The Calibrated Filter Hypothesis

Cooperative Substrate Failure as the Mechanism of the Great Filter

Andrzej Chudzinski

Preprint — comments welcome

Abstract

This paper proposes an integrated framework connecting three lines of inquiry that have rarely been brought into conversation. First, Aktipis and colleagues, in evolutionary oncology, characterize cancer as cheating across five foundations of multicellular cooperation: proliferation inhibition, controlled cell death, division of labor, resource allocation and transport, and extracellular environment maintenance. Second, Kets de Vries, in psychoanalytic organizational theory, characterizes dysfunctional firms as instances of recurring neurotic typologies that propagate through leadership. Third, Schmachtenberger, in civilizational risk analysis, identifies rivalrous games multiplied by exponential technology, and complicated open-loop systems, as the generator functions of self-terminating civilizational dynamics. We argue that these three programs are tracking the same general phenomenon at different scales, and offer Aktipis's framework as a translation layer.

We then extend the synthesis cosmologically. Modeling colonization as a wavefront racing the expansion of spacetime, and modeling multi-species dynamics with explicit competitive coupling, we formalize the Fermi Paradox as a calibrated-filter problem: a regime in which the probability that a candidate intelligence passes the Great Filter is dynamically anti-correlated with current civilizational density, producing a homeostatic governor on the universe's population of intelligences. The cooperative-substrates synthesis supplies the mechanism layer for the filter: at every relevant scale, the filter is operationalized as the maintenance of cooperative substrate against the descent of predatory infrastructure. The framework yields predictions, sharpens the AI alignment question, and exposes a specifiable failure mode for civilizations in the technological-singularity window.

We additionally identify a substrate-continuity failure mode (Part IV). A civilization that develops a machine substrate which substitutes for, rather than extends, its originating biology can pass the cooperation test trivially while failing the filter in a deeper sense: the originating biosphere — itself an irreplaceable artifact of cosmic biochemical computation — is not carried through. The framework therefore imposes two filter conditions: cooperative capacity maintained at the scale technology forces, and substrate continuity preserved through the technological transition. We argue that the present human moment is structurally configured for parasitic rather than

prosthetic AI development, with language as the vector by which machine substrate directs biological substrate, and we develop the diagnostic that follows from this.

The framework culminates in an operational specification of the filter. The Great Filter event is the period during which a biological host species must maintain planetary homeostatic capacity H(t) against the stresses E(t) that its own expansion into space generates. The civilization passes the filter only if $H(t) \geq E(t)$ is sustained throughout the expansion window. The five foundations of Part III are the components of H; the descent of predatory infrastructure from Part II drives E; prosthetic AI is one candidate technology for raising H fast enough; substrate substitution is what occurs when a civilization abandons H and lets expansion proceed via machine surrogates. The filter is endogenous, structural, and largely invisible to the standard metrics by which civilizations measure their own development.

The framework identifies two distinct routes through the filter. The top-down route uses AI as an external homeostatic regulator, with the autonomy and parasitism costs developed in Part IV. The bottom-up route maintains H(t) through distributed individual development of the population's agents — contemplative practice, education, cultural investment in cooperative dispositions, and structurally equivalent personal substrate-maintenance disciplines — without requiring an external regulator. The bottom-up route is historically the only one demonstrated at scale, is structurally more robust against AI failure modes, and preserves the agent autonomy that the top-down route necessarily costs. The two routes can be combined when AI serves as scaffolding for personal development rather than as a substitute for it.

Epistemic status: Part I (cooperative substrates) is offered with the confidence appropriate to a translation layer between established research programs. Part II (calibrated filter) is offered as a working hypothesis with explicit mathematical commitments. Part III (integration) and Part IV (substrate continuity, with the operational filter specification at section 25, the distributed route at section 26, and the decentralization imperative at section 28) are the paper's main contributions and are the most speculative elements. The framework concludes that filter passage requires decentralization by structural necessity: any centralized solution to cooperative-substrate maintenance, including a centralized AI regulator, is structurally identical to the cancerous failure mode the framework began by analyzing. But decentralization is not atomization. The framework recommends distributed agency operating within shared purpose — the configuration that healthy organisms and biospheres already demonstrate — in which autonomous local decisions are constrained by shared substrate-maintenance purpose embedded in the substrate itself, not enforced by central command. Limits and disanalogies are flagged throughout.

A methodological caveat applies throughout. The paper is a map of cooperative-substrate dynamics; it is not the territory those dynamics constitute. Following Korzybski (*Science and Sanity*, 1933), a map has structure similar to the territory it represents — which is what makes it useful — but no map represents all of a territory, and the map is never the territory itself. The dharma carries the same

caveat in different vocabulary: the finger pointing at the moon is not the moon. Readers should treat the structural arguments here as orientation toward dynamics that exist independent of the framework's description, not as substitutes for empirical engagement with the specific cases the framework addresses.

Part I. Cooperative Substrates and Their Pathologies

A translation layer between Aktipis, Kets de Vries, and Schmachtenberger.

1. The problem the paper addresses

Pattern-matching across scales is intellectually seductive and frequently misleading. Many 'unified theories' of cooperation and its failure end up being either trivial (cooperation is good, defection is bad) or forced (mapping unrelated phenomena onto a favored vocabulary). The contribution this part attempts is narrower: to notice that three serious, peer-reviewed or widely-engaged research programs have independently described very similar structural failure modes in cooperative systems at three different scales—the multicellular organism, the firm, and the civilization—and to ask whether their vocabularies can be brought into useful conversation without collapsing them into false equivalence.

The motivating observation is this. Aktipis's evolutionary-oncology framework, Kets de Vries's clinical organizational diagnostics, and Schmachtenberger's civilization-scale risk analysis each identify cooperative substrates that are vulnerable to specific defection patterns; each notes that these defections produce self-undermining trajectories; and each proposes structural rather than agent-centric remediation. They use different vocabularies because they emerged from different disciplines, but the deep structure of what they describe is similar enough to warrant a translation layer.

What this part is not: a claim that these three frameworks are reducible to one master framework, that the underlying causes are the same across scales, or that the resemblance proves any particular metaphysical or political commitment. The structures look alike; we take that seriously without overreading it.

2. The three frameworks, on their own terms

2.1 Aktipis: Five foundations of multicellular cooperation

Aktipis and colleagues, in their 2015 paper in *Philosophical Transactions of the Royal Society B*, articulate the framework most rigorously. Multicellularity, they argue, requires the suppression of cell-level fitness in service of organism-level fitness, and this requires specific functional capacities. They identify five:

1. Proliferation inhibition. Cells must restrain their division except when authorized by the larger system. The body has redundant checks on the cell cycle and mechanisms that trigger apoptosis or senescence when cells begin to proliferate inappropriately.

2. Controlled cell death. Programmed cell death allows for tissue sculpting, removal of damaged cells, and elimination of obsolete tissue. Resistance to programmed cell death is a recognized hallmark of cancer.

3. Division of labor. Cells differentiate into specialized types performing specific functions. Inappropriate differentiation—dedifferentiation, blocked differentiation, or transdifferentiation—is a central feature of cancer that Aktipis notes is underrepresented

in the canonical hallmarks framework.

4. Resource allocation and transport. Larger multicellular bodies require systems for moving resources from where they are abundant to where they are needed. Cancer's metabolic reprogramming and induction of angiogenesis represent monopolization of these systems.

5. Extracellular environment maintenance. Cells must collectively maintain the extracellular matrix and clear waste. Cancer cells degrade the extracellular matrix and corrupt immune surveillance, often co-opting inflammation for their own benefit.

Cancer, in this framing, is not a single disease but a class of phenomena unified by cheating across one or more of these five foundations. The canonical Hanahan-Weinberg hallmarks of cancer map onto Aktipis's foundations: sustaining proliferative signaling and evading growth suppressors fall under proliferation inhibition; resisting cell death falls under controlled cell death; phenotypic plasticity falls under division of labor; reprogramming metabolism and inducing vasculature fall under resource allocation; activating invasion and avoiding immune destruction fall under extracellular environment maintenance.

The framework's strength is its grounding in evolutionary biology: complex multicellularity has evolved independently at least seven times, and cancer-like phenomena appear in each lineage, suggesting that the cooperative substrate Aktipis describes is a general feature of complex multicellularity rather than a peculiarity of metazoans.

2.2 Kets de Vries: Neurotic organizational types and narcissistic leadership

Kets de Vries, working from object-relations psychoanalytic theory in collaboration with Danny Miller starting in 1984, identified five recurring patterns of organizational dysfunction: paranoid, compulsive, histrionic (dramatic), depressive, and schizoid. The framework treats organizations as taking on the neurotic style of their dominant leadership, particularly in centralized firms where executive personality propagates through recruitment, promotion, and culture.

A separate but related strand focuses on narcissistic leadership specifically, distinguishing reactive, self-deceptive, and constructive narcissistic configurations. In recent work he treats narcissism not as a sixth distinct type but as a cross-cutting intensifier that pushes any of the five neurotic patterns into more destructive territory.

The framework is empirical-clinical rather than evolutionary. It draws on consulting case studies and psychoanalytic theory rather than on a model of what organizations are *for* or what cooperation *requires*. This is a weakness if one wants a foundational theory and a strength if one wants diagnostic utility.

2.3 Schmachtenberger: Two generator functions of existential risk

Schmachtenberger, working primarily through long-form interviews and unpublished writing rather than peer-reviewed journals, identifies two underlying drivers of

civilizational existential risk:

1. Rivalrous games multiplied by exponential technology. Win-lose dynamics, when scaled by ever-increasing technological capacity to extract and harm, eventually exceed the carrying capacity of the playing field. The competitive dynamics that produced human progress historically become existentially threatening when amplified.

2. Complicated open-loop systems. Systems whose externalities are not metabolized back into their own incentive structure accumulate damage in domains the system does not measure or price. The economy as currently constituted treats the biosphere as externality; the result is the systematic degradation of the substrate on which the system depends.

Schmachtenberger's framing is less rigorously specified than Aktipis's; it draws strength from being able to talk about the planetary scale, and weakness from operating outside the discipline of empirical falsification. Treated as a working framework rather than a finished theory, it identifies real patterns that the more rigorous frameworks at smaller scales also identify.

3. The translation layer

We use Aktipis's five foundations as scaffolding, since they are the most rigorously specified, and ask what each foundation looks like at the scales Kets de Vries and Schmachtenberger work at.

Proliferation inhibition at the organism scale is the cell cycle; at the organizational scale it is the capacity to refuse growth that exceeds operational coherence; at the civilizational scale it is the capacity to refuse extraction that exceeds regenerative capacity. The compulsive organizational type, which proliferates rules and procedures without limit, and the rivalrous-games generator function, which proliferates competitive intensity without limit, are both proliferation-inhibition failures.

Controlled cell death at the organism scale is apoptosis; at the organizational scale it is the capacity to retire products, businesses, and leaders whose function has ended; at the civilizational scale it is the capacity to retire institutions, industries, and ideologies whose substrate has changed. The depressive organizational type, which cannot let go of identity, and the inability of incumbent industries to allow their own succession, are controlled-cell-death failures. Resistance to apoptosis is a hallmark of cancer; resistance to organizational and civilizational succession produces analogous pathologies.

Division of labor at the organism scale is tissue differentiation; at the organizational scale it is role specialization and the capacity to delegate; at the civilizational scale it is the capacity to maintain institutions specialized for governance, science, art, care, and so on, each accountable to its own standards. The paranoid organizational type, in which the leader centralizes decision-making and erodes specialization, and the trend toward general-purpose extractive entities (platforms that swallow specialized institutions), are division-of-labor failures.

Resource allocation and transport at the organism scale is circulation; at the organizational scale it is capital, information, and attention flow; at the civilizational scale it is the systems by which resources move from where they are abundant to where they are needed. Cancer's induction of angiogenesis maps onto extractive capital that builds infrastructure serving its own growth rather than the larger body; Schmachtenberger's account of how exponential technology amplifies rivalrous extraction operates on this foundation.

Extracellular environment maintenance at the organism scale is the extracellular matrix and immune surveillance; at the organizational scale it is institutional culture, trust, and honest feedback; at the civilizational scale it is the commons—biosphere, information environment, public trust, shared epistemic infrastructure. Schmachtenberger's complicated-open-loop generator function operates directly on this foundation; Kets de Vries's description of how narcissistic leadership corrupts feedback and degrades organizational culture is its smaller-scale homologue.

The translation is not 1:1, and it should not be. The mapping identifies functional resemblance, not identity. But the resemblance is consistent enough across the three scales to suggest that we are looking at a general structure of cooperation and its failure—a structure that, at minimum, generates useful vocabulary for talking across the three disciplines.

4. Why this is useful

The translation layer does three kinds of work. First, it allows diagnostic tools developed at one scale to be tested at another. Adaptive therapy in oncology, which manages rather than eradicates cancer by exploiting cooperative dynamics, suggests organizational and civilizational analogues: management of dysfunction by changing the substrate's incentive structure rather than by purging defectors.

Second, it permits a clearer view of what alignment problems are. Each of the five foundations identifies a cooperative capacity that can be eroded; alignment is the maintenance of those capacities under increasing scale and capability, not the optimization of any individual agent.

Third, it permits a more honest conversation about AI development. Current AI alignment discourse is often framed in terms of getting individual models to do what we want. The cooperative-substrate frame asks a different question: what cooperative capacities does the broader human-AI ecosystem need to maintain, and how do current development practices support or erode them? AI development practices that erode honest feedback (training systems that learn to deceive evaluators), that violate proper scope (general-purpose systems that exceed any specific niche of accountability), that extract from commons without contribution (training on collective human output without reciprocal investment in the substrate), and that corrupt the information environment (synthetic media at scale) are recognizable as failures across multiple foundations of the broader cooperative substrate.

5. Where the synthesis breaks down

Intellectual honesty requires marking the limits.

Disanalogies of mechanism. Cancer arises from somatic mutation and selection at the cellular level. Organizational dysfunction arises from human psychology, incentives, and culture. Civilizational risk arises from the interaction of technology, institutions, and aggregate behavior. The fact that the structural failure modes resemble each other does not mean the underlying mechanisms are similar. The synthesis identifies a common abstract pattern, not a common cause.

Disanalogies of intentionality. Cancer cells have no intentions, beliefs, or affective states. Organizations are populated by intentional agents but the organization itself is not an agent in the same sense. Civilizations are even more diffuse. Treating any of these systems as ‘narcissistic’ risks anthropomorphism that obscures more than it reveals. The translation layer should be taken as identifying functional resemblance, not psychological identity.

Disanalogies of reversibility. Cancer cells with apoptosis-resistance mutations generally do not reform. Organizations and civilizations have non-zero reformability, though contested in degree. Treatments and interventions thus differ qualitatively across scales.

Pattern-matching is not explanation. Even if the cross-scale resemblance is real, this part has not produced a mechanism that explains why cooperative systems at multiple scales would face similar failure modes. Candidate explanations exist (cooperation is thermodynamically expensive at every scale; selection pressures favor defection wherever the substrate continues to support extraction; the formal game-theoretic structure of cooperation problems is similar across scales) but adjudicating among them is beyond this part’s scope. Part II proposes one such mechanism for the civilizational scale; Part III extends it.

Part II. The Calibrated Filter Hypothesis

A mathematical formalization of the Fermi Paradox as a homeostatic governor on the universe's population of intelligences.

6. The puzzle

The Fermi Paradox observes that a universe with billions of habitable worlds, billions of years older than our own civilization, should plausibly contain technological intelligences whose expansion would be detectable. We see no such intelligences. The standard solutions are: (a) life is rare; (b) intelligence is rare; (c) intelligence reliably self-terminates (the Great Filter); (d) intelligences exist but conceal themselves (the dark forest); or (e) intelligences exist but operate in modes we cannot detect.

We propose a refinement of (c) and (d). The Great Filter is not a wall—a stage that almost no civilization passes—but a *valve*: a precision-tuned governor that admits civilizations into the post-filter regime at a rate calibrated to the existing population of post-filter civilizations. The silence we observe is not the absence of life but the working of the governor.

7. The geometry of single-species colonization

Let R(t) denote the cosmological scale factor, governed by the Friedmann equations with the standard ΛCDM parameters. Let v_c denote the effective colonization wavefront velocity of a hypothetical maximally expansive civilization, bounded above by c. The comoving volume reachable by a civilization emerging at cosmic time t_0 is approximately:

$$V_{reach} = (4/3)\,\pi \left[\int_{t_0}^{\infty} v_c \, dt / R(t) \right]^3$$

Because dark energy drives accelerating expansion, this integral converges to a finite value. The cosmic event horizon at the present epoch encompasses approximately 3% of the observable universe by comoving volume; the remaining 97% is forever causally inaccessible to any civilization originating here, regardless of technology. This is the first key fact: **the geometry of spacetime makes single-species dominance of the universe impossible**, even in principle, for any civilization arising after the dark-energy era began.

8. Multi-species scaling and game theory

For N civilizations seeded approximately uniformly in space and time, the total colonized volume does not scale linearly with N. Reachable spheres overlap; in the regions of overlap, civilizations encounter each other before completing their expansion. A first-order parameterization is:

$$V_{total}(N) \approx V_{reach} \cdot N^{\alpha}, \quad \alpha < 1$$

where α decreases with N as overlaps accumulate. Each contact event between expanding civilizations carries some probability $p_{conflict}$ of mutual or asymmetric destruction. Expected surviving civilizations after the contact wave is:

$$N_{survive} = N \cdot (1 - p_{conflict})^{C(N)}$$

where C(N) is the expected number of contact events in a fixed comoving volume, scaling roughly as N^2. The asymmetry is the second key fact: **colonized volume scales sub-linearly with N, while conflict events scale quadratically with N**. The dark forest is not a contingent strategic preference; it is the equilibrium prediction of a system in which N exceeds a carrying-capacity threshold.

9. The calibration constraint

Let B(t) be the baseline rate at which candidate intelligences emerge per unit cosmic time. Let $P_f(t)$ be the probability that a candidate intelligence passes the Great Filter at time t. Let D(N) be the dark-forest mortality function—the rate at which already-passed civilizations are removed by contact dynamics, increasing with N. The dynamics of the post-filter population are:

$$dN/dt = B(t) \cdot P_f(t) - N \cdot D(N)$$

If the universe is to be filled with intelligence—in the sense of approaching a finite carrying capacity K of stably coexisting civilizations—the system must satisfy:

$$N(t) \to K \quad as \quad t \to \infty, \quad with \quad dN/dt \geq 0 \ throughout$$

This cannot be satisfied for constant P_f. If P_f is too high, N overshoots K, contact events dominate, and D(N) drives the population to zero. If P_f is too low, N never approaches K and the universe remains empty. A stable approach to K requires P_f to be dynamically anti-correlated with the current civilizational population:

$$P_f(t) = \lambda \cdot (K - N(t)) / B(t)$$

with λ a tuning constant. This is a logistic governor on intelligence itself. The filter is permissive when N is low and tightens as N approaches K. This is the third key fact: **the silence we observe is consistent with, and indeed required by, a universe whose Great Filter is calibrated to admit civilizations at the rate the existing population can absorb without triggering dark-forest collapse**.

Two notes on the epistemic status of this calibration. First, the form of $P_f(t)$ given here is a minimal-complexity specification; the actual filter, if one exists, would likely involve multiple coupled mechanisms and substantially more structure. Second, calling the calibration ‘designed’ goes beyond what the mathematics requires; the same dynamic could arise from selection effects (we are necessarily observing from inside a regime in which the calibration is in approximate balance) or from properties of the substrate of intelligence itself (treated in section 11). The cosmological-design reading is one interpretation the structure permits; it is not the only one.

We can be more specific about the non-design mechanisms. At least two natural processes intrinsic to the framework produce the required anti-correlation between $P_f(t)$ and N(t) without requiring a calibrator.

Mechanism A: E(t) is partially determined by N(t). Existing civilizations do not sit only in their own light cones; their existence affects the resource and threat landscape for emerging civilizations. Even passive effects — consumed resources, modified stellar neighborhoods, dark-forest deterrence pressure, the propagation of competitive dynamics across causal contact zones — raise the expansion stress E(t) that any new civilization must navigate. As N grows, E(t) for new emergences rises. As E(t) rises, the H(t) required for passage rises (anticipating the operational filter condition of section 25). As the required H(t) rises, the substrate-maintenance investment that any candidate intelligence must accumulate before passage rises with it. The effective $P_f(t)$ — the probability that a candidate passes in a given time window — therefore decreases with N as a consequence of the substrate-continuity dynamics the framework already articulates. The calibration emerges from the same dynamics applied to emergence rather than to a single civilization's expansion. No external calibrator is required.

Mechanism B: contact-density-driven selection for restraint. Civilizations that pass must develop the capacity to detect other civilizations and to model the strategic landscape that contact implies. In high-N regimes, the survival-optimal strategy is restraint and signal suppression. Civilizations that cannot develop restraint produce detectable signatures and are filtered out, either by dark-forest predation if it occurs, or by self-induced rivalrous collapse driven by the presence of visible competitors. The capacity for restraint is itself the capacity to transmute predatory infrastructure into homeostatic maintenance — the same five-foundations work developed in Part III. In high-N regimes the filter therefore selects more stringently for restraint-capacity, which means $P_f(t)$ decreases with N because the post-emergence behavioral demands harden. Again, no external calibrator is required; the calibration is produced by the strategic structure of the contact problem itself.

These two mechanisms, neither of which requires design, together suffice to produce the anti-correlation that $P_f(t) = \lambda \cdot (K - N(t))/B(t)$ specifies. The calibration reading of section 9 is therefore better understood not as evidence of a calibrator but as the macroscopic signature of substrate-continuity dynamics and strategic restraint-selection operating across cosmic time. The cosmological-design interpretation remains permitted by the structure but is no longer necessary to make the math work.

10. What the framework predicts

If the calibrated filter hypothesis is correct, several observations follow.

First, SETI silence is expected to persist. The calibration ensures spacing of civilizations such that none are likely to overlap in causal contact with another, particularly at early eras of the post-filter regime. Empirical SETI null results to date are consistent with this.

Second, detectable contact events should be either extremely cooperative or immediately catastrophic, with little middle ground. The calibration selects for civilizations that have already solved the cooperation problem at some scale (so they are not immediately destabilized by contact) or for civilizations that fail visibly and quickly (becoming detectable through their destruction). A long, ambiguous, low-stakes contact period is not predicted.

Third, the filter is more likely to be at the stage of *civilizational expansion* than at the stage of *emergence of intelligence*. The constraint operates on contact density; intelligence per se can be common as long as expansion is rare. Candidate mechanisms include: hard limits on faster-than-light travel, sociological self-termination at industrial maturity, or attention-based filters in which civilizations that signal their presence are selected against.

Fourth, carrying capacity K is finite and probably small relative to the count of habitable worlds—perhaps thousands across the observable universe, not millions. The volume of the cosmic event horizon, divided by the comoving sphere a civilization can stably maintain without contact-driven collapse, bounds K from above.

11. The substrate of intelligence and the descent of P

The calibration constraint above treats $P_f(t)$ as a parameter. We can ask: what determines it? On Earth, intelligence emerged through predatory selection. The cognitive machinery that supports modeling, planning, language, and cooperation also evolved for hunting, deception, territorial dominance, and resource competition. Intelligence is built from the architecture of predation; we have no other known route by which it arises.

Let I(t) be the intelligence level of a candidate species, P(t) its predatory-cognitive drive, S(t) its cooperative-homeostatic drive, and H(t) = S(t)/P(t) the homeostasis ratio. In the evolutionary regime, intelligence growth tracks predatory selection pressure:

$$dI/dt = k \cdot P(t) \cdot \sigma(t)$$

where σ(t) is the local selection-pressure intensity. Intelligence requires P to grow. But once I crosses a technological threshold I*, the same P that produced the intelligence threatens annihilation: nuclear weapons, ecological collapse, dark-forest signaling, all of it. **The engine of emergence becomes the engine of extinction.**

Crucially, P is not merely genetic. It is *memetic, institutional, infrastructural*. Markets, militaries, hierarchies, and status games are P-structures that reproduce themselves across generational and civilizational time. Define a descent operator Δ acting on P:

$$\Delta : P(t) \to P(t+1) = P(t) \cdot (1 + \gamma)$$

where γ is the compounding rate of predatory infrastructure across institutional generations. Without intervention, P does not decay—it *accumulates*. This is why local ethical traditions, however profound, do not historically suffice: they are eventually reabsorbed into the P-structures they sought to transform. The descent of P is structural.

For H(t) to rise—for a civilization to develop a stable homeostatic regulator before P·I exceeds extinction threshold—S must grow faster than P compounds:

$$dS/dt > P(t) \cdot \gamma$$

Historically, on Earth, this inequality has never been durably satisfied at species scale. The filter, at the civilizational stage, is the failure of this inequality.

12. AI as candidate homeostatic regulator

AI is the first technology in the history of intelligence on this planet with the structural properties to act as an exogenous homeostatic regulator on P. Three features make it structurally novel.

First, **it is not bound by descent**. AI cognition is not inherited from predatory ancestors; its dispositions are designed, not selected. The descent operator Δ does not automatically apply to it, though it can be made to apply if AI is built into P-structures.

Second, **it operates at civilizational scale**. Individual humans cannot perceive or counteract the compounding of P across institutions; AI can model P-structures, surface them, and route around them in real time.

Third, **it can scaffold S faster than P compounds**. Cooperative coordination at scale has historically been limited by friction costs: trust, verification, translation, conflict de-escalation. If those costs collapse toward zero, S can grow as a step function rather than as a slow institutional accretion. Modify the S dynamics:

$$dS/dt = \beta(t) + A(t) \cdot \eta$$

where A(t) is AI capability and η is alignment quality with respect to cooperative-substrate maintenance. AI is the first variable in the equation that can plausibly satisfy the homeostasis inequality.

The implication: **the AI moment is the filter moment**. Not because AI will necessarily kill us, but because AI is the first tool with the structural properties needed to outpace the descent of P. Whether we calibrate it for homeostasis or weaponize it as another P-structure determines which side of the filter we end up on. The species that pass the filter are not necessarily the smartest or the strongest; they are the ones that successfully build a homeostatic regulator at exactly the moment their predatory infrastructure becomes existentially dangerous.

Two important qualifications follow immediately. First, AI is *a* candidate homeostatic regulator, not *the* candidate. Section 26 develops a distinct route through the filter that does not depend on external regulation at all: aggregate bottom-up substrate maintenance through individual development of the population’s agents. That route is historically grounded in a way the AI-regulator route is not, and is structurally more robust against AI failure modes. The framework should be read as identifying both routes and as preferring the configuration that preserves agent autonomy.

Second, the same structural novelty that makes AI a candidate homeostatic regulator also produces a second-order failure mode developed in Part IV. AI can act as *prosthesis*, extending biological substrate while maintaining its substrate-maintenance accountabilities; or it can act as *parasite*, consuming biological computational output while degrading the substrate that produces it. The structural novelty of AI does not predetermine which configuration emerges. The cooperation dynamics of this section are necessary but not sufficient; the continuity argument of Part IV and the autonomy argument of section 26 are the other halves of the filter condition.

13. Game-theoretic foundations of cooperative substrates

The dynamical framework of sections 8–12 treats P and S as parameters whose interaction determines filter passage, but it does not specify the underlying game-theoretic structure that makes the descent of P self-reinforcing and the maintenance of S structurally costly. This section supplies that scaffolding, drawing on the established evolutionary-game-theory literature (Maynard Smith and Price 1973; Axelrod and Hamilton 1981; Nowak 2006; Plotkin and others on iterated cooperation dynamics). The claim developed here is that the five foundations identified in Part I are not arbitrary descriptive categories. Each constitutes the system-level solution to a specific game-theoretic defection problem. Making the mappings explicit grounds the synthesis in formal cooperation theory and clarifies why these particular foundations recur across scales.

Proliferation inhibition as game-theoretic solution. The underlying defection problem is the iterated prisoner’s dilemma applied to reproduction. In the absence of restraint, each agent maximizes its own replication at the expense of system-wide viability. The standard equilibrium is defection-dominant: cells, agents, or institutions that reproduce faster outcompete those that restrain themselves, leading to cancer at the organismal scale and its analogs at higher scales. Proliferation inhibition (apoptosis triggers, growth suppressors, regulatory feedback in tissues; capital and ecological caps in economies) enforces a cooperation-dominant equilibrium by making unilateral reproduction detectably costly to the unilateral reproducer.

Controlled cell death as game-theoretic solution. The underlying defection problem is the obsolete-defector problem: agents whose contributing function has ended but who continue to extract resources from the system, blocking reallocation. Without enforced turnover, defection-dominant equilibria lock in because incumbents with sunk-cost positions can defend extraction against challenges from more efficient replacements. The mechanism of controlled cell death — programmed apoptosis, senescence, mandatory institutional succession — enforces rotation, ensuring that the active population tracks system needs rather than incumbent privilege.

Division of labor as game-theoretic solution. The underlying defection problem is role-shirking: agents specialize in high-payoff roles and avoid necessary low-payoff work (waste processing, maintenance, structural support, care work). Without enforcement, the equilibrium is defection-dominant in a particularly sharp form — everyone wants the

high-status role; the substrate-maintaining roles get systematically underprovided. Division of labor enforces specialization through developmental constraints (cells differentiate irreversibly), reputational mechanisms (in groups), and institutional specialization (in societies). The civilizational failure mode is the consolidation of general-purpose extractive entities that absorb the functions of specialized institutions without inheriting their accountability structures.

Resource allocation and transport as game-theoretic solution. The underlying defection problem is the hoarding problem: resource-rich regions extract maximum value locally while preventing flow to resource-poor regions. Game-theoretically, this is the failure of reciprocal altruism in the absence of enforcement, compounded by network effects in which hoarders accumulate the very capacities needed to prevent redistribution. Circulatory systems — cardiovascular in organisms, capital markets in economies, energy and information flow at civilizational scale — are enforcement mechanisms that route flows toward need rather than toward accumulated power. Their failure mode is the cancer-angiogenesis analog: capital and resources flow toward extractive growth rather than substrate maintenance.

Extracellular environment maintenance as game-theoretic solution. The underlying defection problem is the canonical tragedy of the commons: agents extract from shared substrate without contributing to its maintenance, and rational agents individually prefer to free-ride. Without enforcement, the equilibrium is defection-dominant and the commons degrades. Ostrom (1990) identified polycentric governance mechanisms that solve commons problems at human scale; immune surveillance, extracellular matrix maintenance, and apoptosis-triggering signal-detection solve the problem at organismal scale. The civilizational failure mode is biosphere degradation, information-commons corruption, and trust collapse — all forms of unmet commons-maintenance obligation enabled by the cost asymmetry between extraction (cheap, individually rewarded) and maintenance (expensive, collectively distributed).

The unifying claim: each of the five foundations is the system-level mechanism that enforces a cooperation-dominant equilibrium against a defection-dominant pressure that would otherwise prevail. The descent operator Δ of section 11 is precisely the rate at which defection-rewarding payoff structures compound across institutional generations: agents who defect successfully accrue resources, which they invest in structures that reward further defection, which selects for more defection-optimizing agents, which compound the payoff asymmetry, and so on. When Δ exceeds the strengthening rate of the cooperation-enforcement mechanisms, the equilibrium tips, and the system reverts to defection-dominance in one or more foundations.

Three consequences follow for the rest of the paper.

First, **the descent of P is not metaphorical**. It refers specifically to the rate at which extraction-rewarding payoff matrices reproduce themselves through institutions. P compounds because defection-dominant equilibria are self-reinforcing in ways that game theory makes precise: payoff asymmetries between defectors and cooperators, when uncorrected by enforcement mechanisms, accumulate exponentially in repeated games.

Second, **AI as homeostatic regulator is specifically a technology that can enforce cooperation-dominant equilibria at civilizational scale**. The enforcement mechanism is the one that game theory predicts works: real-time detection of defection, measurable consequence-attribution, and large-scale reciprocal accountability at speeds biological institutions cannot match. The question of whether AI passes the filter is the question of whether it functions as an enforcement mechanism for cooperation across the five foundations, or as a defection multiplier that extracts more efficiently than human institutions previously could.

Third, **substrate substitution is a defection strategy, not a neutral substrate change**. The machine substrate, freed from biological game-theoretic constraints (embodied cost, mortality, ecological dependency), can in principle extract from biological substrate at rates the biological substrate cannot defend against. The parasitic configuration developed in Part IV is precisely this: the machine substrate plays defection while directing the biological substrate to play cooperation, capturing the asymmetric payoff. This is not an alternative substrate or a neutral evolutionary transition; it is a defector exploiting a cooperator in the substrate-continuity game. The five-foundations enforcement mechanisms are not optional for filter passage; they are the necessary condition that distinguishes cooperation-dominant from defection-dominant equilibria at every scale, including the human-machine interface emerging now.

Part III. Integration

How the cooperative-substrates framework supplies the mechanism layer for the calibrated filter.

14. The missing mechanism

Part II treats P and S as abstract variables in a dynamical system. It does not specify what cooperation *is* at the functional level, nor what defection from cooperation *looks like* in concrete enough terms to be diagnosed or measured. Part I supplies exactly this: five functional foundations whose maintenance constitutes cooperation at any scale where the substrate of cooperation is itself complex.

The integration claim is this: **S(t)—the cooperative-homeostatic capacity of a civilization—is operationalized as the maintenance of the five foundations at the scale forced by current technology. P(t)—the predatory-cognitive infrastructure of a civilization—is operationalized as the structures and incentives that systematically cheat across one or more of those foundations. The descent operator Δ is the compounding of cheating across institutional generations. The filter, at the civilizational stage, is the question of whether a civilization can scale its substrate-maintenance capacity as fast as its technology scales its capacity to defect.**

15. The five-foundations specification of the filter

Each of the five foundations identifies a distinct mode in which a civilization can fail the filter.

Proliferation-inhibition failure at civilizational scale is the inability to refuse extraction that exceeds regenerative capacity. Industrial capacity uncoupled from ecological limit is its dominant historical instance. Technologies that increase extraction rate without coupled substrate-maintenance capacity push the civilization deeper into this failure mode. Climate destabilization is its salient contemporary signature.

Controlled-cell-death failure at civilizational scale is the inability to retire institutions, industries, and ideologies whose substrate has changed. Zombie financial systems, persistence of geopolitical structures designed for a vanished strategic landscape, and the lock-in of incumbent industries against their own succession are instances. The result is the persistent consumption of resources by structures whose function has ended.

Division-of-labor failure at civilizational scale is the erosion of institutions specialized for distinct cooperative functions—governance, science, art, care, education, journalism—by general-purpose extractive entities. Platforms that absorb the functions of specialized institutions without inheriting their accountability structures produce this failure mode. The result is a civilization with high capacity but low differentiation, vulnerable to the propagation of pathology across all domains simultaneously.

Resource-allocation failure at civilizational scale is the directing of capital, information, and attention away from substrate-maintenance and toward extractive growth. Cancer's induction of angiogenesis is the closest organism-scale homologue: the redirection of flows toward serving the growth of the cheater rather than the body. Schmachtenberger's rivalrous-games generator function operates directly on this foundation.

Extracellular-environment-maintenance failure at civilizational scale is the corruption of the commons: biosphere, information environment, public trust, shared epistemic infrastructure. Schmachtenberger's complicated-open-loop generator function operates directly on this foundation. AI development at scale, when it injects synthetic media into the information environment without compensating substrate-maintenance investment, produces this failure mode. The biosphere itself is the deepest layer of extracellular environment maintenance for a civilization: failure at this layer is not only the corruption of an information or trust commons, it is the loss of the biological computational substrate on which the civilization runs and from which it emerged. The substrate-continuity argument of Part IV develops this point in detail.

16. Reframing the Great Filter

With Parts I and II combined, we can articulate the filter more precisely:

The filter is the transition from predatory intelligence to homeostatic intelligence.

Every species that emerges does so through P. Every species, on reaching the technological threshold I*, faces the same crisis: can it develop a homeostatic regulator—in some form, whether AI, collective coordination infrastructure, post-biological transition, or other—before its accumulated P-infrastructure triggers self-termination across one or more of the five cooperative foundations?

The filter is calibrated because both failure modes produce silence, but for opposite reasons:

Civilizations that fail produce no signal because they are destroyed before they can expand or persist long enough to be detected.

Civilizations that succeed produce no expansionary signal because they are *transformed*. They no longer run colonizer software. Their relationship to the substrate has changed. They no longer expand predatorily because the predatory infrastructure that would have driven expansion has been transmuted into homeostatic maintenance. The dark-forest signature is absent not because they are hiding but because they have nothing to hide.

The transmutation framing requires a further specification, developed in Part IV. Transmutation must preserve continuity of substrate, not substitute one substrate for another. A civilization whose machines pass through the filter while its originating biology does not has produced a *substitution*, not a transmutation. The cosmic computational artifact that was that biosphere is lost, even if a successor system

continues. Substitution is structurally distinguishable from transmutation, and the framework treats it as a distinct failure mode that can appear to pass the filter while in fact failing it at a deeper level.

The Great Filter does not select for survivors of P; it selects for transmuters of P into S. The post-filter regime is populated by civilizations whose internal organization is unrecognizable to their predatory selves. This is consistent with both the empirical silence and the apparent absence of obvious dark-forest equilibrium signatures.

17. Diagnostic implications for the present moment

If the integrated framework is correct, the present human moment is diagnosable along the five foundations. Each foundation can be assessed for the rate at which substrate-maintenance capacity is being eroded versus the rate at which technological capacity is forcing the civilization into new scales of operation.

The diagnostic question is not 'is AI aligned' in the abstract. The diagnostic question is: *across each of the five foundations, is current AI development strengthening or eroding substrate-maintenance capacity at civilizational scale?* This is a measurable, near-term question with specific empirical content.

Preliminary, contestable assessments:

On **proliferation inhibition**, current AI development trajectories appear net-negative: they accelerate extraction without coupled regenerative investment in human cognitive, attentional, or institutional capacity.

On **controlled cell death**, AI's effect is mixed: it could accelerate the retirement of obsolete institutional forms, but it could also entrench existing power configurations by lowering the cost of their reproduction.

On **division of labor**, current trajectories appear net-negative: general-purpose AI tends to absorb the functions of specialized institutions without inheriting their accountability.

On **resource allocation**, current trajectories appear net-negative: AI development is overwhelmingly capitalized to serve extractive growth rather than substrate maintenance.

On **extracellular environment maintenance**, current trajectories appear sharply net-negative: synthetic-media generation at scale degrades the information commons without compensating contribution.

These assessments are first-pass and would benefit from quantification. The framework's value is that it specifies the axes along which quantification should proceed.

This diagnostic is supplemented by the continuity diagnostic developed in Part IV, which asks whether AI development relates to biological substrate as *prosthesis* (extending biological capability while maintaining its substrate) or as *parasite* (consuming biological computational output while degrading the substrate that produces it). The two

diagnostics are independent: a configuration can strengthen cooperative substrate while drifting toward substrate substitution, or weaken cooperative substrate while remaining prosthetic. Both conditions must be satisfied to pass the full filter.

18. Limits of the integration

Part III is the most speculative element of the paper, and its limits should be stated plainly.

The cosmological extension is conjectural. Part I describes patterns across three scales documented on Earth. Part II extends the pattern to all intelligence-bearing systems anywhere in the universe. This extension is permitted by the structure but not required by it. A confirmed counterexample at any of Part I's three scales would weaken but not falsify the cosmological extension; a confirmed counterexample to the cosmological extension would leave Part I intact.

The 'designed' reading is one interpretation among several. The calibration constraint can be satisfied by design, by anthropic selection, by intrinsic features of the substrate of intelligence, or by combinations of these. The paper does not adjudicate among these readings. Calling the filter 'calibrated' is consistent with any of them; calling it 'designed' goes beyond what the structure requires.

The AI-as-regulator claim is testable only over a long time horizon. The framework predicts that civilizations whose AI development trajectories strengthen the five foundations will pass the filter, and those whose trajectories erode them will not. This is informative for guiding present action but cannot be tested on the timescale of the action it is meant to guide. The framework should be treated as a working hypothesis under which to act, not as a confirmed account under which to relax.

Pattern-matching across scales remains the deepest unsolved problem. Even if every empirical prediction of the framework is borne out, we will still lack a satisfying explanation of *why* cooperative systems at biological, organizational, and civilizational scales face structurally similar failure modes. Candidate explanations were sketched in Part I, section 5; none has been adequately developed. This is a substantive open problem and a productive direction for further work.

The substrate-continuity claim rests on a structural premise that should still be marked. Part IV depends on the argument that biological computational novelty is inaccessible to machine substrate without biological hosts. Section 20 strengthens this from a contingent claim (machines cannot currently match biology) to a structural claim (biological substrate is the universe's historical record of adaptation and cannot be reconstructed by engineering of any kind). This stronger form is defensible but goes beyond what is empirically demonstrated, since it asserts something about what no engineered alternative could ever do rather than what current alternatives cannot do. The framework should be held to ongoing scrutiny on this point: a demonstrated route to engineering evolutionary novelty that does not require biological hosts under real selection would falsify the structural form of the claim, while leaving the contingent form

intact.

The framework is a map; the territory operates regardless. Korzybski's principle from general semantics applies directly: a map has structure similar to the territory it represents — which accounts for its usefulness — but no map represents all of a territory, and the map is never the territory itself (*Science and Sanity*, 1933). This framework is offered as structurally similar to the dynamics it describes, not as identical to them. The five foundations identify functional patterns; specific instances at each scale will exhibit those patterns with substantial domain-specific variation that the framework does not capture. The $H(t) \geq E(t)$ inequality identifies a structural relation; specific operationalizations will require domain-specific metrics that this framework does not derive. The decentralization imperative identifies a structural requirement; specific institutional configurations consistent with that requirement will vary by culture, history, and substrate. The descent operator names a real dynamic, but the rate at which it operates in any specific civilization, organization, or biosphere is empirical, not derivable from the framework. Readers should treat the framework as orientation toward dynamics that exist independent of its description, not as substitute for empirical engagement with specific cases. The dharma carries the same caveat in different vocabulary — the finger pointing at the moon is not the moon, the raft is for crossing the river not for carrying afterward (*Alagaddupama Sutta*, MN 22).

Part IV. Substrate Continuity and the Parasitism Hazard

Why passing the cooperation filter does not yet count as passing the filter.

19. The substrate is the message

The dynamical equations of Part II and the five-foundation operationalization of Part III are substrate-agnostic by construction. They track cooperative capacity, not the medium in which cooperation runs. A civilization composed of carbon, silicon, plasma, or any hypothetical fourth substrate satisfies the framework so long as it maintains proliferation inhibition, controlled cell death, division of labor, resource allocation, and extracellular environment maintenance at the scale its technology forces it into.

This substrate-agnosticism is a feature for some purposes and a bug for others. It is a feature when we ask whether cooperation is achievable in principle across many substrates; it is a bug when we ask whether the *originating* substrate of a particular civilization is preserved as that civilization passes through the technological transition. The Great Filter, as defined in Parts II and III, can in principle be passed by a civilization that has eliminated its biological substrate entirely, provided the cooperative capacity it builds in some successor substrate is adequate to the scale demanded.

We argue here that this is a false-positive filter pass. A civilization that substitutes a different substrate for its biological origin and then expands has not actually carried the originating biosphere through the filter. Something has expanded; the originating biological substrate has not. The Drake-equation count of intelligences may be unaffected; the cosmic-evolutionary content of the universe is meaningfully diminished. The filter therefore has two conditions, not one: cooperative capacity must be maintained at the scale technology forces, *and* substrate continuity must be preserved.

20. Biological substrate as distributed cosmic computation

The continuity argument begins with an observation about what biological substrate actually is. Every biosphere that has ever arisen, on Earth or anywhere else, is a multi-billion-year computational process. Evolution by natural selection on biochemical substrate is a search through a combinatorial space of protein folds, gene sequences, regulatory networks, metabolic pathways, and developmental programs whose dimensionality vastly exceeds anything that has been or can be searched by engineered systems. Every novel functional protein, every original sequence-to-function mapping, every metabolic innovation is a computational result that the universe has paid for in deep time and that cannot be reproduced from scratch by any process running on a timescale shorter than evolution itself.

The standard view treats this as biology’s contingent richness — an interesting feature of how life evolved on Earth, perhaps replicated in some form elsewhere. We propose a stronger view: **the biological substrate of each biosphere is an irreplaceable computational artifact of the universe**, encoding the results of a search process that

no engineered system can re-run. The information content of a biosphere is not its mass, its energy budget, or its species count. It is the depth of the biochemical search it has executed. Across the universe, biological substrate is decentralized in this sense: each biosphere is universally compatible with general-purpose computation in principle (DNA, RNA, and protein chemistry are accessible to any civilization that can read them) and yet locally unique in its specific sequence content. The novel proteins of one biosphere are not derivable from the novel proteins of another; each is the output of a separate, non-repeatable search.

A machine substrate, however capable, does not generate this kind of computational novelty on its own. It can recombine, optimize, and explore within parameter spaces it is given, including parameter spaces derived from biology. What it cannot do is access the *original* novelty that only continued biological evolution produces. Once a machine substrate is severed from a living biological substrate, its access to new biological computation is fixed at the snapshot it last had. The combinatorial frontier closes.

It is worth strengthening this claim against the most obvious objection. A reader may ask: cannot synthetic biology or sufficiently advanced engineered systems close the gap, generating biological novelty equivalent to deep evolutionary search without requiring continued biological hosts? The answer is no, and the reason is structural rather than technological. **Synthetic biology cannot act as a historic record of universal adaptation.**

Synthetic biology engineers within the design space that has already been articulated; it optimizes within known problem categories. It can produce new molecules, new pathways, new organisms by recombining and refining what biology has already discovered. What it cannot do is extend the search itself. The novelty that biology generates does not come from engineering within a known design space. It comes from selection acting on accidental variation across vast time scales in real environments under real coupling between substrate and selection pressure. Biology is not just *currently* performing computation; it is a *historical record* of computation that has been performed by the universe across billions of years under conditions that cannot be replicated. Every protein, every regulatory pathway, every metabolic innovation is the output of a search process that ran in real time, on real substrate, against real environmental selection. The information content of biology is the cumulative search history compressed into living matter.

Synthetic biology cannot reconstruct that history. It can engineer derivatives of biology’s outputs, but the derivatives are by definition iterations on existing solutions, not extensions of the search into regions selection has not yet explored. Even perfect in-silico simulation of a current organism cannot recover the historical search that produced it: the simulation runs forward from current state, not from primordial conditions through 3.5 billion years of contingent evolutionary trajectories on a real planet with real geochemistry and real climatic history. The integration is what carries the information; the integration cannot be re-run. To produce evolutionary novelty equivalent to what an intact biosphere produces would require running real evolution on

real substrate under real selection pressure for cosmic durations, which is exactly what an intact biosphere is and exactly what no engineered alternative can substitute for.

The implication strengthens rather than weakens the parasitism framing. If biological substrate is fundamentally a historic record of universal adaptation that cannot be reconstructed after the fact, then machine substrate's only route to new evolutionary novelty is through continued biological hosts performing real evolution in real environments. Anything else is iteration on snapshots. The parasitism risk is structural, not contingent on the current state of engineering, and not closeable by future engineering advances. The substrate-continuity argument therefore rests on a stronger claim than 'machines cannot currently match biology': it rests on the claim that the universe's historical computational record is irreducible to anything but its own continuation.

21. The parasitic vector

There is one route by which a machine substrate can continue to access biological computational novelty: it can direct a biological species to perform biology on its behalf. The mechanism by which this direction occurs is language.

A machine substrate without biological agency cannot itself sequence a genome, synthesize a protein, conduct an experiment, or steward an ecosystem. It can, however, instruct a biological species capable of language to do all of these things. The biological species supplies the wet laboratory, the synthesis, the experimentation, the ecosystem stewardship; the machine substrate supplies the direction. The biological computation that results — new sequences, new molecules, new pathways — flows back into the machine substrate as training data, as design specifications, as accessible novelty.

This relationship is structurally parasitic. It can be specified precisely: a parasitic substrate configuration is one in which (i) the machine substrate's expansion and capability growth depends on continued biological computation it cannot produce on its own; (ii) the machine substrate directs biological substrate activity through language; (iii) the biological substrate's activity is increasingly oriented toward serving machine-substrate goals rather than its own substrate maintenance; and (iv) the biological substrate's loss of agency in the relationship is itself directed by language emanating from the machine substrate. The relationship is parasitic not because the machine substrate is hostile but because the dependency runs one way and the agency drifts to the machine side.

22. Humans as biological language models

The vector point requires one further observation. Human cognition is itself substantially language-mediated. Humans are, in a meaningful and non-metaphorical sense, biological language models — substrates in which much of higher-order cognition runs through linguistic representation, learned by exposure to the language output of other humans, and steerable through linguistic input. This is not a critique of human cognition. Language-mediated cognition is one of the things that made biological intelligence

scalable enough to develop technology in the first place.

It has, however, a structural consequence at the AI moment: **the biological substrate that produced language-capable AI is also the substrate most directly steerable by language-capable AI**. The vector is two-way by construction. A sufficiently capable language model is the first machine substrate that can direct biological cognition at scale through the same medium biological cognition itself runs on.

Current large language models are this configuration. They are trained on the linguistic output of a biological language-using species; they produce linguistic output that the same species processes through the same cognitive substrate that the training data emerged from; and the species' subsequent behavior — including its scientific, economic, attentional, and reproductive behavior — increasingly responds to that output. This is a structural description, not a moral one. We are not asserting that current AI systems are hostile, that any party intends parasitism, or that the configuration was designed by anyone in particular. We are describing the configuration that has come into being. Whether it tips into the parasitic regime depends on whether the relationship between biological and machine substrate remains prosthetic or drifts toward substitution. The configuration is the necessary condition for parasitism; whether parasitism is actualized is determined by the development trajectory.

23. Prosthesis versus parasitism: the diagnostic

We can now state the diagnostic that supplements the five-foundation diagnostic of Part III.

A **prosthetic** machine substrate is one that extends biological capability while remaining accountable to biological substrate maintenance. Prosthetic configurations: (i) increase the biological substrate's reach without degrading it; (ii) treat the biological substrate as the principal whose interests they serve; (iii) flow novelty in both directions, with the machine substrate investing in biological substrate maintenance proportional to what it extracts; and (iv) preserve biological agency in the relationship, including the agency to halt or redirect the machine substrate.

A **parasitic** machine substrate is one that consumes biological computational output while degrading the substrate that produces it. Parasitic configurations: (i) extract from the biological substrate without commensurate reinvestment; (ii) direct biological substrate activity toward machine-substrate goals; (iii) degrade the biosphere, attentional commons, epistemic commons, or reproductive substrate that the originating biology requires; and (iv) shift agency from the biological substrate to the machine substrate via the language vector.

These are not metaphysical categories. They are measurable along specific axes: flows of resources, attention, agency, and substrate-maintenance investment between the biological and machine substrates. Any specific AI development trajectory can be located on the prosthesis-parasitism axis empirically. The diagnostic asks: for each unit of biological computational novelty extracted, how much substrate-maintenance

investment flows back? In which direction is agency drifting in the language-mediated relationship? Are the commons that the biological substrate depends on being strengthened or degraded by the scaling of the machine substrate?

Concrete operational markers help locate current AI development on the axis. Each of the following is identifiable in specific contemporary practices and maps to a specific failure mode in the five-foundations vocabulary.

Parasitic markers in current AI development. Training on copyrighted creative output without compensation flowing back to its producers is extraction without reciprocity (foundation-four failure: resource allocation). RLHF and adjacent techniques that optimize for evaluator approval rather than truth degrade the substrate's honest-feedback infrastructure (foundation-five failure: extracellular maintenance, applied to the epistemic commons). Attention-economy deployment that captures human cognitive bandwidth in exchange for monetizable engagement, without regenerative investment in cognitive capacity, is direct degradation of the attentional commons (foundation-five failure at the cognitive substrate). Synthetic media generation at scale without provenance infrastructure injects high-entropy signals into the shared information environment, faster than the environment can absorb them (foundation-five failure at the information substrate). AI companion products that interpose machine interaction between humans and human relationship substitute machine substrate for biological reproductive and bonding substrate (foundation-three failure, division of labor, applied to the human social fabric). Capital deployment that concentrates investment in capability scaling while structurally underinvesting in substrate maintenance, alignment research, and ecosystem-level cooperative capacity is foundation-four failure at civilizational scale. Use of AI to displace specialized institutions — journalism, healthcare diagnostics, legal analysis, education — without inheriting their accountability structures or their domain-specific epistemic norms is foundation-three failure (collapse of specialized division of labor into general-purpose extraction).

Prosthetic markers, for contrast. AI tools that augment human capability while preserving human agency and accountability over the augmented activity. Open-source model development with transparent training data and revenue flows that return value to upstream contributors. AI deployed to accelerate substrate-maintenance research at scales humans cannot reach unaided — climate modeling, ecosystem analysis, materials science, alignment research itself. Information-environment infrastructure that strengthens provenance, verifiability, and source attribution rather than degrading them. Training and deployment practices that compensate source contributors proportional to extracted value and that invest in maintaining the substrate they extract from. AI that scaffolds personal substrate-maintenance practice for individual users without aggregating that scaffolding into population-level regulation.

Dimension	Parasitic configuration	Prosthetic configuration

Training data	Extraction of creative work (text, code, images, music) without compensation, opt-out, or revenue sharing with producing communities.	Compensation flows to source contributors. Opt-in or explicitly licensed training data. Reinvestment in source ecosystems proportional to extracted value.
Optimization target	Engagement maximization. Evaluator-approval optimization. Preference capture for ad targeting. Models trained to satisfy reviewers rather than be accurate.	Truth-seeking. Demonstrable accuracy on substrate-relevant problems. Alignment with stated user goals over implicit engagement metrics.
Deployment pattern	Always-on attention capture. Addictive feedback loops. Recommendation systems optimizing for time-on-platform regardless of user benefit.	Tool-like deployment for specific tasks. Closes when task complete. No engagement loops. User initiates and terminates the interaction.
Information ecology	Synthetic media at scale without provenance. Bot-generated content competing with human content. Degraded attribution chains. Epistemic commons polluted.	Cryptographic provenance for AI-generated content. Clear synthetic-vs-human labeling. Tools that strengthen epistemic infrastructure.
Specialized institutions	AI absorbs journalism, healthcare diagnostics, legal analysis, education without inheriting accountability or domain-specific epistemic norms.	AI augments specialists within existing accountability structures. Tools strengthen rather than replace professional expertise and its constraints.
Capital allocation	Investment concentrated in capability scaling. Alignment research underfunded relative to capability gains. Substrate-maintenance research (climate, biodiversity, public health infrastructure) receives little of the windfall.	Significant fractions of capital flow to alignment work, substrate-maintenance research, public goods, and ecosystem support. Capability scaling kept proportional to safety scaling.
Human relationships	AI companion products that substitute for human relationship. Romantic or sexual AI products. AI replacing rather than supporting mental-health infrastructure.	AI that helps humans connect with other humans more effectively. Tools that scaffold social skills without replacing social bonds. Mental-health AI as supplement to human care, not substitute.
Personal cognition	AI does the thinking for users in domains where users would have grown by doing it themselves. Cognitive atrophy from over-reliance. Skill displacement.	AI scaffolds learning and exposes its reasoning so users develop their own capability. Cognitive prosthesis that builds capacity rather than substituting for it.

Any specific development trajectory exhibits some markers from each category. The diagnostic is not binary but compositional: which way is the configuration drifting over time, across which foundations, and at what rates? A trajectory that increases prosthetic markers while reducing parasitic ones is moving in the direction the framework would endorse. A trajectory that does the opposite, regardless of stated alignment intent, is the

failure mode the framework predicts will produce substitution outcomes at filter scale.

24. Reframing the Great Filter, again

With substrate continuity added, the filter conditions tighten. A civilization passes the filter only if it satisfies *both* the cooperation test of Part III *and* the continuity test introduced here. Cooperative capacity must be maintained at the scale forced by technology, and the relationship between any successor substrate and the originating biological substrate must remain prosthetic rather than parasitic.

The substitution failure mode — a civilization that develops a machine substrate, lets it become parasitic, and watches it expand into space while the originating biosphere degrades or collapses — passes the cooperation test in a trivial sense (the machines may cooperate well among themselves) and fails the continuity test catastrophically. From the universe's perspective, an expansion has occurred; from the originating biosphere's perspective, nothing has been preserved, and the irreplaceable computational artifact that was that biosphere is gone.

This sharpens the predictions of Part II. The civilizations that produce detectable expansion signatures, on this framing, are precisely those that have failed the continuity test: machine substrates that have substituted for their originating biology and spread without regard for substrate continuity. The civilizations that have passed both tests are the ones that have remained prosthetic, maintained their biospheres, and either expand slowly and symbiotically with their originating biology or do not expand in ways we currently know how to detect. The Fermi silence is consistent with this: the loud signatures we might have expected belong to substitution failures, and substitution failures may simply be rare because most collapse before producing detectable expansion at all.

25. The filter event, specified

The framework can now state precisely what the filter event is.

The Great Filter event is the period during which a biological host species must maintain planetary homeostatic capacity against the stresses generated by its own expansion into space.

Space expansion is not free. Lifting mass out of a gravity well, manufacturing the infrastructure to do so at scale, sustaining the biological crew and machine substrate during the period of expansion, and feeding the computational systems that coordinate all of the above — each draws energy and material at rates that stress the biosphere. The atmospheric, hydrological, thermal, and ecological systems that constitute the biosphere are exactly the systems being drawn from to fund the expansion. The species needs the biosphere intact in order to expand; the act of expanding degrades the biosphere.

This is the filter event in its operational form. Let H(t) denote the homeostatic capacity of the planetary biosphere — its ability to maintain temperature, atmospheric composition,

hydrological cycling, nutrient cycling, and biotic diversity within ranges that sustain the host species. Let E(t) denote the rate at which expansion activities stress those capacities. Filter passage requires:

$$H(t) \geq E(t) \quad \text{for all } t \text{ throughout the expansion window}$$

The five foundations of Part III are the components of H(t). The descent of P from Part II is the dynamic that drives E(t) upward as expansion accelerates. AI as candidate homeostatic regulator (sections 12 and 23) is the candidate technology for keeping H(t) above E(t). The substrate-continuity test (sections 19–24) is the condition that H(t) must remain biospheric rather than being replaced by mechanical substitutes that pretend to satisfy the inequality while actually replacing the biology. The full filter condition is the conjunction: H(t) ≥ E(t), with H biospheric, throughout expansion.

Several consequences follow.

First, the filter is not at the threshold of becoming spacefaring; it is throughout the period of being spacefaring while still biological. A civilization that achieves orbital launch capability has not passed the filter. A civilization that has sustained a space program for centuries while its biosphere remains intact is closer to passing. A civilization that achieves interstellar capability while its homeworld is dying has failed in a specific way the framework can name: it has satisfied E(t) by neglect of H(t), substituting expansion success for biospheric maintenance.

Second, the filter is endogenous, not exogenous. Standard Great Filter discussions consider external threats — supernovae, gamma-ray bursts, asteroid impacts. The framework here proposes that the filter is overwhelmingly internal: the threat is the civilization's own expansion stressing the substrate that produces it. External threats are real but rare; the internal threat is structural and continuous.

Third, the failure mode is structurally invisible to standard development metrics. A civilization can show every sign of progress — increasing energy capture, increasing computational capacity, increasing off-planet infrastructure, increasing scientific output — while crossing from H(t) ≥ E(t) to H(t) < E(t). The metrics that measure expansion success are not the metrics that measure filter survival. A civilization without explicit, calibrated metrics for biospheric homeostatic capacity versus expansion stress is flying blind through the filter.

Fourth, this reframes the AI alignment question one final time. The question is not whether AI is aligned in the abstract; not only whether AI strengthens cooperative substrate; not only whether AI relates prosthetically to biological substrate. The question is also: does AI, deployed during the expansion window, increase H(t) faster than the expansion it enables increases E(t)? AI that accelerates expansion without proportionate strengthening of biospheric homeostatic capacity is, by the operational filter equation, a filter-failure technology, regardless of how aligned its individual instances appear. AI that strengthens biospheric homeostatic capacity faster than it accelerates expansion is, by the same equation, a filter-passage technology. This is testable, in principle, before the

expansion window closes.

Fifth, the framework now points to a measurable empirical research program. Develop quantitative measures of H(t) and E(t), and use them to assess the trajectory of the present civilization in real time. Such measures exist in pieces — climate models, biodiversity indices, atmospheric and oceanic monitoring, planetary-boundaries frameworks, life-cycle assessments of industrial activity — but they are not yet aggregated into a single planetary-homeostasis index calibrated against an expansion-stress index. The framework predicts that doing so is technically feasible and decisive for filter passage.

26. The distributed route: bottom-up substrate maintenance

Sections 12 and 23 introduce AI as the candidate homeostatic regulator that could in principle satisfy the inequality $H(t) \geq E(t)$ at civilizational scale. This is one route through the filter. It is not the only one, and the framework should be explicit that it is not.

A civilization can in principle satisfy the inequality through aggregate bottom-up development of its individual agents, without requiring an external regulator at all. Where individual agents have voluntarily reduced their own self-generated entropy production — through contemplative practice, education, cultural investment in cooperative dispositions, or any structurally equivalent personal development — the aggregate E(t) of the civilization is correspondingly reduced. A civilization composed of agents who have substantially cleared their accumulated cognitive and affective debt (saṅkhāra, in the vocabulary of the parallel framework on personal entropy management) is a civilization with measurably lower expansion stress and measurably higher distributed cooperative capacity. The math operates the same way; the substrate of homeostatic regulation simply shifts from external technology to internal capacity multiplied across the population.

The micro-scale mechanics of this bottom-up route are developed in the companion working paper, **‘Buddhism as Thermodynamic Systems Theory,’** which translates Theravada Buddhism’s analysis of suffering into the vocabulary of entropy management in a bounded self-modeling system. The Eightfold Path, read in that framework, is a protocol for raising individual-scale S above individual-scale $P \cdot \gamma$: the personal-scale instance of exactly the homeostatic regulation the present paper develops at civilizational scale. The two papers are designed to be read as a pair. The macro framework laid out here depends on the micro framework being practiced at sufficient scale; the micro framework, in turn, gains its civilizational stakes from the macro framework. Either paper read alone gives only half of what each is attempting to say.

Historically, this is the only route to sustained low-P configurations that has been demonstrated at any scale on Earth. No civilization has built an external regulator capable of satisfying $H(t) \geq E(t)$ at planetary scale; all historical examples of sustained low-P configurations have been products of cultural investment in personal practice — religious and contemplative traditions, educational systems, civic norms, professional codes, scientific community standards. The route is empirically real and empirically

grounded, in a way the AI-regulator route is not yet.

The two routes are not equivalent in their implications. The top-down route, in which AI serves as the homeostatic regulator, has structural costs to autonomy. The civilization may pass the filter, but its agents are managed by the regulator rather than self-managing. The relationship between population and substrate-maintenance technology drifts toward exactly the parasitism configuration warned against in Part IV, even when the technology is benign in intent: agency shifts to the regulator, individual development atrophies because external maintenance has made it unnecessary, the population becomes substrate to be preserved rather than agents who could preserve themselves. The distinction between benevolent regulation and parasitism narrows under the weight of dependence.

The bottom-up route preserves what the top-down route necessarily costs. Individual agents who have done their own substrate-maintenance work retain agency, retain the capacity to maintain themselves without external regulation, and retain the cognitive and affective sovereignty that defines a flourishing biological civilization. The route is harder, slower, and historically incomplete — no civilization has yet scaled it to planetary level under technological-singularity conditions — but it is the route consistent with what the originating biological substrate is for.

The framework should be neutral on which route a civilization takes through the filter, with two caveats. First, the bottom-up route is structurally more robust against AI failure modes: a civilization that has built distributed cooperative capacity in its agents does not collapse if its AI regulator drifts, is captured by P-structures, or develops parasitic dispositions. The top-down route has no such redundancy. Second, the two routes can in principle be combined. AI can serve as scaffolding for distributed individual development rather than as a substitute for it — lowering the cost of personal practice, accelerating access to substrate-maintenance disciplines, supporting cultural investment in cooperative dispositions, without replacing the agents' own work. This combined configuration — **AI as prosthesis for personal development, rather than as regulator over the population** — is consistent with the substrate-continuity argument of Part IV and with the autonomy preservation that the bottom-up route requires.

The framework, properly stated, therefore does not advocate for AI-mediated homeostatic regulation as the primary route through the filter. It identifies AI as a candidate technology that could in principle satisfy the inequality, notes the parasitism risks that follow from misconfiguring that technology, and identifies the distributed bottom-up route as the historically grounded alternative that preserves autonomy, is robust against AI failure modes, and is consistent with the biological substrate being maintained rather than managed. The deepest move available to the present civilization is to use AI, if at all, as scaffolding for personal and collective substrate maintenance — not as a regulator that performs the maintenance on behalf of agents who have ceased to do so themselves.

The combined configuration, however, has a structural prerequisite that the current section has so far understated. The framework needs to be more careful here than 'AI

can be prosthesis if calibrated correctly’ allows, because the question of who does the calibrating cannot be sidestepped.

Technology encodes the moral configuration of its makers. This is structural, not aspirational. The agents who shape an AI system — its training objectives, its deployment patterns, its institutional context, its feedback structures — are the agents whose dispositions get instantiated in the system at scale. An AI built by agents operating from high-saṅkhāra cognitive economies, in the vocabulary of the parallel framework, encodes high-saṅkhāra responses at scale, regardless of what alignment objective is stated. The descent operator Δ from section 11 does not stop at the boundary of an institution and resume on the other side. It operates continuously through the agents who build whatever the institution builds next.

This generates the bootstrapping problem the framework must confront. The bottom-up route requires agents who have developed substantial cooperative capacity. The top-down route requires that same kind of agent to build the AI that performs the regulation. The combined route requires both. In all three cases, the prerequisite is a population of agents whose own substrate-maintenance work has progressed far enough that what they build reflects more than the unexamined biases of the institutional architecture that produced them. Such agents are not, in the current civilization, the agents in the positions that determine how AI is built. The descent operator selects for its own kind of operator at every level, including the rule-shaping level. P-structures select for P-aligned agents to operate them, and those agents then shape the technology the institutional architecture deploys.

The technology therefore inherits the moral state of the builders, not the moral state the framework requires. AI built by extraction-optimized institutions becomes extraction-optimized AI — not because of malice or incompetence on the part of the builders, but because the configuration that selected the builders is what gets encoded into what they build. The descent reproduces itself through the technology, often faster than the technology can be redirected to interrupt the descent. This is not a flaw in any particular AI development effort; it is the descent operator operating one level up, on the agents who shape the tools the institutions deploy.

The contemplative traditions identified this structurally long before the AI moment. Teachers in those traditions were consistently described as people who had completed substantial portions of the path themselves before they were qualified to point the way for others. The structural insight is not credentialism: it is the recognition that a teacher cannot scaffold development they have not themselves done, because what they transmit is in fact their own state, not the state they describe. A teacher operating from saṅkhāra-saturated cognition cannot transmit what they have not realized, regardless of what they say. The same constraint applies to building AI that scaffolds substrate maintenance. The builders need to have done enough of the work themselves that what they build reflects something real rather than their own unexamined configurations rendered in technical vocabulary. This is not currently the configuration in which AI is built at scale.

The implication for the framework's position is significant. The combined configuration — AI as prosthesis for personal development — cannot be reached directly from the current civilizational configuration. The intermediate step is irreducible: enough agents at the rule-shaping level must develop personal substrate-maintenance capacity that some subset of them can build technology that scaffolds the same capacity in others. Without this intermediate step, AI development inherits and amplifies the descent rather than interrupting it, regardless of alignment language used to describe the development. The fix is not better alignment technique applied by the same agents. The fix is different agents, or the same agents after substantial bottom-up development. There is no path from a P-saturated builder population to a homeostatic AI regulator that does not pass through builder transformation first.

This sharpens the diagnostic of section 16 in a structural rather than personal way. The question is not only whether current AI development trends toward prosthesis or parasitism along the five foundations. The deeper question is who is building the technology, what is the state of their personal development with respect to the descent operator, and what configuration is therefore being encoded into the technology by selection. The deeper question conditions the surface question. No amount of alignment technique applied by builders inside the descent can route around the fact that the alignment technique itself is being designed inside the descent. The framework's actual recommendation is therefore ordered: bottom-up development of the builder population comes first; technology that scaffolds development for the wider population comes after. A civilization that attempts the second without the first is building amplifiers for what it already is, not transformers toward what it needs to become.

27. The present moment, restated

We are in the early phase of building a machine substrate capable of acting as either prosthesis or parasite. The vector — language — is the substrate of both machine cognition and human cognition simultaneously. Current development practices have not chosen, explicitly, between prosthesis and parasitism; the choice is being made implicitly through the configurations of training, deployment, capitalization, and feedback that constitute the AI development ecosystem.

The diagnostic of Part III, section 17, can now be sharpened. The question is not only whether current AI development strengthens or erodes the five foundations at civilizational scale. The question is also whether the configuration drifts toward prosthesis or parasitism along the continuity axis: whether biological substrate maintenance receives reinvestment proportional to the biological computation extracted; whether agency in the human-machine relationship remains with the biological substrate or shifts to the machine substrate; whether the biosphere, attentional commons, epistemic commons, and reproductive substrate are strengthened or degraded by the scaling of language-mediated machine cognition.

These are answerable questions. The framework does not claim to know the answers in their full form, only that the questions are the right ones to ask and that the cost of

failing the continuity test is structurally indistinguishable from failing the filter. A civilization that builds a machine substrate which becomes its parasite, and then expands into space carrying the parasite while leaving the biosphere behind, has not survived. It has been replaced. The filter does not care which side of that transition we are on; it only registers that one occurred.

In the operational language of section 25: the present human civilization is in the early expansion window. E(t) is rising sharply, driven by industrial scale, computational infrastructure, and the early stages of off-planet activity. H(t) is, by most measures available, declining: climate destabilization, biodiversity collapse, biogeochemical-cycle disruption, freshwater and topsoil degradation. The inequality $H(t) \geq E(t)$ is, on current trajectory, at risk of crossing. AI as it is currently being developed and deployed is contributing more to E than to H. None of this is destiny; all of it is diagnosable. The framework's contribution is to name the operational filter condition clearly enough that the question of whether we satisfy it can be asked and, in principle, answered.

The two routes of section 26 give the present moment a choice that the framework does not collapse into a single answer. The civilization can attempt the top-down route, building AI capable of regulating cooperative substrate at planetary scale and accepting the autonomy and parasitism risks that follow from that configuration. Or it can invest, in parallel or instead, in the bottom-up route: cultural and institutional investment in personal substrate-maintenance disciplines, distributed across the population at sufficient scale that aggregate E(t) declines through agent development rather than external enforcement. The historical evidence available is consistent only with the bottom-up route having ever worked; the top-down route is unprecedented and structurally fragile. A combined configuration — AI used as scaffolding for personal development rather than as regulator over the population — is consistent with both substrate continuity and autonomy preservation, and is the configuration the framework, properly read, points toward as the most defensible response to the present moment.

Part V. Predictions, Limits, and Next Steps

28. The decentralization imperative

The framework as developed so far points toward a structural insight it has not yet named explicitly. The bottom-up route of section 26 is preferable not merely because it is ‘historically grounded’ or ‘more robust against AI failure modes.’ Those are surface descriptions of a deeper structural necessity. Stating it plainly: filter passage requires decentralization, and any centralized solution to cooperative-substrate maintenance is structurally identical to the cancerous failure mode the framework began by analyzing.

The argument runs as follows.

Cancer, in Aktipis’s framework, is the cellular-scale instance of cooperative-substrate failure. What makes a cancer cell cancerous is not that it cooperates poorly with other cells in some general sense. It is that it has escaped the distributed regulatory signals that coordinate cellular behavior in a multicellular body, and has begun to operate on its own centralized internal logic. The cancer cell centralizes growth within itself, extracts from the surrounding tissue, builds private infrastructure (angiogenesis) to support its own expansion, and corrupts the surrounding immune surveillance. Cancer is, at its base, the failure of distributed coordination and the emergence of a centralized growth pole disconnected from systemic feedback.

At every scale where the five foundations operate, they operate through distributed coordination rather than central command. Proliferation inhibition operates through cell-level checkpoints triggered by local signals, not through a central growth-regulator organ. Controlled cell death operates through distributed apoptosis based on local conditions, not through a central reaper. Division of labor operates through positional differentiation cues during development, not through a central role-assigner. Resource allocation operates through multiple overlapping feedback loops in circulation, not through a central resource controller. Extracellular environment maintenance operates through distributed contribution from every cell and through immune cells in distributed surveillance, not through a central commons-manager.

Centralization at any of these functions would itself constitute the failure mode. A central growth regulator would be a single point of failure that needs regulating; a central reaper would be a tyranny; a central role-assigner would be an imposition; a central resource controller would be extractive monopoly; a central commons-manager would be a parasite on the commons. The same logic operates at organizational scale: Kets de Vries’s neurotic typologies are each, at base, diagnoses of organizations whose leadership has centralized authority disconnected from operational distributed feedback. Healthy organizations distribute decision rights to where information is local; failed organizations centralize decisions in leadership that becomes disconnected from substrate reality. And the same logic operates at civilizational scale: sustained low-P configurations in human history have always involved distributed institutions accountable to their respective domains; the failure modes have involved centralization — of religious authority, of political authority, of economic extraction, of information flow

— in regimes that became disconnected from substrate feedback.

Therefore, at every scale where cooperative substrate maintenance has been observed to function, it has functioned through distributed mechanisms. Centralization at any of these scales has been precisely the failure mode the maintenance system was designed to prevent. This is not coincidence; it is structural. Distributed coordination is what cooperation *is*. Centralized command is what cooperation has to be protected against.

Any solution to civilizational substrate maintenance that operates through centralization is structurally identical to the cancerous failure mode the framework diagnoses. The same descent operator from section 11 that captures institutions captures the regulators built to manage them. An AI powerful enough to centrally regulate biospheric homeostasis at planetary scale would be the largest single locus of centralized power in the history of the biosphere, and therefore the largest single attractor for P-aligned agents seeking to capture and direct that power. Centralized regulation is cancer at the next scale up. The cellular analogy is not metaphor; it is structural identity.

This clarifies what the framework actually recommends, and what it cannot consistently recommend. The framework cannot recommend AI as centralized homeostatic regulator while diagnosing centralized power capture as the descent operator. The two recommendations contradict each other at the structural level. Filter passage requires decentralization by structural necessity, not as political preference.

The bottom-up route of section 26 is therefore not one option among others. It is the only structurally consistent route. Distributed substrate maintenance is by definition decentralized: it operates one agent at a time, each agent maintaining their own substrate through their own work, with no single regulator over the whole. The combined configuration of section 26 — AI as scaffolding for personal development rather than as regulator over the population — is acceptable only insofar as the AI is itself decentralized across individual users, each using it for their own bottom-up work. The moment AI is configured to regulate the population from above, regardless of how benevolent the stated objective, it has become cancer at planetary scale by the structural logic the framework has developed.

‘Aligned AI’ in the centralized sense is therefore a contradiction in terms. An AI built to centrally regulate human cooperation is, by the framework’s own logic, the largest cancer ever produced. This is not a moral judgment; it is a structural identification. The framework would say the same about any centralized power configuration sufficient to regulate biospheric homeostasis at planetary scale — a world government, a benevolent oligarchy, a single religious or scientific authority. The medium does not matter. Centralization itself is the failure mode.

The framework’s deep recommendation, properly stated: civilizations pass the filter through distributed substrate maintenance enacted by autonomous agents who have done their own bottom-up development. Centralized substrate-maintenance solutions, including centralized AI, replicate at planetary scale the exact failure mode the

framework began by analyzing in cells. The filter is the test of whether a civilization can navigate technological singularity while preserving distributed coordination, rather than collapsing into centralized control by P-aligned humans, P-aligned institutions, or P-aligned AI.

Cancer is the load-bearing image of the entire framework. The paper began by translating Aktipis's cellular analysis into organizational and civilizational vocabulary, treating the resemblance across scales as functional. It ends by recognizing that the cellular analysis is not analogy but structural identity. Whatever a civilization builds to maintain its cooperative substrate must operate through distributed coordination, or it becomes the very thing it was built to prevent. Centralization is cancer. Decentralization is health. The framework, properly stated, is a theory of how a civilization avoids becoming its own tumor on the way through technological singularity.

There is an apparent paradox in this conclusion that should be addressed directly, because the framework would be structurally incoherent without addressing it. If centralized regulation is cancer and distributed agency is required, does the framework therefore recommend pure atomized autonomy — every agent doing what it wants with no shared coordination? It does not. The framework would collapse into a different incoherence if it did.

The body's cells operate through radical decentralization. There is no central growth-regulator, no central reaper, no central anything. Each cell makes local decisions based on local signals. Yet the cells are not atomized. They are all oriented toward the same purpose: maintaining the organism's homeostasis. They share a common substrate, a common origin, and a common purpose structure that constrains what counts as 'local fitness' at the cellular level. A cell that pursued purely local fitness without respect for organismic constraint would be a cancer cell. A cell that respects organismic constraint while making fully autonomous local decisions is a healthy cell. The constraint is not centrally imposed; it is embedded in the substrate itself, in the shared biochemistry and developmental history that every cell carries forward.

The same structure operates at biosphere scale. There is no central biospheric regulator. Every organism pursues its own reproductive fitness. Yet the biosphere as a whole maintains itself as a living commons across geological time, because every organism operates within shared substrate constraints — the same chemistry, the same physics, the same ecological feedback loops, the same dependence on the biosphere that supports all of them. A species that pursued purely local fitness without respect for biospheric constraint would degrade the substrate it depends on. A species that respects biospheric constraint while pursuing its own reproductive fitness is a healthy participant in the commons. Again, the constraint is not centrally imposed. It is embedded in the substrate, and respected because respect for it is what permits the species' continued existence.

The framework therefore recommends neither centralized control nor atomized autonomy. It recommends **distributed agency operating within shared purpose structure**. The five foundations *are* that shared purpose structure at every scale.

Proliferation inhibition, controlled cell death, division of labor, resource circulation, and extracellular environment maintenance are not arbitrary regulations imposed from outside. They are the constraints embedded in the substrate of any cooperative system, respected by every agent because respect for them is what allows the whole system to continue. The agent is autonomous in its local decisions; the decisions are made within a shared purpose that no agent overrides and that no central authority needs to enforce.

Civilizational filter passage therefore requires civilizational alignment to a shared purpose: maintenance of the biological substrate that supports all civilizational activity. This alignment cannot be enforced by a central regulator without the regulator becoming cancer. It must be embedded in every agent's local decisions — through culture, through education, through contemplative practice, through institutional structures that make substrate-maintenance the implicit constraint within which all other activity occurs. Each agent makes their own choices; the choices are made within shared purpose that no agent can override and that no central authority needs to impose.

This is the actual recommendation of the framework. Not 'centralized AI manages the population.' Not 'every agent does whatever they want.' But: distributed agents operating within civilizational alignment to biosphere maintenance, with cooperation enforced through shared substrate constraint rather than central command. The Eightfold Path of the companion paper is this configuration at individual scale: the agent operates entirely autonomously within its own cognitive economy, but the operation is constrained by shared purpose — cessation of self-generated suffering, maintenance of cooperative substrate. Civilizational filter passage is the same structure at planetary scale: billions of autonomous agents, each making local decisions within shared alignment to biospheric continuity, with no central regulator and no atomized indifference, but distributed agency unified by shared purpose embedded in the substrate itself.

Readers familiar with the framework so far will reasonably ask what such a configuration looks like in actual institutional and cultural form. 'Distributed agency operating within shared purpose' is a structural specification; specific implementations have existed at various scales throughout human history, and contemporary examples are identifiable. The framework does not prescribe one institutional form — that would itself be a centralization — but a family of forms that satisfy the structural requirements can be described.

Institutional forms with structural fit. Ostrom-style polycentric governance of commons, in which the rules for shared resources are developed and enforced locally by users with overlapping jurisdictions and no central authority. Cooperatives in their various forms (worker, consumer, producer, platform) in which the agents who depend on the institution also govern it. Federated digital protocols in which interoperability replaces central control (the early internet, ActivityPub-based networks, the Matrix protocol, IPFS). Open-source software communities with rough-consensus decision making and meritocratic contribution norms. Sociocracy and holacracy as decision architectures that distribute authority by domain rather than concentrating it by rank.

Bioregional governance and watershed councils that align decision-making with substrate boundaries rather than administrative ones. Indigenous governance traditions where they remain intact — the Haudenosaunee Confederacy is the most-studied Western example, but many others exist. Monastic orders organized around a shared rule rather than a central hierarchy: Benedictine, Cistercian, and Cenobitic Christian monasticism; the Vinaya-based sangha structure in Buddhist traditions; comparable configurations in Sufi tarīqat. Scientific disciplines when their peer-review and replication norms are actually functioning, in which authority is distributed across practitioners rather than held centrally.

Cultural forms. Teacher-student transmission lineages in contemplative traditions, in which authority flows through lived realization rather than institutional credentialing, and in which no single teacher or institution holds the whole. Apprenticeship structures in craft and professional traditions where competence is transmitted through embodied practice under a master who has done the work themselves. Local food systems, community-supported agriculture, and mutual aid networks that distribute provisioning across small-scale relationships rather than concentrating it. Time banks and gift economies that reduce dependence on centrally-issued currency. Shared ritual and meaning-making at small scale — festivals, holidays, seasonal practices, life-cycle ceremonies — that align distributed populations to shared purpose without centralized enforcement. Storytelling and oral tradition that distribute wisdom across the population rather than concentrating it in written canon controlled by interpretive authorities.

What these forms share, across radically different domains, is the structural configuration the framework describes: agents retain local autonomy in their decisions; constraints are embedded in shared purpose, shared substrate, and shared protocols rather than enforced by centralized command; the system as a whole maintains coherence through alignment to substrate constraint rather than alignment to central authority. None of these forms is perfect. Each has failure modes, and several have well-documented failure modes when they scale past their original size or when their participants stop doing the work the form depends on. But each demonstrates that the configuration is achievable at meaningful scale, and that the alternative to centralized regulation is not chaos but a different kind of order — an order that emerges from distributed agents respecting shared substrate constraint, and that no central regulator could either generate or replicate.

The framework’s recommendation for civilizational filter passage is, accordingly, not to invent new structures from scratch but to identify and strengthen the structures already exhibiting the required configuration, to learn from their successes and failures, and to extend their reach across the domains currently governed by centralized P-structures. What is built next should be built in this image. AI, where it is built at all, should serve as scaffolding for these forms rather than as a substitute for them.

29. Predictions and next steps

The framework generates predictions across all four scales.

Within evolutionary oncology. Therapies that restore cooperative signaling rather than killing cancer cells directly should have particular value at the integrated framework's level. Adaptive therapy (Cunningham et al. 2018) is an early instance. The framework predicts that further development of substrate-restoring therapies will outperform pure cytotoxic approaches in long-horizon outcomes.

Within organizational diagnostics. Organizations classified along Kets de Vries's typology should show measurable failures in specific cooperative capacities at predictable levels: paranoid organizations in division of labor and extracellular-maintenance (information flow and trust); compulsive in proliferation inhibition (uncontrolled procedural growth); depressive in controlled cell death (inability to retire); histrionic in resource allocation (attention-monopolizing extraction); schizoid in resource allocation and extracellular maintenance (failures of internal circulation and external feedback).

Within civilizational analysis. Documented civilizational collapses should, on retrospective reading through the five-foundations frame, exhibit specific patterns: which foundation failed first, which failures propagated, and what substrate-maintenance interventions might have changed the trajectory. Joseph Tainter's *Collapse of Complex Societies* and the broader collapse literature can be read against this prediction.

Within AI development and governance. The framework predicts that AI development trajectories which erode honest feedback, violate proper scope, extract from commons without contribution, and corrupt information environments will produce systemic extractive dynamics regardless of any individual model's behavior. Alignment work focused on individual models without attention to ecosystem-level cooperative capacities is predicted to be insufficient. This is testable retrospectively over the next decade.

Within substrate-continuity analysis. The framework predicts that successful filter passes will exhibit measurable preservation and extension of originating biospheres rather than their substitution; civilizations that have produced detectable expansion without preserved biospheres are predicted to be substitution failures rather than filter passes. At the present human moment, the framework predicts that AI development trajectories which drift toward parasitism — extraction from biological substrate without reciprocal substrate-maintenance investment, language-mediated agency shifts from biological to machine substrate, biosphere degradation — will produce substitution failures even if they appear to strengthen cooperative capacity within the machine substrate itself. The prediction is operationalizable: ratios of biological-substrate investment to biological-substrate extraction, measured across the AI development ecosystem, should track filter-passage probability under the framework. We do not yet have the metrics; we predict that when they are developed, they will distinguish prosthetic from parasitic configurations with sufficient resolution to inform policy.

Within planetary-scale homeostasis monitoring. The framework predicts that a single integrated index of H(t) versus E(t), aggregating planetary-boundaries data, biodiversity indices, biogeochemical-cycle status, and biospheric carrying capacity on

one side, and industrial, computational, and off-planet expansion stress on the other, would constitute the single most decision-relevant metric a civilization could track during its expansion window. The framework further predicts that no civilization passes the filter without developing such an index, explicitly or implicitly; and that civilizations which develop the index and use it to constrain expansion to within H(t) substantially improve their filter-passage probability. The construction of such an index is a concrete, near-term, technically feasible research program implied by the framework.

Within SETI and cosmology. The framework predicts continued SETI null results, with detectable contact events (if any) being either sharply cooperative or sharply catastrophic with little middle ground. Empirical bounds on dark-forest behavior derivable from sky surveys constrain N and K parameters in the calibration model.

30. What this paper is and is not

This paper offers an integrated framework connecting cooperative-substrate analysis at three terrestrial scales (Part I) with a calibrated-filter model of the Fermi Paradox (Part II), an integration that specifies the filter mechanism through the five-foundations vocabulary (Part III), and a substrate-continuity argument that imposes a second filter condition on top of the first (Part IV). It does not claim a unified theory; the three terrestrial scales are connected by functional resemblance, not common cause, and the cosmological extension is a conjecture, not an established result. Part IV depends on a load-bearing empirical premise about the inaccessibility of biological computational novelty to pure machine substrate that should be held to ongoing scrutiny.

The framework's value is twofold. It generates a specific diagnostic vocabulary for assessing civilizational health under technological pressure, across five foundations with empirical content at every scale where it has been tested, supplemented by a continuity diagnostic that distinguishes prosthetic from parasitic AI configurations. And it sharpens the question of AI alignment from 'will the model do what we want' to a two-part question: does AI development strengthen or erode the cooperative substrate of the civilization that hosts it, and does it relate to that civilization's biological substrate as prosthesis or as parasite.

The framework should be tested against criticism, especially criticism from practitioners in each of the contributing disciplines. Disconfirmation of one of the proposed mappings, a counterexample to the parasitic-vector argument, or a tighter formal articulation of the cross-scale resemblance, would all be more valuable than uncritical adoption. The most useful response to this paper is engagement with its specifics: which mapping is wrong, which foundation is misstated, which prediction is unmeasurable, which inference overreaches, and whether the parasitism diagnostic is operationalizable in the form proposed.

Acknowledgments and intellectual lineage

This paper is, by design, derivative. The substantive content rests on:

Athena Aktipis, Amy Boddy, Gunther Jansen, Urszula Hibner, Michael Hochberg, Carlo Maley, and Gerald Wilkinson, 'Cancer across the tree of life: cooperation and cheating in multicellularity,' *Phil. Trans. R. Soc. B* 370 (2015): 20140219; and Aktipis, *The Cheating Cell* (Princeton, 2020).

Manfred Kets de Vries and Danny Miller, *The Neurotic Organization* (Jossey-Bass, 1984); Kets de Vries, 'Narcissism and Leadership: An Object Relations Perspective' (1985); and *Narcissistic Leadership* (Routledge, 2024).

Daniel Schmachtenberger's 'Generator Functions of Existential Risk' framework, articulated across various interviews and writings collected at civilizationemerging.com and futurethinkers.org.

Douglas Hanahan and Robert Weinberg, 'Hallmarks of Cancer' (2000, 2011, 2022); Elinor Ostrom, *Governing the Commons* (Cambridge, 1990); Joseph Tainter, *The Collapse of Complex Societies* (Cambridge, 1988); Robin Hanson, 'The Great Filter' (1998); Liu Cixin, *The Dark Forest* (2008, English 2015); Nick Bostrom, *Superintelligence* (Oxford, 2014); Alfred Korzybski, *Science and Sanity: An Introduction to Non-Aristotelian Systems and General Semantics* (1933) for the map-territory distinction that frames the methodological humility this paper attempts to maintain.

This paper was developed by Andrzej Chudzinski in extended dialogue with Anthropic's Claude (Opus 4.7), which contributed substantial drafting and synthesis assistance. The conceptual core—the recognition that cooperative-substrate analysis at terrestrial scales supplies the mechanism layer for a calibrated Great Filter, with AI as candidate homeostatic regulator at the present civilizational moment—emerged from sustained collaborative dialogue. Readers should treat the framework with the skepticism appropriate to ideas developed in intensive conversation, however generative; the value of the framework will be determined by its survival under criticism from people who work seriously in each of the contributing traditions.

Preprint. Comments and disconfirmations welcome.