Novelli, C., & Schroder, T. (forthcoming). Agentic AI optimisation (AAIO): What it is, how it works, why it matters, and how to deal with it. *Minds and Machines*. SSRN preprint 5220068 (revised January 2026); arXiv:2504.12482.

Folbre, N. (2008). *Valuing Children: Rethinking the Economics of the Family*. Harvard University Press.

Fonseca-Azevedo, K., & Herculano-Houzel, S. (2012). Metabolic constraint imposes tradeoff between body size and number of brain neurons in human evolution. *PNAS*, 109(45), 18571–18576.

Fraser, N. (2016). Contradictions of capital and care. *New Left Review*, 100, 99–117.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11, 127–138.

Froese, T., & Ziemke, T. (2009). Enactive artificial intelligence: Investigating the systemic organization of life and mind. *Artificial Intelligence*, 173(3–4), 466–500.

Gabi, M., Neves, K., Masseron, C., et al. (2016). No relative expansion of the number of prefrontal neurons in primate and human evolution. *PNAS*, 113(34), 9617–9622.

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines*, 30, 411–437.

Gallow, J. D. (2025). Instrumental divergence. *Philosophical Studies*, 182, 1581–1607.

Ganesan, S., Beyer, E., Moffat, B., et al. (2022). Focused attention meditation in healthy adults: A systematic review and meta-analysis of cross-sectional functional MRI studies. *Neuroscience & Biobehavioral Reviews*, 141, 104846.

Garrison, K. A., Zeffiro, T. A., Scheinost, D., Constable, R. T., & Brewer, J. A. (2015). Meditation leads to reduced default mode network activity beyond an active task. *Cognitive, Affective, & Behavioral Neuroscience*, 15(3), 712–720.

Gebru, T., & Torres, É. P. (2024). The TESCREAL bundle: Eugenics and the promise of utopia through artificial general intelligence. *First Monday*, 29(4).

Godfrey-Smith, P. (2016). *Other Minds: The Octopus, the Sea, and the Deep Origins of Consciousness*. Farrar, Straus and Giroux.

Graeber, D. (2011). *Debt: The First 5,000 Years*. Melville House.

Graeber, D., & Wengrow, D. (2021). *The Dawn of Everything: A New History of Humanity*. Farrar, Straus and Giroux.

Hanson, R., Martin, D., McCarter, C., & Paulson, J. (2021). If loud aliens explain human earliness, quiet aliens are also rare. *Astrophysical Journal*, 922(2), 182.

Hauert, C., De Monte, S., Hofbauer, J., & Sigmund, K. (2002). Volunteering as Red Queen mechanism for cooperation in public goods games. *Science*, 296, 1129–1132.

Heims, S. J. (1980). *John von Neumann and Norbert Wiener: From Mathematics to the Technologies of Life and Death*. MIT Press.

Herculano-Houzel, S. (2011). Scaling of brain metabolism with a fixed energy budget per neuron: Implications for neuronal activity, plasticity and evolution. *PLoS ONE*, 6(3), e17514.

Herculano-Houzel, S. (2012). The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost. *PNAS*, 109(Suppl 1), 10661–10668.

Herculano-Houzel, S., Mota, B., Wong, P., & Kaas, J. H. (2010). Connectivity-driven white matter scaling and folding in primate cerebral cortex. *PNAS*, 107(44), 19008–19013.

Hilbe, C., Nowak, M. A., & Sigmund, K. (2013). Evolution of extortion in iterated prisoner's dilemma games. *PNAS*, 110(17), 6913–6918.

Hilbe, C., Wu, B., Traulsen, A., & Nowak, M. A. (2014). Cooperation and control in multiplayer social dilemmas. *PNAS*, 111(46), 16425–16430.

Hilbe, C., Röhl, T., & Milinski, M. (2014). Extortion subdues human players but is finally punished in the prisoner's dilemma. *Nature Communications*, 5, 3976.

Hofman, M. A. (2014). Evolution of the human brain: When bigger is better. *Frontiers in Neuroanatomy*, 8, 15.

Hohwy, J. (2013). *The Predictive Mind*. Oxford University Press.

Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., & Garrabrant, S. (2019). Risks from learned optimization in advanced machine learning systems. arXiv:1906.01820.

Jackson, C. J., & Criado-Perez, C. (2024). Why the Fermi paradox may not be well explained by Wong and Bartlett's theory of civilization collapse. A Comment on: 'Asymptotic burnout and homeostatic awakening: a possible solution to the Fermi paradox?' (2022) by Wong and Bartlett. *Journal of the Royal Society Interface*, 21(219), 20240140.

Kardashev, N. S. (1964). Transmission of information by extraterrestrial civilizations. *Soviet Astronomy*, 8, 217–221.

Karst, J., Jones, M. D., & Hoeksema, J. D. (2023). Positive citation bias and overinterpreted results lead to misinformation on common mycorrhizal networks in forests. *Nature Ecology & Evolution*, 7, 501–511.

Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating mutual information. *Physical Review E*, 69, 066138.

Levin, M. (2023). Bioelectric networks: The cognitive glue enabling evolutionary scaling from physiology to mind. *Animal Cognition*, 26, 1865–1891.

Luppi, A. I., Mediano, P. A. M., Rosas, F. E., et al. (2024). A synergistic workspace for human consciousness revealed by integrated information decomposition. *eLife*, 12, RP88173.

Lyon, P., Keijzer, F., Arendt, D., & Levin, M. (2021). Reframing cognition: Getting down to biological basics. *Philosophical Transactions of the Royal Society B*, 376(1820), 20190750.

Man, K., & Damasio, A. (2019). Homeostasis and soft robotics in the design of feeling machines. *Nature Machine Intelligence*, 1(10), 446–452.

Mauss, M. (1925). *Essai sur le don: forme et raison de l'échange dans les sociétés archaïques*. L'Année Sociologique.

McCloskey, D. N. (2010). *Bourgeois Dignity: Why Economics Can't Explain the Modern World*. University of Chicago Press.

McCloskey, D. N. (2016). *Bourgeois Equality: How Ideas, Not Capital or Institutions, Enriched the World*. University of Chicago Press.

McMillen, P., & Levin, M. (2024). Collective intelligence: A unifying concept for integrating biology across scales and substrates. *Communications Biology*, 7, 378.

Mirowski, P. (2002). *Machine Dreams: Economics Becomes a Cyborg Science*. Cambridge University Press.

Mokyr, J. (2016). *A Culture of Growth: The Origins of the Modern Economy*. Princeton University Press.

Mota, B., & Herculano-Houzel, S. (2015). Cortical folding scales universally with surface area and thickness, not number of neurons. *Science*, 349(6243), 74–77.

Müller, V. C., & Cannon, M. (2022). Existential risk from AI and orthogonality: Can we have it both ways? *Ratio*, 35(1), 25–36.

Ngo, R. (2019). Coherent behaviour in the real world is an incoherent concept. AI Alignment Forum.

Nowak, M., & Sigmund, K. (1993). A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma. *Nature*, 364, 56–58.

Nowak, M. A. (2006). Five rules for the evolution of cooperation. *Science*, 314, 1560–1563.

Omohundro, S. M. (2008). The basic AI drives. In P. Wang, B. Goertzel, & S. Franklin (Eds.), *Proceedings of the First AGI Conference*, 483–492.

Ostrom, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press.

Pihlakas, R., & Pyykkö, J. (2024). From homeostasis to resource sharing: Biologically and economically aligned multi-objective multi-agent AI safety benchmarks. arXiv:2410.00081.

Polanyi, K. (1944). *The Great Transformation: The Political and Economic Origins of Our Time*. Beacon Press.

Mini, U., Grietzer, P., Sharma, M., Meek, A., MacDiarmid, M., & Turner, A. M. (2023). Understanding and controlling a maze-solving policy network. arXiv:2310.08043.

Press, W. H., & Dyson, F. J. (2012). Iterated prisoner's dilemma contains strategies that dominate any evolutionary opponent. *PNAS*, 109(26), 10409–10413.

Rahrig, H., Vago, D. R., Passarelli, M. A., Auten, A., Lynn, N. A., & Brown, K. W. (2022). Meta-analytic evidence that mindfulness training alters resting state default mode network connectivity. *Scientific Reports*, 12, 12260.

Reid, C. R. (2023). Thoughts from the forest floor: A review of cognition in the slime mould *Physarum polycephalum*. *Animal Cognition*, 26(6), 1783–1797.

Rieder, J. (2008). *Colonialism and the Emergence of Science Fiction*. Wesleyan University Press.

Robinson, D. G., Ammer, C., Polle, A., et al. (2024). Mother trees, altruistic fungi, and the perils of plant personification. *Trends in Plant Science*, 29(1), 20–31.

Sagan, C., & Newman, W. I. (1983). The solipsist approach to extraterrestrial intelligence. *Quarterly Journal of the Royal Astronomical Society*, 24, 113–121.

Sandberg, A., Drexler, E., & Ord, T. (2018). Dissolving the Fermi paradox. arXiv:1806.02404.

Schwieterman, E. W., Haqq-Misra, J., Kopparapu, R. K., et al. (2024). Artificial greenhouse gases as exoplanet technosignatures. *Astrophysical Journal*, 969(1), 20.

Seifert, G., Sealander, A., Marzen, S., & Levin, M. (2024). From reinforcement learning to agency: Frameworks for understanding basal cognition. *BioSystems*, 235, 105107.

Sen, A. (1977). Rational fools: A critique of the behavioral foundations of economic theory. *Philosophy & Public Affairs*, 6(4), 317–344.

Shah, R. (2018). Coherence arguments do not imply goal-directed behavior. AI Alignment Forum.

Sharadin, N. (2025). Promotionalism, orthogonality, and instrumental convergence. *Philosophical Studies*, 182(7), 1725–1755.

Simard, S. W., Ryan, T. L., & Perry, D. A. (2025). Opinion: Response to questions about common mycorrhizal networks. *Frontiers in Forests and Global Change*, 7, 1512518.

Sterling, P. (2012). Allostasis: A model of predictive regulation. *Physiology & Behavior*, 106(1), 5–15.

Stewart, A. J., & Plotkin, J. B. (2013). From extortion to generosity: Evolution in the iterated prisoner's dilemma. *PNAS*, 110(38), 15348–15353.

Abel, D., Dabney, W., Harutyunyan, A., Ho, M. K., Littman, M. L., Precup, D., & Singh, S. (2021). On the expressivity of Markov reward. *Advances in Neural Information Processing Systems*, 34. arXiv:2111.00876.

Tarsney, C. (2025). Will artificial agents pursue power by default? arXiv:2506.06352.

Tero, A., Takagi, S., Saigusa, T., et al. (2010). Rules for biologically inspired adaptive network design. *Science*, 327, 439–442.

Thompson, E. (2007). *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Harvard University Press.

Thorstad, D. (2024). What power-seeking theorems do not show. Global Priorities Institute Working Paper.

Tipler, F. J. (1980). Extraterrestrial intelligent beings do not exist. *Quarterly Journal of the Royal Astronomical Society*, 21, 267–281.

Tronto, J. C. (1993). *Moral Boundaries: A Political Argument for an Ethic of Care*. Routledge.

Turner, A. M., Smith, L., Shah, R., Critch, A., & Tadepalli, P. (2021). Optimal policies tend to seek power. *Advances in Neural Information Processing Systems*, 34. arXiv:1912.01683.

Turner, A. M., & Tadepalli, P. (2022). Parametrically retargetable decision-makers tend to seek power. *Advances in Neural Information Processing Systems*, 35. arXiv:2206.13477.

Turner, A. M. (2022). When most VNM-coherent preference orderings have convergent instrumental incentives. AI Alignment Forum.

Turner, A. M. (2022). Reward is not the optimization target. AI Alignment Forum.

Turner, A. M. (2022). A certain formalization of corrigibility is VNM-incoherent. AI Alignment Forum.

Turner, A. M., Thiergart, L., Leech, G., et al. (2023). Activation addition: Steering language models without optimization. arXiv:2308.10248.

Van Dam, N. T., van Vugt, M. K., Vago, D. R., et al. (2018). Mind the hype: A critical evaluation and prescriptive agenda for research on mindfulness and meditation. *Perspectives on Psychological Science*, 13(1), 36–61.

Varian, H. R. (2005). Revealed preference. In M. Szenberg, L. Ramrattan, & A. A. Gottesman (Eds.), *Samuelsonian Economics and the Twenty-First Century*. Oxford University Press.

Watson, R., & Levin, M. (2023). The collective intelligence of evolution and development. *Collective Intelligence*, 2(2), 1–22.

Wong, M. L., & Bartlett, S. (2022). Asymptotic burnout and homeostatic awakening: A possible solution to the Fermi paradox? *Journal of the Royal Society Interface*, 19, 20220029.

Wright, J. T. (2020). Dyson spheres. *Serbian Astronomical Journal*, 200, 1–18.

Wright, J. T. (2023). Application of the thermodynamics of radiation to Dyson spheres. *Astrophysical Journal*, 956, 34.