Recursive Continuation

Autopoiesis, Sympoiesis, and the Compression of Self-Modification

Flyxion

2026

Abstract

Discussions of recursive self-improvement (RSI) in artificial intelligence typically treat it as a threshold event: a future moment at which a sufficiently capable system begins inspecting and rewriting itself, triggering an accelerating spiral of capability. This essay argues that the underlying phenomenon — a system modifying the rule that governs its own future modification — is neither future nor rare. Autopoietic systems (cells, organisms) and sympoietic systems (ecosystems, languages, scientific communities, civilizations) have exhibited rule-level recursive modification for as long as life and culture have existed. What is misleading is the word *improvement*: autopoiesis and sympoiesis are not, in general, optimization processes. Their recursion is aimed at continuation — at preserving the conditions under which the system's activity remains admissible — and improvement, where it occurs, is one trajectory among many available to a continuation-preserving system, not the defining one. This reframing yields a formal hierarchy, Recursive Self-Improvement $\subset$ Admissible Recursive Modification $\subset$ Recursive Continuation Dynamics, and a small number of theorems governing when recursive modification remains viable: a complexity–repair relation bounding sustainable capability growth, a repair-dominance result showing that mature systems become repair-limited rather than capability-limited, and a compression principle locating the genuine novelty of AGI not in recursion itself but in the concentration of recursive modification into a single, low-diversity substrate. Artificial general intelligence is then reframed as one case, not the central case, of a general theory of recursive continuation under admissibility constraints.

Contents

Part I

Recursion Before Improvement

1. The Misleading Grammar of "Self-Improvement"

The contemporary discussion of artificial general intelligence (AGI) treats recursive self-improvement as a discrete, future threshold. The canonical version of the story runs roughly as follows: a system reaches a level of general capability sufficient to understand its own architecture; it uses that understanding to design a better version of itself; the better version is, among other things, better at the specific task of designing better versions of itself; and the loop compounds, each iteration shortening the time to the next, until capability grows far faster than any external process could track or correct. The event is typically presented as historically unprecedented — a departure from anything biological, cultural, or technological systems have previously done, and precisely for that reason treated as uniquely difficult to reason about, regulate, or prepare for.

This essay's central claim is the opposite. Recursive modification of the rule governing a system's own future behavior is not new; it is close to definitional of autopoietic and sympoietic organization, and has characterized cells, organisms, ecosystems, languages, scientific communities, and civilizations for as long as those systems have existed. A cell does not merely persist through time; it continuously rebuilds the membranes, replaces the proteins, and regulates the metabolic pathways that are themselves responsible for continuing to rebuild, replace, and regulate. A language does not sit still while its speakers use it; every act of speech is also, cumulatively, an act of quiet revision to the grammar future speakers will inherit. Neither the cell nor the speech community is doing anything resembling intelligence-explosion-style optimization, and yet both are engaged in exactly the structural operation — a system's own activity altering the rule by which its future activity proceeds — that the AGI literature treats as novel.

What *is* misleading in the standard framing is the word *improvement*, which smuggles in an optimization-oriented reading before any argument has been given for it. Autopoiesis, in Maturana and Varela's original sense, is a claim about organizational closure — a system continuously regenerating the components that regenerate it — with no built-in commitment to increasing capability. Sympoietic systems, a term due to Beth Dempster and developed further by Donna Haraway, more often evolve toward robustness, redundancy, and coexistence than toward any scalar notion of improvement: an ecosystem that becomes more resilient to disturbance has not thereby become more capable of anything in particular, and a language that develops a richer system of evidential marking has not thereby become a better language by any standard its speakers would recognize as improvement rather than mere change.

The thesis of this essay is therefore not that recursive self-improvement is old. It is that recursive self-improvement is a narrow, optimization-biased special case of a more general and older phenomenon: *recursive continuation*, in which a system's activity modifies the very rule by which it continues, without that modification being required to increase anything at all. The formal spine of the essay

develops this containment,

Recursive Self-Improvement $\subset$ Admissible Recursive Modification $\subset$ Recursive Continuation Dynamics,

and argues that what is genuinely novel about AGI is located in Part III of this essay — not in the presence of recursion, but in its compression into a single, low-diversity substrate. Parts I and II are devoted to earning the containment above properly: showing, first, that rule-level recursive modification is already everywhere once it is defined precisely (Sec. 2 onward), and second, that whether such modification is *improvement* in any interesting sense is a separate and much harder question than whether it occurs at all (Part II).

2. Recursive Modification

2.1 Three Levels of Adaptation

Discussions of self-modifying systems routinely conflate three distinct phenomena. Making the distinction precise is the essay's first formal task, because much of the popular urgency around RSI depends on treating sophisticated instances of the first two levels as though they were instances of the third. The conflation is understandable: from the outside, a system whose behavior is changing looks the same whether the change is happening at Level 1, 2, or 3. The levels differ not in how much behavior changes but in *what kind of object* is doing the changing.

Definition 2.1 (Levels of adaptation)**.** Let X be a state space, Θ a parameter space, and $\mathcal{F}$ a space of maps $X \to X$.

- **Level 1 (state update).** $x_{t+1} = F(x_t)$, for fixed F. The system moves through X under an unchanging rule.
- **Level 2 (parameter update).** $x_{t+1} = F_{\theta_t}(x_t)$, where $F : \Theta \times X \to X$ is fixed but $\theta_t \in \Theta$ varies. The system's behavior changes, but only within the range expressible by the fixed parametric family F.
- **Level 3 (rule update).** $F_{t+1} = G(F_t, x_t, E_t)$, where G acts on $\mathcal{F}$ itself, E_t an environment. The mapping governing future behavior is now the object being modified, not merely a parameter feeding into a fixed mapping.

Only Level 3 is recursive self-modification in the strong sense this essay is concerned with. A neural network updating weights by gradient descent is Level 2: the architecture and the update rule are fixed, and θ_t moves through Θ. Treating Level 2 flexibility as evidence of Level 3 dynamics is the single most common category error in popular discussions of RSI.

2.2 A Worked Example: Three Systems, Three Levels

The distinction is easiest to see by holding the domain fixed and varying only the level. Consider three systems that all "get better at playing chess."

A chess engine following a fixed evaluation function and fixed search algorithm, whose only per-game variation is the sequence of board states it visits, is a Level 1 system: F (evaluate position, search to depth d) never changes, only x_t does. A system such as AlphaZero's training process is Level 2: the network architecture and the reinforcement-learning update rule are fixed by the researchers in advance, and what changes over training is θ_t, the weight vector, moving through a fixed parametric family to better approximate a value function within that family's expressive range. No amount of additional training turns a fixed architecture into a different architecture. A neural architecture search process that evaluates candidate network topologies and selects better-performing structures, or a self-hosting compiler that rewrites the algorithm it uses to compile programs (including the next version of itself), is Level 3: the object being modified is F itself, not a coordinate within a fixed F. In the architecture-search case, G maps a topology together with performance data to a new topology, altering the very functional form future "training" will operate on; in the compiler case, G maps a compilation algorithm together with its own source to a revised compilation algorithm, which can now compile itself differently than before.

This example also makes the strictness of Theorem 2.2's inclusions concrete rather than merely existential: the AlphaZero case witnesses $\mathcal{P} \setminus \mathcal{S}$ directly (its parameter trajectory θ_t is manifestly non-constant, and no fixed F describes its behavior at the state level alone), and the self-hosting compiler witnesses $\mathcal{R} \setminus \mathcal{P}$ directly (no fixed parametric family F_θ describes a system whose compilation algorithm's functional form, not merely a numeric setting within it, is what changes between versions).

Theorem 2.2 (Adaptation Hierarchy)**.** *Let $\mathcal{S}$, $\mathcal{P}$, $\mathcal{R}$ denote the classes of dynamical systems expressible at Levels 1, 2, and 3 respectively. Then*

$$\mathcal{S} \subsetneq \mathcal{P} \subsetneq \mathcal{R}.$$

Proof sketch. $\mathcal{S} \subseteq \mathcal{P}$: any Level-1 system $x_{t+1} = F(x_t)$ is a Level-2 system with $\theta_t \equiv \theta_0$ constant. $\mathcal{P} \subseteq \mathcal{R}$: any Level-2 system is a Level-3 system in which G is restricted to updating only the θ-slot of a fixed functional form F_θ, i.e. $G(F_{\theta_t}, x_t, E_t) = F_{\theta_{t+1}}$ for some θ_{t+1} depending on (x_t, E_t), leaving the functional form itself untouched. Strictness of each inclusion is witnessed concretely in Sec. 2.2 above. □

The chess example is deliberately mundane. Nothing about neural architecture search or self-hosting compilers is exotic or futuristic; both exist today, are widely deployed, and are rarely described using the language of recursive self-improvement at all. This is itself evidence for the essay's thesis: Level-3 rule modification is not a threshold that has yet to be crossed. It is a known engineering technique, already in ordinary use, that only becomes alarming once combined with the further conditions this essay develops in Part III.

3. Recursive Continuation

Level-3 recursive modification, on its own, says nothing about whether the system persists. A rule can change and become worse; a system can self-modify into extinction. A corporation that recursively

restructures its own supply chain to cut costs, where each restructuring is itself informed by the cost savings of the last, is engaged in genuine Level-3 recursion — the rule generating future restructuring decisions is itself a product of prior restructuring — and can recursively optimize its way into a supply chain so brittle that a single disruption ends the company. Nothing about the recursion was insufficiently sophisticated; the modification simply was not required, by anything internal to the process, to preserve the conditions under which the company could keep operating. This section isolates the property that actually matters, prior to any question of improvement.

Definition 3.1 (Viability manifold and admissible continuation)**.** Let $\mathcal{A} \subset X$ be a *viability manifold*: the set of states under which the system's characteristic activity remains possible. A recursive process (x_t, F_t, E_t) is *continuation-preserving* (admissible) if

$$x_t \in \mathcal{A} \implies x_{t+1} = F_t(x_t) \in \mathcal{A} \qquad \text{for all } t.$$

Proposition 3.2 (Recursion does not imply continuation)**.** *There exist Level-3 recursive systems (x_t, F_t, E_t), with $F_{t+1} = G(F_t, x_t, E_t)$, that are not continuation-preserving for any nontrivial $\mathcal{A}$.*

Proof sketch. Immediate by construction: let G modify F_t so that F_{t+1} maps every state outside a shrinking neighborhood of a fixed point to a point outside $\mathcal{A}$. Rule-level recursion places no constraint on the image of F_{t+1} relative to $\mathcal{A}$ unless G is specifically constrained to respect it. The brittle-supply-chain corporation above is an informal instance: each successive F_t is a strictly more cost-efficient rule than its predecessor by the metric the company was optimizing, and the sequence nonetheless converges toward a state with no tolerance for perturbation, i.e. outside any reasonable $\mathcal{A}$ for continued operation. □

This is the essay's pivot. Recursive modification is a large, permissive class; recursive *continuation* is the much smaller, and much more interesting, subclass in which modification of the rule does not undermine the conditions that make future modification possible at all. Sections 4–6 argue that autopoiesis, sympoiesis, and evolution are best understood as three different mechanisms for achieving admissible recursive continuation, none of which presuppose or require improvement — and, unlike the corporation above, all three have been doing this successfully for a very long time, which is itself worth explaining.

4. Autopoiesis as Local Recursive Continuation

Maturana and Varela introduced autopoiesis to characterize the minimal organization distinguishing a living system from a merely persistent physical structure: not metabolism or reproduction as such, but the continuous production, by the system's own components, of the network of processes that produces those components. A cell membrane is not a fixed container within which chemistry happens; it is itself synthesized, maintained, and repaired by the metabolic processes it encloses, which in turn depend on the membrane to maintain the concentration gradients that make those processes possible. Neither the membrane nor the metabolism is primary; each is the ongoing condition of the other's

continued production. This closure, not any particular chemistry, is what autopoiesis names.

Definition 4.1 (Autopoietic system, formalized)**.** Let P_t denote the components of a system at time t and O_t its organization (the network of relations among components that determines which further components get produced). An autopoietic system satisfies

$$P_t \longrightarrow O_t, \qquad O_t \longrightarrow P_{t+1},$$

i.e. components produce the organization, and the organization governs production of the next generation of components.

Theorem 4.2 (Autopoietic Recursion)**.** *Any system satisfying Definition 4.1 is a Level-3 recursive system in the sense of Definition 2.1, with O_t playing the role of F_t.*

Proof sketch. Identify $x_t := P_t$ and $F_t := O_t$ (organization-as-production-rule: O_t maps current components to next-generation components). Since $P_t \to O_t \to P_{t+1}$, the organization at $t+1$ is determined by the components produced under O_t, i.e. $O_{t+1} = G(O_t, P_t)$ for some G — exactly the Level-3 schema $F_{t+1} = G(F_t, x_t, E_t)$ with E_t absorbing external perturbation. □

Remark 4.3. Nothing in Definition 4.1 or Theorem 4.2 involves a capability functional K. Autopoiesis is rule-level recursion aimed at reproducing O_t, i.e. at satisfying Definition 3.1 for whatever $\mathcal{A}$ characterizes the organism's continued organizational identity. A tree does not spend a growing season redesigning itself into a better tree; it spends the season reproducing the conditions under which it remains a tree. The immune system offers a sharper case: it does not merely fight pathogens, it continuously revises, at the level of somatic hypermutation and clonal selection, the very repertoire of antibody-producing rules by which it will fight future pathogens — a textbook Level-3 update, occurring constantly, in every healthy adult, with no relation whatsoever to anything resembling an intelligence explosion. Autopoiesis is recursive self-*maintenance*; self-improvement, where it occurs (immune learning, neural plasticity directed at improved future learning), is an additional property some autopoietic systems have, not a consequence of autopoiesis itself.

Remark 4.4 (When local recursion escapes $\mathcal{A}$). Autopoiesis is not automatically admissible; Definition 4.1 describes a structure, not a guarantee. Cancer is the clearest biological counterexample: a cell lineage whose Level-3 production rule O_t has itself been altered (by mutation, itself a form of G acting on O_t) so that $P_t \to O_t \to P_{t+1}$ continues to hold perfectly well *for the lineage*, while the lineage's continued production actively destroys the $\mathcal{A}$ of the organism that contains it. The lineage's local recursive continuation and the organism's are no longer the same viability manifold. This anticipates the relationship between capability and repair developed formally in Part II: unconstrained local proliferation is exactly the failure mode Sec. 10 treats in general terms.

5. Sympoiesis as Distributed Recursive Continuation

Autopoiesis locates recursive continuation inside a single bounded system. Many of the systems this essay is ultimately interested in — languages, markets, scientific communities, ecosystems — have

no such boundary, and no component that could be said to house the relevant O_t on its own. Beth Dempster coined *sympoiesis* for exactly this case: collectively-producing systems that lack self-defined boundaries, in which the very distinction between producer and product, or between system and environment, is continuously renegotiated by the activity of many interacting agents rather than settled in advance.

Definition 5.1 (Sympoietic system)**.** Let $a_1, \ldots, a_n$ be agents interacting through a shared medium E_t (an environment, language, market, or institutional substrate). Suppose

$$E_{t+1} = H(E_t, a_1^t, \ldots, a_n^t), \qquad a_i^{t+1} = J_i(a_i^t, E_{t+1}).$$

The medium is co-produced by the agents, and the agents are in turn shaped by the medium they co-produced.

A language is the clearest illustration. No speaker holds, or updates, a representation of the grammar; each act of speech is a Level-1 or at most Level-2 act from the speaker's point of view, an utterance produced according to a grammar the speaker did not consciously choose and mostly cannot articulate. Yet the aggregate of those utterances is precisely what determines which constructions become more frequent, which irregular forms get regularized, and which new constructions become grammatical for the next generation of speakers — a Level-3 modification of the grammar itself, occurring continuously, with no individual speaker as its locus. A currency market exhibits the identical structure: individual traders follow fixed or slowly-adapting strategies (Level 1 or 2), while the aggregate of their trades continuously reshapes the price-formation dynamics — the effective rule by which future prices are set — that every trader's strategy takes as given.

Theorem 5.2 (Distributed Recursion)**.** *A sympoietic system as in Definition 5.1 exhibits Level-3 recursive modification at the level of the collective system $(a_1, \ldots, a_n, E)$, even when each individual a_i is only a Level-1 or Level-2 system.*

Proof sketch. Define the collective continuation rule F_t^{coll} as the joint map $(a_1^t, \ldots, a_n^t, E_t) \mapsto (a_1^{t+1}, \ldots, a_n^{t+1}, E_{t+1})$ induced by H and $\{J_i\}$. Because E_{t+1} depends on the full agent profile at t, and each a_i^{t+1} depends on E_{t+1}, the effective transition rule for the collective state changes shape as the agent profile changes — i.e. $F_{t+1}^{\text{coll}} = G(F_t^{\text{coll}}, \cdot)$ for a nontrivial G — even where every individual J_i is a fixed, Level-1 or Level-2 map. □

Remark 5.3 (Composition elevates level). This is the essay's first genuinely non-obvious technical point: rule-level (Level-3) recursion can be an *emergent property of composition*, present in a collective system without being present in any individual component. Intuitively, no single agent needs to represent or reason about F_t^{coll} for the collective to be modifying it; the modification is carried entirely by the aggregate effect of many independently unremarkable individual updates. This is the formal basis for treating science, language, and markets as recursively self-modifying systems without requiring that any scientist, speaker, or trader personally rewrite their own cognitive architecture.

6. Evolution as Pre-Intelligent Recursive Continuation

If Level-3 recursion can be an emergent property of many interacting cognitive agents (Sec. 5), the natural next question is whether it requires cognition at all. Evolution by natural selection, together with the more specific phenomenon of niche construction, shows that it does not.

Definition 6.1 (Niche-constructing evolutionary loop)**.** Let G_t be a genotype distribution, P_t the resulting phenotype distribution, and E_t the selection environment. Suppose

$$G_t \to P_t, \qquad P_t \to E_{t+1} \quad \text{(niche construction)}, \qquad E_{t+1} \to G_{t+1} \quad \text{(selection)}.$$

Beavers provide the standard illustration of niche construction: the phenotype (dam-building behavior) alters the environment (converting flowing streams into ponds), and the altered environment changes the selection pressures acting on the beaver population's future genotypes, favoring further adaptations to still-water rather than stream habitats. Earthworms alter soil chemistry and structure in ways that change which plant and microbial genotypes are favored in the same location for generations afterward. Human dairy farming, a purely cultural and phenotypic innovation, altered the selection environment for the human genome itself sufficiently to drive the spread of lactase-persistence alleles in populations with a sufficiently long history of pastoralism — culture recursively modifying the genetic rule governing future populations, with no genotype anywhere containing a representation of that fact.

Theorem 6.2 (Evolutionary Recursion)**.** *The loop of Definition 6.1 is a Level-3 recursive system in which the selection rule mapping populations to next-generation populations is itself modified by the population's own prior activity (via niche construction), without requiring foresight, intention, or intelligence in any individual organism.*

Proof sketch. Let F_t be the effective selection map $G_t \mapsto G_{t+1}$ induced by composing phenotype expression, niche construction, and selection. Since E_{t+1} (and hence the fitness landscape defining F_t) depends on P_t, which depends on G_t, the selection map itself is a function of the population's history: $F_{t+1} = G(F_t, G_t)$, the Level-3 schema, with no agent in the loop possessing a representation of F_t at all. □

Remark 6.3. Evolution is the cleanest available demonstration that Level-3 recursion requires neither intelligence nor intention. This matters for the essay's overall argument: if rule-level recursive modification is achievable by a process with no cognition whatsoever, then citing an AI system's cognitive sophistication cannot be what makes its recursive self-modification distinctively dangerous. Whatever is genuinely new about AGI (Part III) has to be found elsewhere.

7. Recursive Continuation Without Identity

Every mechanism examined so far preserves something recognizable across the recursion: a cell preserves its organization, a species preserves a genetic lineage, an institution preserves a charter or a

name. Some of the clearest cases of recursive continuation preserve none of these. A shipbuilding tradition can continue for centuries while every named component — the specific hull design, the materials, the tools, the guild structure teaching the craft — changes beyond recognition; what continues is not any object handed down but the *practice of building ships in a way informed by how ships were previously built*. A language continues while every word, every grammatical construction, and every speaker eventually changes completely; nothing that existed in the language a thousand years ago need exist in it now for it to be, uncontroversially, a continuation of it rather than a replacement. A scientific field can continue while its central theories are overturned, its instruments replaced, and its founding practitioners forgotten by name.

This suggests that Definition 3.1's viability manifold $\mathcal{A}$ has been doing double duty, and it is worth separating the two jobs explicitly.

Definition 7.1 (Object versus process admissibility)**.** Let $\mathcal{A}_{\text{obj}}$ denote a viability condition on the persistence of some designated object or component across the recursion (a cell's organization, a species' gene pool, an institution's charter). Let $\mathcal{A}_{\text{proc}}$ denote a viability condition on the persistence of the *recursive process itself* — that each F_{t+1} remains recognizably a modification of F_t by the system's own characteristic activity, whether or not any object present at F_t survives into F_{t+1}.

Autopoiesis (Sec. 4) is closest to a pure $\mathcal{A}_{\text{obj}}$ case: organizational closure is precisely a claim about a persisting object (the organization O_t) being reproduced. Language change and craft traditions are closer to a pure $\mathcal{A}_{\text{proc}}$ case: $\mathcal{A}_{\text{obj}}$ for any particular word, tool, or technique is regularly violated — most specific words, tools, and techniques do not survive — while $\mathcal{A}_{\text{proc}}$ holds continuously, because each stage's modification is still recognizably produced by applying the tradition's own characteristic activity to its own prior stage, rather than by an unrelated process starting fresh. $\mathcal{A}_{\text{obj}} \neq \mathcal{A}_{\text{proc}}$ in general: a system can satisfy one while violating the other, and most real cases of sympoiesis (Sec. 5) sit somewhere between the two poles, preserving some objects (a language's core grammar changes far more slowly than its vocabulary) while continuously replacing others.

Remark 7.2. This refinement matters for how Part II's results should be read. The Complexity–Repair Theorem and Repair Dominance corollary were stated in terms of a single viability manifold $\mathcal{A}$; a $\mathcal{A}_{\text{proc}}$-type system can appear to be failing badly by an $\mathcal{A}_{\text{obj}}$-type measure — most of its components are, at any given moment, in the process of being discarded — while remaining perfectly admissible by the standard that actually governs its continuation. A shipbuilding tradition that stopped changing its designs and materials entirely would not thereby become *more* admissible as a tradition; freezing the object can be precisely what destroys the process. Which of the two conditions is the one that matters is a modeling choice this essay's formalism does not make automatically, and getting it backwards — treating a healthy $\mathcal{A}_{\text{proc}}$-system's object turnover as evidence of failing viability — is a readily available mistake the distinction above is meant to prevent.

Part II

Repair and Admissibility

8. Repair Requires Recoverable History

Every repair mechanism examined so far — DNA proofreading, peer review, a software rollback, an audited institution — has silently presupposed something this essay has not yet stated as a requirement: a repair map ρ cannot correct a state toward $\mathcal{A}$ unless something persists, somewhere, that records what an admissible state of that kind looks like. A cell's repair machinery does not decide de novo what a healthy protein structure is; it checks candidate structures against a template encoded in DNA, itself a persisting record of prior admissible configurations. Peer review does not evaluate a claim against nothing; it evaluates it against an accumulated, citable record of prior findings the claim must be consistent with or must explicitly overturn. A version-control rollback is only possible because prior commits were retained rather than overwritten.

Remark 8.1 (The dependency made explicit). Let H_t denote whatever record of prior states — genomic, institutional, archival, or literal version history — a system retains at time t. The repair map ρ of Corollary 12.2 is not, in general, a function of the damaged state alone; it is a function $\rho(x_t^{\text{cand}}, H_t)$ of the candidate together with whatever historical record is available to judge it against. Where H_t is impoverished or absent, ρ degrades toward guessing: a system with no record of its own prior admissible states cannot distinguish a damaged configuration it should correct from a novel configuration it should accept, because it has no reference to compare either against. This is the sense in which repair requires recoverable history rather than merely requiring computational effort: the effort has nothing to operate on without a persisting trace of what came before.

This gives Part I's history-preservation mechanisms — DNA (Sec. 4), archives and citation networks (Sec. 5), legal precedent (Sec. 6's civilizational cousin, developed further in Sec. 17) — a role in Part II's apparatus that has so far only been implicit: they are not simply evidence that these systems value the past, they are the substrate H_t without which the ρ this essay's admissibility results depend on could not function at all. A civilization's libraries, legal archives, accounting systems, and standards documents are, on this reading, best understood as *distributed memory* in exactly the sense this section requires — the persisting H_t that lets civilizational repair (Sec. 17.1) operate on something rather than on nothing. This essay does not develop H_t's own dynamics — how historical records themselves degrade, get selectively retained, or get reconstructed from partial traces is a substantial further theory this author has developed elsewhere under a different name, and is not re-derived here — but the dependency is worth stating as a standing requirement of everything Part II builds on top of it, rather than left as an unexamined background assumption.

9. Improvement as a Special Case

Part I established that rule-level recursive modification is old and common. It did not establish that such modification is generally beneficial to the systems undergoing it, and Remark 4.4 already flagged a case — a cell lineage whose local recursion actively destroys the organism containing it — where it plainly is not. This section makes the resulting distinction precise: between a recursive process that merely continues, one that continues while also becoming more capable by some measure, and the much larger space of recursive processes that neither.

Definition 9.1 (Admissible recursive modification and self-improvement)**.** Let $K : X \to \mathbb{R}$ be a capability functional. A recursive process is:

- **Admissible** if it satisfies Definition 3.1 relative to $\mathcal{A}$.
- **Self-improving** if, in addition, $K(x_{t+1}) > K(x_t)$ for all (or almost all) t.

Proposition 9.2 (Strict containment)**.**

$$\{\textit{self-improving trajectories}\} \subsetneq \{\textit{admissible trajectories}\} \subsetneq \{\textit{all Level-3 trajectories}\}.$$

Proof sketch. Right inclusion is Proposition 3.2. Left inclusion: a stable homeostatic system with $K(x_{t+1}) = K(x_t)$ for all t — an adult organism maintaining constant metabolic function, neither growing nor declining — is admissible without being self-improving, witnessing strictness in one direction. The cancerous cell lineage of Remark 4.4 witnesses the other: local capability K (proliferation rate) strictly increases at every step, satisfying the letter of "self-improving," while the lineage's own trajectory exits the organism's $\mathcal{A}$ in finite time. Self-improvement, as defined, is therefore neither necessary nor sufficient for admissibility unless the definition explicitly nests it inside admissibility as a further condition — which is exactly what Definition 9.1 does, and exactly what the informal notion of "recursive self-improvement" in the AGI literature typically fails to do, treating $K(x_{t+1}) > K(x_t)$ as though it were the whole of the relevant claim. □

This gives the essay's governing hierarchy its final, precise form:

Recursive Self-Improvement $\subset$ Admissible Recursive Modification $\subset$ Recursive Continuation Dynamics $\subset$ Recur

Autopoiesis and sympoiesis (Secs. 4–6) populate the second and third of these classes far more often than the first: a tree, a language, and an ecosystem are, for most of their history, admissible without being especially self-improving by any capability metric worth naming. The remainder of Part II asks what determines whether a given trajectory of capability growth — the narrower, more demanding case — stays inside the admissible class at all, rather than repeating the cancer lineage's mistake of confusing local capability increase with genuine continuation.

10. Capability Is Not Free

Every domain in which capability accumulates also accumulates a cost of keeping that capability working. A software system that grows in feature count must also track a growing web of interactions between features, any of which can silently break when another part of the system changes; the discipline of software engineering has a standard name for the resulting burden, *technical debt*, precisely because it behaves like a debt — it does not go away on its own, and it compounds. A biological organism that grows in size and metabolic complexity must devote a growing share of its resources to DNA repair, protein quality control, and immune surveillance simply to keep its existing capability from degrading. Neither case is a story about insufficient capability. Both are stories about a maintenance burden that scales with capability and has to be paid out of some separate budget.

10.1 Reachability as Capability

Before capability can be put to formal use, it needs to be said what licenses treating intelligence, territorial extent, code complexity, and metabolic sophistication as instances of one quantity K rather than four unrelated ones dressed in the same letter. Leaving this unstated is the single largest source of ambiguity in what follows, and it is worth resolving with an actual definition rather than a passing remark.

Definition 10.1 (Capability as reachability)**.** Let $\mathcal{R}(x) \subseteq X$ denote the set of states reachable from x under the system's available dynamics within some fixed horizon. Define

$$K(x) \;=\; \log \mathrm{Vol}\,\big(\mathcal{R}(x)\big),$$

the log-volume of the reachable-state set, under whatever measure on X is natural to the domain (a counting measure for discrete state spaces, a volume form for continuous ones).

On this reading, a more capable organism is one that can occupy more metabolic and behavioral states; a more capable software system is one that can be driven through more distinct execution paths; a more extensive territory is one that admits more distinct administrative configurations; a more knowledgeable field is one whose accumulated results open more distinct lines of further inquiry. $C(K)$ — the maintenance burden of Definition 10.3 below — is then naturally read as the cost of keeping a state space of that size navigable at all: distinguishing which of the many reachable states the system is actually in, verifying that transitions between them still function as intended, and preventing the sheer combinatorial size of $\mathcal{R}(x)$ from becoming, itself, a source of untracked and potentially $\mathcal{A}$-violating drift.

Remark 10.2 (Scope of the identification). Definition 10.1 is offered as this essay's working answer to what K measures, not as a claim that every capability functional used informally throughout Parts I–IV has been shown to reduce to it. Territorial extent and scientific knowledge are the least immediately obvious cases: a territory's reachable-state volume would need to be read as the space of distinct administrative and economic configurations it can be organized into, not its physical area, and a field's reachable-state volume as the space of distinct further results its accumulated methods and findings

open up, not the raw count of its past publications. Both readings are defensible but neither has been derived here from anything more basic; where the identification is not immediate (as in Sec. 18's use of K for AI capability), Definition 10.1 should be read as a modeling hypothesis this essay adopts for uniformity, not as an established equivalence the essay has proved.

Definition 10.3 (Maintenance burden and repair capacity)**.** Let $K_t = K(x_t)$ be capability at time t per Definition 10.1, $C : \mathbb{R} \to \mathbb{R}_{\geq 0}$ the maintenance burden imposed by a given capability level (with $C' > 0$), and $R_t \geq 0$ the repair capacity available to the system at t. Model viability by the difference equation

$$V_{t+1} = V_t - C(K_t) + R_t.$$

Theorem 10.4 (Complexity–Repair Theorem)**.** *If there exists T such that $C(K_t) > R_t$ for all $t > T$, then $V_t \to -\infty$, i.e. the system eventually leaves any bounded viability threshold.*

Proof. Immediate from Definition 10.3: for $t > T$, $V_{t+1} - V_t = R_t - C(K_t) < 0$ by hypothesis, so V_t is eventually strictly decreasing without bound below. □

Corollary 10.5 (Sustainability condition)**.** *A capability trajectory K_t is sustainable only if $R_t \geq C(K_t)$ for all sufficiently large t.*

Proposition 10.6 (Robustness beyond the linear model)**.** *Let $V_{t+1} = \Psi(V_t, K_t, R_t)$ for any Ψ that is non-decreasing in R_t, non-increasing in $C(K_t)$, and satisfies $\Psi(V, K, R) \leq V - c(K) + R$ for some baseline burden function c with $c' > 0$. Then the conclusion of Theorem 10.4 still holds — $V_t \to -\infty$ whenever $R_t < c(K_t)$ persistently — with c in place of C.*

Proof sketch. The domination hypothesis gives $V_{t+1} \leq V_t - c(K_t) + R_t$ directly at each step, for the system's actual realized values. Whenever $R_t < c(K_t)$, this forces $V_{t+1} < V_t$; under the same persistent-shortfall hypothesis as Theorem 10.4 (applied now to c rather than C), the same argument gives $V_t \to -\infty$, since V_t is bounded above by the linear comparison sequence L_t defined by $L_{t+1} = L_t - c(K_t) + R_t$, $L_0 = V_0$, which diverges to $-\infty$ by Theorem 10.4 itself. □

Remark 10.7. Proposition 10.6 is the essay's answer to the objection that Definition 10.3's linear difference equation is an artificial modeling choice. It is a modeling choice, but the qualitative conclusion — sustained excess of maintenance burden over repair capacity is fatal to viability — does not depend on linearity, only on the much weaker requirement that Ψ respond monotonically to its two inputs and be no better than some baseline convex-in-K cost. Nonlinear, stochastic, or domain-specific versions of Ψ (a logistic viability model, a stochastic difference equation with a repair-capacity-dependent survival probability) inherit the same divergence result by the same comparison argument, provided the monotonicity and domination conditions hold.

10.2 Repair as Information Production

Repair capacity R_t has so far been treated as a resource the system simply spends, on a par with any other budget item. This undersells what repair actually requires. To repair anything, a system

must first be able to tell the difference between an intact and a damaged component — DNA repair machinery must distinguish correct from incorrect base pairing before it can act, a software test suite must distinguish passing from failing behavior before a fix is meaningful, an institution's audit process must distinguish compliant from non-compliant conduct before correction is possible. Repair, in other words, presupposes a prior act of distinction, and distinction production is not free.

Remark 10.8 (Repair's diagnostic dependency). Let I_t denote the diagnostic information available to a system at time t — the resolution at which it can distinguish intact from damaged states within its own component space. Repair capacity is bounded by diagnostic information rather than independent of it: R_t cannot exceed what I_t makes distinguishable, since a system cannot correct a fault it has no means of detecting. Informally, $R_t \lesssim g(I_t)$ for some non-decreasing g. This has an immediate and slightly uncomfortable consequence for Sec. 10: because I_t itself has to be produced — sensors built, tests written, audits conducted — producing diagnostic information is itself an activity with a cost, and that cost is plausibly part of $C(K_t)$ rather than external to it. A system that grows in capability without growing its diagnostic resolution is not simply failing to invest in repair; it may be losing the ability to even perceive that repair is becoming necessary, which is a strictly worse position than knowingly under-repairing. This essay does not develop a full account of I_t's dynamics — that would require formal apparatus (a theory of distinguishability) this essay has not built — but the dependency is worth flagging explicitly rather than leaving R_t looking like an unconditional resource a system can simply choose to allocate more of.

10.3 Repair Entropy

Diagnostic information's role can be sharpened slightly further. A set of observed symptoms is rarely compatible with only one possible fault; it is compatible with some set $\mathcal{H}$ of candidate fault histories, any of which could have produced what was observed. Define

$$S_R \;=\; \log |\mathcal{H}|,$$

the *repair entropy* of a diagnostic situation: the log-size of the set of fault histories consistent with available symptoms, in the same spirit Shannon entropy measures the log-size of a set of messages consistent with a received, noisy signal. Low S_R means the symptoms narrow the space of possible faults to a small set — a stack trace pointing at one function is low-entropy diagnosis — while high S_R means many distinct, mutually incompatible fault histories remain consistent with what has been observed — an intermittent failure with no reproducible trigger is high-entropy diagnosis, and repair effort under high S_R is disproportionately effort spent narrowing $\mathcal{H}$ rather than effort spent correcting anything.

Remark 10.9. S_R gives Remark 10.8's bound a sharper form: I_t is, in effect, whatever evidence reduces $|\mathcal{H}|$, and $R_t \lesssim g(I_t)$ can be restated as repair capacity being bounded by how far diagnostic evidence has driven S_R toward zero rather than by raw effort expended. Two systems with identical R_t in the crude budgetary sense can differ enormously in effective repair capacity if one operates at low S_R (its diagnostics reliably narrow $\mathcal{H}$ to a small set) and the other at high S_R (its diagnostics leave the fault space nearly as large as before any diagnostic effort was spent at all).

Proposition 10.10 (Repair–Distinction Dependency)**.** *Let g be the non-decreasing bound of Remark 10.8, restated via repair entropy as $R_t \lesssim g(-S_R)$ (repair capacity bounded by a non-decreasing function of how far diagnostic evidence has driven S_R toward zero, equivalently a non-increasing function of S_R itself). If the distinguishability between intact and damaged states decreases — formally, if the set $\mathcal{H}$ of fault histories consistent with observed symptoms grows, so that S_R increases — then the upper bound on effective repair capacity R_t is non-increasing.*

Proof. Immediate from the definitions: g is non-decreasing in $-S_R$ by construction (equivalently non-increasing in S_R), so S_R increasing forces $g(-S_R)$ non-increasing, and $R_t \lesssim g(-S_R)$ carries the bound down with it. □

Remark 10.11. This is a deliberately modest formalization of a claim this essay had, until now, only discussed informally: repair depends on distinction, and distinction loss degrades repair. The proposition does not derive S_R's own dynamics — what causes distinguishability to erode in a given system is a domain-specific question this essay does not answer — it only makes precise what follows *once* distinguishability has eroded, given the diagnostic-dependency structure already established in Remark 10.8. A fuller treatment connecting this essay's admissibility apparatus to a general theory of distinction and its preservation is one of this essay's open problems (see the closing section), not a claim resolved here.

10.4 Failure Modes of Repair

Nothing so far has considered the possibility that repair itself goes wrong. Corollary 10.5 treats $R_t \geq C(K_t)$ as the condition to satisfy, implicitly assuming that repair capacity, once allocated, functions as intended. Four distinct failure modes are worth separating, because they have different signatures and different remedies.

- **Under-repair.** R_t is simply insufficient relative to $C(K_t)$ — the case Theorem 10.4 already covers.

- **Over-repair.** Repair activity itself becomes a source of burden rather than relief: a software team that runs so many defensive checks and reviews that little capacity remains for anything else, or an immune system whose response to a minor irritant causes more tissue damage than the irritant would have. Over-repair is not covered by Theorem 10.4 as stated, since the theorem treats R_t as pure offset; a fuller model would let R_t itself impose a burden $C_R(R_t)$, with autoimmune disease as the biological limiting case where C_R dominates the benefit R_t was meant to provide.

- **Misrepair.** The diagnostic step of Sec. 10.2 misfires: a fault is misidentified, and correction is applied to the wrong component, or a correct component is altered on the mistaken belief that it was faulty. A software hotfix that patches a symptom while leaving the underlying defect in place, and then requires a cascade of further hotfixes to patch the side effects of the first, is a case where each repair step appears locally successful while $\mathcal{A}$ continues to erode.

- **Repair delay.** Diagnostic information I_t correctly identifies a fault, and repair capacity R_t is sufficient, but the interval between detection and correction is long enough for the fault to compound — institutional overcorrection often has this shape, where a problem is identified promptly but the corrective response is delayed by process until the original problem has changed character, and the eventual correction addresses a version of the problem that no longer exists.

Remark 10.12. None of these four failure modes are visible in the single-variable model of Definition 10.3, where R_t appears only as a scalar offset. They become visible only once repair is treated as a process with its own diagnostic prerequisites (Sec. 10.2) and its own possible costs and timing, which is offered here as a direction for extending the model rather than as a claim this essay's formal results already cover.

11. The Repair Dominance Theorem

Corollary 10.5 says only that repair must keep pace with maintenance burden. It does not yet say which of the two variables — increasing capability, or increasing repair capacity — is the better use of a marginal unit of whatever resource the system has to spend. The next result answers that question, and the answer changes as capability grows.

Theorem 11.1 (Repair Dominance)**.** *Under the dynamics of Definition 10.3, suppose C is convex with $C'(0) < 1$ and $C'(K) \to \infty$ as $K \to \infty$. Then there exists a threshold K^* (defined by $C'(K^*) = 1$) such that for all $K > K^*$,*

$$\left|\frac{\partial V_{t+1}}{\partial K_t}\right| \;=\; C'(K_t) \;>\; 1 \;=\; \frac{\partial V_{t+1}}{\partial R_t}.$$

Proof. From $V_{t+1} = V_t - C(K_t) + R_t$, direct differentiation gives $\partial V_{t+1}/\partial R_t = 1$ and $\partial V_{t+1}/\partial K_t = -C'(K_t)$. By convexity of C and $C'(0) < 1 < \lim_{K\to\infty} C'(K)$, the intermediate value theorem gives a unique K^* with $C'(K^*) = 1$, and C' increasing thereafter gives $C'(K) > 1$ for all $K > K^*$. □

Remark 11.2. Past K^*, a marginal unit of repair capacity buys more viability than a marginal unit of capability. This is the formal content behind the informal observation that aging organisms, aging software systems, and aging institutions become repair-limited rather than capability-limited: the theorem shows this is not an empirical curiosity but a structural consequence of any convex maintenance-cost model, independent of the specific domain.

11.1 Three Domains Past K^*

The convexity assumption on C is doing real work, and it is worth checking that it is not an artifact of the model but an observed feature of the systems the theorem is meant to describe.

Senescence. Cellular damage — DNA lesions, misfolded proteins, dysfunctional mitochondria — accumulates as a roughly fixed byproduct of ordinary metabolic activity, but the resources available for DNA repair, autophagy, and proteostasis do not scale up to match a growing organism's growing metabolic rate; if anything, repair-pathway efficiency itself declines with age, which is a second,

compounding mechanism this section's single-variable model does not even need to invoke to make its point. An organism well past developmental maturity is a system for which K (sustained metabolic and functional capacity) is no longer profitably increased by any available means, and for which the marginal return on repair investment (caloric restriction's effect on lifespan being one of the better-studied empirical instances) is comparatively large.

Software. Technical debt is convex almost by definition: adding the n-th feature to a codebase risks interacting with any of the $n-1$ existing features, so the surface area of possible unintended interaction grows combinatorially, not linearly, in the size of the system, even though each individual feature took roughly constant effort to add. Mature software organizations recognize K^* operationally — it is the point at which teams shift from being feature-constrained to being refactoring-constrained, and organizations that fail to make this shift accumulate exactly the divergent-V_t trajectory Theorem 10.4 describes, typically labeled a systemic rewrite becoming unavoidable.

Empires. Administrative and logistical overhead in pre-modern states scaled worse than linearly with territorial extent, since the cost of maintaining communication, tax collection, and military presence across a periphery grows with both the number of provinces and the interactions among them; historians of imperial decline have long treated a growing gap between administrative capacity and territorial extent as a leading indicator of collapse, independently of any theory of recursive systems. The same K^* structure describes a very different kind of system: capability (territorial extent) growing past the point at which repair capacity (administrative capacity) can keep pace with the convex maintenance burden that extent imposes.

These three cases are not offered as proof that the linear-C model is correct in any of them, only as evidence that the qualitative shape the model predicts — a marginal-return crossover past which repair dominates — is not an artifact of the formalism but a pattern independently visible wherever complexity accumulates a maintenance burden.

12. Admissibility as Viable Recursion

Theorem 12.1 (Admissibility Theorem)**.** *A rule-update sequence G (Definition 2.1, Level 3) is a viable continuation strategy relative to $\mathcal{A}$ only if, for all t, the induced capability and repair trajectories satisfy the sustainability condition of Corollary 10.5. The Complexity–Repair relation is a sufficient mechanism by which admissibility (Definition 3.1) can fail; it is not offered as the only such mechanism.*

Proof sketch. If $R_t < C(K_t)$ persistently, Theorem 10.4 gives $V_t \to -\infty$, which for any viability manifold $\mathcal{A}$ defined by V_t remaining above some threshold entails $x_t \notin \mathcal{A}$ for large t, violating Definition 3.1. The qualifier "sufficient, not necessary" is required because $\mathcal{A}$ can be exited by mechanisms independent of the (K, C, R) model — e.g. a discrete external shock — that this section's apparatus does not model. □

Corollary 12.2 (Repair Precedes Continuation)**.** *Let $x_t^{\text{cand}} = F_t(x_t)$ denote the raw candidate produced*

by applying the current rule, prior to any check against $\mathcal{A}$. Define the validated predecessor

$$x_t^+ = \begin{cases} x_t^{\text{cand}} & \text{if } x_t^{\text{cand}} \in \mathcal{A}, \\ \rho(x_t^{\text{cand}}) & \text{otherwise, for some repair map } \rho \text{ with } \rho(x_t^{\text{cand}}) \in \mathcal{A} \text{ when repair succeeds.} \end{cases}$$

Only x_t^+, not x_t^{cand}, may serve as x_{t+1} in the next iteration of the recursion.

Proof sketch. Immediate from Definition 3.1: the admissibility condition is a condition on x_{t+1}, not on any intermediate candidate the system happens to generate en route to it. Setting $x_{t+1} := x_t^{\text{cand}}$ without the case split risks violating $x_{t+1} \in \mathcal{A}$ whenever ρ would have been needed; setting $x_{t+1} := x_t^+$ enforces the condition by construction. □

Remark 12.3. This is a small point with a disproportionate consequence for how the recursion in Definition 2.1 should actually be implemented, and it sharpens something Part II has been assuming rather than stating: a rule-update sequence is not simply an alternation between applying F_t and checking membership in $\mathcal{A}$ after the fact. The object that gets fed back into the next iteration is the *validated* predecessor x_t^+, and the recursion is honest only if x_t^{cand} is never allowed to silently stand in for it. This is the same structure this author's companion work on telemetry and boundary repair names $\Gamma \to \Gamma^* \to \Gamma^+$ (raw fragments, repaired candidate, admitted history) — arrived at independently, for a different kind of system, but resolving to the identical requirement: repair and admission are not optional bookkeeping performed on a history that continues regardless, they are the gate that determines what the next round of continuation is even allowed to continue *from*.

12.1 Boundary Repair and Admissibility Recovery

Corollary 12.2 introduced a repair map ρ without asking what happens when ρ is actually needed — that is, what happens once a trajectory has already left $\mathcal{A}$, rather than merely produced a single candidate step that would have left it. The question is not idle: Theorem 10.4 describes divergence as an eventual, asymptotic fate under persistent shortfall, but real trajectories can exit $\mathcal{A}$ at a finite time and, depending on what ρ can do, either recover or not.

Definition 12.4 (Repaired continuation)**.** A trajectory (x_t) exhibits *admissible continuation* on an interval if $x_t \in \mathcal{A}$ throughout. It exhibits *repaired continuation* at t if $x_t \notin \mathcal{A}$ but $\rho(x_t) \in \mathcal{A}$ for the repair map of Corollary 12.2, so that continuation resumes from $\rho(x_t)$ rather than from x_t itself.

Whether repaired continuation is available at all depends on properties of ρ that this essay has not needed to specify until now. If ρ is defined on all of X and $\rho(x) \in \mathcal{A}$ for every x, recovery is always available and $\mathcal{A}$-exit is never actually fatal, only temporarily costly — a strong and generally unrealistic assumption. More plausibly, ρ has a limited domain of effectiveness: a repair map can return a mildly damaged state to $\mathcal{A}$ but fails, or is simply undefined, once damage exceeds some threshold. This motivates distinguishing the *recoverable region* $\mathcal{A}^+ = \{x \notin \mathcal{A} : \rho(x) \in \mathcal{A}\}$ from the remainder $X \setminus (\mathcal{A} \cup \mathcal{A}^+)$, states from which no continuation, repaired or otherwise, is available. Cellular apoptosis is the biological instance of a system detecting that it occupies this remainder and halting rather than

continuing to execute a rule-update that repair cannot rescue; software systems that fail closed rather than continuing to run on corrupted state are the engineered analogue.

Remark 12.5. This refinement does not change any theorem already proved; Theorem 10.4 and Theorem 11.1 are unaffected by whether exit from $\mathcal{A}$ is instantaneous or recoverable, since both describe the asymptotic consequence of ρ being persistently insufficient rather than the short-run dynamics of any single excursion. What Definition 12.4 adds is a vocabulary for a question the rest of the essay had been answering only implicitly: a system that exits $\mathcal{A}$ and recovers via ρ has not thereby shown that its recursion is admissible in the sense Definition 3.1 originally specified — that definition describes trajectories that never leave $\mathcal{A}$ at all. A system that survives only by repeated recovery through $\mathcal{A}^+$ is admissible in a strictly weaker, recovery-dependent sense, and whether that weaker sense is good enough depends on costs (time, resources, accumulated damage during the excursion) this section's formalism does not yet price.

This closes Part II's argument: admissibility, not capability, is the governing property, and Secs. 10–11 identify one general, structural reason (convex maintenance cost) why capability-maximizing trajectories tend to be exactly the trajectories that risk leaving $\mathcal{A}$.

It is worth returning briefly to Part I in light of this result, because the systems examined there turn out, with hindsight, to already be organized around exactly this constraint. Autopoiesis (Sec. 4) devotes the overwhelming majority of a cell's activity to R_t-type maintenance — membrane repair, protein turnover, metabolic regulation — rather than to increasing any capability functional, which by Theorem 11.1 is exactly the correct allocation for a system that is not near the low-K region where capability investment still dominates. Sympoietic systems (Sec. 5) go further: science's peer review and replication, civilization's courts and archives, and a market's clearinghouses and audits are, functionally, distributed repair capacity R_t maintained by institutions rather than by any single agent, which is precisely what allows the collective's K (accumulated knowledge, accumulated capital, accumulated law) to keep growing without the collective as a whole drifting outside $\mathcal{A}$. Evolution (Sec. 6) selects directly on the joint trajectory, since a lineage whose repair mechanisms (DNA proofreading, immune function, wound healing) fail to keep pace with its own metabolic or morphological complexity is a lineage that does not survive to reproduce — natural selection is, among other things, a filter that removes exactly the trajectories Theorem 10.4 predicts will diverge.

None of this is coincidence, and none of it required intelligence to arise: it is what remains after billions of years, or centuries, of variation and selection have removed the alternative — the systems visible today are disproportionately the ones whose recursive modification happened to respect Corollary 10.5, whether or not anything inside them ever represented that constraint explicitly. Part III asks what happens to this picture once recursive modification is moved into a substrate that has not been filtered by any comparable process, and where the distributed repair architecture that Sec. 5 showed to be doing much of the real work is no longer guaranteed to be present at all.

Part III

Compression

13. The Compression Principle

Part II identified admissibility, not capability, as the property that determines whether recursive modification is sustainable. It said nothing about *when* an admissible-looking trajectory is actually at risk of quietly drifting out of $\mathcal{A}$ before anyone notices. This section develops the essay's account of that risk, and — unlike the informal treatments this idea has received elsewhere in this author's notes — attempts to derive rather than merely assert it, under a stated and clearly flagged model.

Remark 13.1 (Which $\mathcal{A}$ is at risk?). Section 7 distinguished $\mathcal{A}_{\text{obj}}$ from $\mathcal{A}_{\text{proc}}$, and the distinction does not become irrelevant simply because Part III's variables (n, Δt, D_R) were introduced without mentioning it. A compressed system failing by an $\mathcal{A}_{\text{obj}}$ standard and one failing by an $\mathcal{A}_{\text{proc}}$ standard are different events with different signatures: object-failure looks like a specific component breaking (a corrupted weight, a lost record), while process-failure looks like the recursive modification itself ceasing to be recognizably produced by the system's own characteristic activity, even if every individual component along the way looks fine. The risk apparatus developed below is agnostic between the two — $\mathcal{A}$ appears as an unanalyzed viability manifold throughout — and applying it to a real system requires deciding, first, which of the two failure types is actually the one being guarded against, since the answer changes what counts as a locus, what counts as a repair channel, and what counts as recovery.

Definition 13.2 (Iteration period and locus count)**.** Let n denote the number of independent, only weakly-coupled loci at which Level-3 recursive modification occurs, and let Δt denote the characteristic period between successive rule-updates $F_t \to F_{t+1}$ at a given locus. Let τ denote the characteristic period on which an external, independent verification or correction process can detect and respond to a rule error.

Remark 13.3 (Resolving an ambiguity). Earlier informal treatments of this idea describe Δt inconsistently, sometimes as "repair latency" (where small Δt is safe) and sometimes as "iteration speed" (where small Δt is dangerous). Definition 13.2 resolves this: Δt is the period of the *rule-update* loop, not of repair. The quantity that matters is the ratio $N^* = \tau/\Delta t$: the number of unmonitored rule-iterations that can occur within one external-verification window. Danger scales with N^*, not with Δt alone.

13.1 What Counts as a Locus?

Definition 13.2 treats n as given, and the entire apparatus of this and the following section presupposes that a system's recursive modification can be partitioned into countable, individuable loci. This is a substantive assumption, not a bookkeeping detail, and it is worth asking directly what licenses it. Is a research lab a locus, or is each scientist within it one? Is a codebase a locus, or each contributor's fork?

Is a training paradigm a locus at all, or is it closer to the rule F_t itself — the thing being modified, rather than a site at which modification occurs?

That last case is worth resolving explicitly, because it is a category confusion rather than a hard case. A paradigm, a shared methodological convention, or a common software dependency is not a locus in this essay's sense; it is part of what makes candidate loci correlated, i.e. it is exactly the kind of common-cause mechanism the exchangeable model of Sec. 13.3 represents through the shared factor Θ. Two labs running different experiments under the same statistical convention are two loci with high $\bar{\rho}$, not one locus; the convention is not itself a third thing requiring its own count. Once this confusion is set aside, the remaining question — lab versus individual scientist, codebase versus fork — is genuinely hard, and does not have a domain-independent answer: the right grain depends on which rule-update process is under discussion, since the same physical system can be one locus with respect to one recursive process (a lab is a single locus with respect to institutional funding decisions that govern it as a whole) and many loci with respect to another (the same lab is many loci with respect to individual experimental replications performed independently by its members).

The essay's apparatus is more robust to this difficulty than it might first appear, because D_R was derived (Sec. 13.4) from a correlation statistic, $\bar{\rho}$, rather than from a locus count taken as metaphysically primitive. This matters operationally: if an analyst individuates loci too finely — treating what is really one institution's shared convention as many independent scientists — the resulting $\bar{\rho}$ estimated across those units will simply come out high, and D_R will correctly collapse back toward the coarser, more accurate effective count. If the analyst individuates too coarsely — treating a diverse institution as a single locus — $\bar{\rho}$ within it, were it measured at finer grain, would be low, and treating the institution as one locus rather than several under-credits real internal diversity that a finer analysis would reveal. Neither error is fatal to the framework in the way it would be fatal to a theory built on raw n alone: D_R is, by construction, a statement about how much independent evidence a set of channels actually provides, and it is comparatively insensitive to which candidate partition an analyst starts from, provided $\bar{\rho}$ is estimated honestly at whatever grain is chosen. What the framework cannot do is supply the grain itself; that remains an empirical and domain-specific judgment, informed by which rule-update process is actually in question, not a further theorem this essay can derive.

13.2 An Operational Model

Suppose each rule-iteration at a locus independently introduces an uncorrected, $\mathcal{A}$-violating error with probability p (a rough idealization: real error processes are neither independent nor identically distributed across iterations, a point returned to below). Within one verification window, a locus undergoes $N^* = \tau/\Delta t$ iterations, so the probability that the locus accumulates *at least one* undetected error before the window's external check occurs is

$$q(N^*) \;=\; 1 - (1-p)^{N^*},$$

strictly increasing in N^*. This already recovers half the Compression Principle without further assumptions: shortening Δt relative to τ increases N^* and hence increases q, the per-locus probability of an undetected error surviving to the next verification point.

The n-dependence requires a further assumption about how the system as a whole tolerates a failure at any one locus. Suppose n independent loci operate with cross-checking such that a genuinely system-wide, $\mathcal{A}$-violating drift requires a *majority* of loci — some fraction $\theta \in (1/2, 1]$ — to fail within the same window, on the idea that a minority of erroneous loci can be identified and corrected by comparison against the uncorrupted majority (this is the quantitative content behind Sec. 5's informal claim that distributed institutions provide error correction). Under this assumption, system-wide risk is

$$\text{Risk}(n, N^*) \;=\; \Pr\big[\text{Binomial}(n, q(N^*)) \geq \theta n\big].$$

Proposition 13.4 (Compression bound under the quorum model)**.** *For fixed $q(N^*) < \theta$,* $\text{Risk}(n, N^*) \leq \exp\big(-2n(\theta - q(N^*))^2\big)$*, which tends to* 0 *as* $n \to \infty$*. For fixed* n*,* $\text{Risk}(n, N^*)$ *is non-decreasing in* $q(N^*)$*, hence non-decreasing in* N^**.*

Proof sketch. The bound is Hoeffding's inequality applied to a sum of n i.i.d. Bernoulli(q) variables: $\Pr[\text{Binomial}(n, q) \geq \theta n] \leq \exp(-2n(\theta - q)^2)$ whenever $\theta > q$. Monotonicity in q (and hence in N^* via $q = q(N^*)$) is immediate since increasing the success probability of each Bernoulli trial stochastically increases the binomial sum, and $\Pr[\text{Binomial} \geq \theta n]$ is non-decreasing under stochastic dominance. □

Conjecture 13.5 (Compression Principle)**.** Beyond the specific quorum model of Prop. 13.4, risk of inadmissible recursive drift increases with $N^* = \tau/\Delta t$ and decreases with effective, uncorrelated locus count: concentrating recursive modification into few loci with an iteration period much shorter than the available external-verification period increases the number of self-modifications that occur before any independent check can occur.

Remark 13.6. Proposition 13.4 is a theorem, not a conjecture, but only about the idealized quorum model — independent per-locus failure, i.i.d. error probability, and majority-vote correction. Conjecture 13.5 is the claim that the qualitative shape of this result (risk falling exponentially in locus count, rising in N^*) survives relaxing those idealizations to something resembling real distributed systems, where failures are correlated, error probabilities vary, and correction is not a clean majority vote. The rest of this section relaxes the independence assumption directly, rather than leaving the relaxation as an unexamined promissory note.

13.3 Relaxing Independence: An Exchangeable Failure Model

Independent per-locus failure is the least realistic part of Sec. 13.2's model: real loci sharing a codebase, a training corpus, or an institutional culture do not fail independently. The standard way to model correlated binary outcomes without abandoning tractability is a common-shock construction: let $\Theta \sim \text{Beta}(\alpha, \beta)$ with mean $q = \alpha/(\alpha + \beta)$, and conditional on Θ, let the n failure indicators $X_1, \ldots, X_n$ be i.i.d. Bernoulli(Θ). This is the Beta-Binomial model, and it has two properties that make it the right tool here: the marginal failure probability of each locus is still q, and the pairwise

correlation between any two loci's failures is exactly

$$\bar{\rho} \;=\; \frac{1}{1+\alpha+\beta},$$

a standard identity for this model. Varying $\alpha+\beta$ sweeps $\bar{\rho}$ continuously from 0 (independent) to 1 (fully correlated) while holding the marginal risk q fixed, which is exactly the axis Sec. 14 needs.

Proposition 13.7 (Two exact limits)**.** *Let $S_n = \sum_i X_i$ under the Beta-Binomial model above, and let* Risk $= \Pr[S_n \geq \theta n]$ *for fixed $\theta \in (0,1)$.*

1. *As $\bar{\rho} \to 0$ ($\alpha+\beta \to \infty$ with q fixed), Θ degenerates to the point mass at q, $X_1, \ldots, X_n$ become i.i.d.* Bernoulli(q)*, and* Risk $\leq \exp(-2n(\theta-q)^2)$ *exactly as in Prop. 13.4.*

2. *As $\bar{\rho} \to 1$ ($\alpha+\beta \to 0$ with $\alpha/(\alpha+\beta) \to q$ fixed), Θ converges in distribution to a two-point law,* $\Pr[\Theta=1] = q$*,* $\Pr[\Theta=0] = 1-q$*, so $S_n = n$ with probability q and $S_n = 0$ with probability $1-q$, giving* Risk $\to q$*, independent of both n and θ.*

Proof sketch. (1) is Prop. 13.4 applied to the degenerate limit directly. (2) uses the standard fact that Beta(α,β) converges weakly to a two-point distribution at $\{0,1\}$ with masses $(1-q, q)$ as $\alpha, \beta \to 0$ with $\alpha/(\alpha+\beta) \to q$ held fixed; conditioning on each atom gives $S_n \equiv n$ or $S_n \equiv 0$ respectively, and $\Pr[S_n \geq \theta n]$ for $\theta \in (0,1)$ then equals $\Pr[\Theta=1] = q$ in the limit. □

Part 2 of Proposition 13.7 is the Monoculture Corollary (Sec. 15) proved rather than asserted: in the fully correlated limit, additional loci buy *no* reduction in risk whatsoever — not a smaller exponential benefit, but exactly none, because the system either fails everywhere at once or nowhere at all, and which of those two outcomes occurs is settled by the shared factor Θ before any individual locus's behavior is even realized.

13.4 Deriving D_R Rather Than Importing It

The Beta-Binomial model also gives Sec. 14's effective-diversity formula a derivation instead of an analogy. A direct calculation gives

$$\mathrm{Var}(S_n) = n\,q(1-q)\,\big[1+(n-1)\bar{\rho}\big].$$

Matching this to the variance an i.i.d. system of n_{eff} trials would need to reproduce the same variance in the *sample mean* S_n/n — i.e. solving $q(1-q)/n_{\text{eff}} = q(1-q)[1+(n-1)\bar{\rho}]/n$ — gives

$$n_{\text{eff}} \;=\; \frac{n}{1+(n-1)\bar{\rho}},$$

which is exactly Definition 14.1's D_R with $m = n$. The formula used in Sec. 14 is therefore not an imported survey-statistics convention applied to a new domain by analogy; it is what falls out of demanding that a correlated and an independent system agree on how noisy their failure rate estimate is.

Remark 13.8 (Where the approximation breaks). Substituting D_R for n in Prop. 13.4's exponent, Risk $\lesssim \exp(-2D_R(\theta - q)^2)$, is a reasonable interpolation for intermediate $\bar{\rho}$, but Proposition 13.7 shows it is not exact at $\bar{\rho} \to 1$: there, $D_R \to 1$ for any n, and the substituted bound gives $\exp(-2(\theta - q)^2)$, a fixed constant — whereas the true limiting risk is q, which need not equal that constant. The variance-matching heuristic correctly captures the *first moment* of how correlation degrades effective sample size, which is enough to derive D_R's functional form, but tail probabilities depend on higher moments the heuristic does not track. The two exact limits of Prop. 13.7 should be treated as the reliable endpoints; the D_R-substituted bound in between is a useful approximation whose error has not been bounded here, and Conjecture 13.5 should be read as covering that unbounded middle region, not the endpoints, which are now theorems.

13.5 Compression and Monotony

The exchangeable model of Sec. 13.3 was built to answer a question about risk, but it has a second consequence worth drawing out on its own. As $\bar{\rho} \to 1$, the n loci do not merely lose their ability to protect each other from correlated failure (Prop. 13.7); their trajectories $x_1, \ldots, x_n$ become, in the limit, literally the same trajectory, since $X_i \equiv X_j$ almost surely once Θ is degenerate at $\{0, 1\}$. High compression is not merely a risk condition; it is, in the same limit, a condition under which the system's n nominally distinct histories converge toward one history.

Remark 13.9. Call this convergence *monotony*: a system exhibits monotony to the extent that its component histories, which could in principle diverge, instead move together. The exchangeable model gives a direct route to this notion: $\bar{\rho}$ is simultaneously the correlation that degrades effective repair diversity (Sec. 14) and a measure of how monotonic the system's histories are, since $\bar{\rho} \to 1$ is exactly the condition under which distinct loci stop generating distinguishable trajectories at all. On this reading, compression's danger has two faces that are really one: $D_R \to 1$ removes the redundancy that would otherwise contain an error (Prop. 13.7), and the same limit is the condition under which the system's histories become too similar to constitute independent evidence about anything, which is the concern this author's other work on monotony and convergence addresses in a different register. This essay does not attempt to establish that identification formally — doing so would require importing definitions this essay has not developed and cannot verify are compatible with how monotony is defined elsewhere — but the two notions sharing $\bar{\rho}$ as a common variable, arrived at from the risk side here and (by report) from a different side elsewhere, is worth recording as a candidate bridge for future work rather than treating as coincidence.

Remark 13.10 (Compression as loss of counterfactual diversity). Definition 10.1 read K as the volume of a system's reachable-state space. Under that reading, $D_R \to 1$ has a semantic gloss beyond the statistical one: since the n loci's trajectories converge toward a single trajectory in the $\bar{\rho} \to 1$ limit, the set of *counterfactual futures* the system as a whole could still reach — the futures available if one locus had gone differently than the others — collapses along with D_R. A system with high $\bar{\rho}$ does not merely fail to catch its own errors; it has, in the same limit, fewer genuinely distinct futures available to it than its raw component count n would suggest, because most of the branches that would have made those futures distinct have already been foreclosed by the correlation. Compression is, on this

reading, a statement about reachability (how large a space of futures remains open) as much as it is a statement about risk (how likely an error is to go uncaught) — and the two are related rather than coincidental, since a system with little reachable diversity left has correspondingly little room for a corrective response to occupy even when an error is caught in time.

14. Repair Diversity and the Recursive Stability Principle

The quorum model of Sec. 13.2 assumed n *independent* loci. Real repair and verification channels are rarely independent: ten servers running identical software share a bug the moment it is introduced; ten peer reviewers trained in the same paradigm share its blind spots. Raw count overstates real redundancy whenever failures are correlated, and the Compression Principle needs a variable that accounts for this.

Definition 14.1 (Repair-path diversity)**.** Let a system have access to repair or verification channels $c_1, \ldots, c_m$ (replication, peer review, competing institutions, independent physical copies, etc.), and let $\bar{\rho} \in [0, 1]$ denote the average pairwise correlation among their failure events. Define the *effective repair diversity*

$$D_R \;=\; \frac{m}{1 + (m-1)\bar{\rho}},$$

the standard design-effect correction for effective sample size under intra-class correlation, applied here to repair channels rather than survey respondents — and, as Sec. 13.4 shows, recoverable directly from a variance-matching argument under the exchangeable failure model of Sec. 13.3 rather than assumed by analogy.

Proposition 14.2 (Boundary behavior of D_R)**.** *$D_R = m$ when $\bar{\rho} = 0$ (fully independent channels), and $D_R \to 1$ as $\bar{\rho} \to 1$ for any fixed m (fully correlated channels), regardless of how large m is.*

Proof. Direct substitution: $\bar{\rho} = 0$ gives $D_R = m/1 = m$. As $\bar{\rho} \to 1$, $D_R \to m/(1 + (m-1)) = m/m = 1$. □

14.1 A Worked Example

Consider two systems, each with $m = 10$ repair channels. System A consists of ten replica servers running identical, unmodified software images with $\bar{\rho} = 0.9$ (a bug in the shared image is very likely to be present, and to fail, in all ten simultaneously): $D_R = 10/(1 + 9 \times 0.9) = 10/9.1 \approx 1.10$. Ten nominal copies buy barely more effective diversity than one. System B consists of ten independently developed, competing implementations — different teams, different codebases, different assumptions — with $\bar{\rho} = 0.1$: $D_R = 10/(1 + 9 \times 0.1) = 10/1.9 \approx 5.26$. The same raw count $m = 10$ yields nearly five times the effective diversity once correlation is accounted for. This is the essay's answer to a question Sec. 13 left open: n in the quorum model of Prop. 13.4 should be read as D_R, not as raw locus count, whenever the loci in question can fail together.

Conjecture 14.3 (Recursive Stability Principle)**.** Stability of a recursive adaptive system is increasing in D_R, not merely in the raw count m of agents or copies: $S \propto D_R$.

Remark 14.4. Definition 14.1 discharges half of this section's earlier skeleton obligation — a concrete, standard construction for D_R rather than a placeholder — and Sec. 14.1 discharges the other half, a numeric example showing correlated copies buying little real diversity. What remains conjectural is only the proportionality $S \propto D_R$ itself, because "stability" S has not been given an operational definition independent of the risk model of Sec. 13; substituting D_R for n in Prop. 13.4's bound would turn this conjecture into a theorem of the same kind, at the cost of assuming the quorum model's idealizations apply with D_R playing the role of effective independent trial count, which is itself an approximation rather than an exact reduction.

14.2 Diversity Beyond Count

A single scalar $\bar{\rho}$ can understate a real failure this section's model is otherwise built to catch. Ten channels can have low measured failure-correlation — they fail on different inputs, at different times, for different proximate causes — while still sharing a common architecture, a common training corpus, or a common methodological assumption that would cause them to fail together on exactly the class of input that matters most. Failure correlation, estimated from ordinary operating conditions, is not guaranteed to reveal a shared vulnerability that has simply never yet been triggered. This is a real limitation of Definition 14.1 as stated, not a gap in its derivation: $\bar{\rho}$ is a property of the failure process actually observed, and a channel set can have measured $\bar{\rho} \approx 0$ under ordinary conditions while having architectural, methodological, or epistemic similarity that would produce $\bar{\rho} \approx 1$ under a stress the channels have not yet encountered. Distinguishing *architectural diversity* (different underlying mechanisms), *methodological diversity* (different procedures for reaching conclusions), and *epistemic diversity* (different background assumptions and training) as separate axes, rather than collapsing them into one observed correlation statistic, would let an analyst ask not only "how correlated have these channels' failures been" but "how similar are the mechanisms that would need to fail for them to fail together" — a harder, more forward-looking question this section's formal apparatus does not yet answer, and one operational $\bar{\rho}$ estimates can only ever answer retrospectively.

Remark 14.5 (Estimating $\bar{\rho}$ in practice). This essay has treated $\bar{\rho}$ as a given parameter throughout, and it is worth being honest about how far that is from a measurement procedure. Three imperfect proxies are available where direct failure-correlation data is sparse or entirely retrospective, and each captures a different part of what $\bar{\rho}$ is meant to represent. *Shared-dependency analysis* — treating two channels as more correlated the more of their software dependencies, training data provenance, or supply chains overlap — is the most operationalizable proxy and the one closest to how software supply-chain risk is already assessed in practice, but it is a structural proxy for correlation rather than correlation itself, and can miss correlated failure modes that share no traceable dependency at all. *Historical co-failure rates*, where available, are the most direct estimate of $\bar{\rho}$ as this essay defines it, but require a large enough sample of past failures to be statistically meaningful, which is precisely what is missing for the failure modes with the highest stakes (a first-of-its-kind AGI failure has, definitionally, no historical co-failure rate to estimate from). *Architectural and methodological distance metrics* — how different two systems' underlying mechanisms are, independent of whether either has yet failed — are the most forward-looking proxy, in the spirit of Sec. 14.2's distinction, but require a domain-specific notion of distance this essay does not supply. None of the three is adequate on its own, and combining them into

a single operational estimate of $\bar{\rho}$ for a real system such as the current AI ecosystem is, honestly, an unsolved measurement problem rather than a straightforward application of Definition 14.1 — the theorem tells an analyst what to look for; it does not yet tell them how to look.

15. Monocultures and Recursive Failure

Corollary 15.1 (Monoculture Corollary)**.** *As $D_R \to 1$, a single rule-error is no longer contained by independent cross-checking and propagates without an external correction opportunity, producing system-wide rather than local failure.*

Proof sketch. Direct from Conjecture 14.3 together with Prop. 13.4: substituting $D_R \to 1$ for n in the quorum-model bound removes the exponential suppression of risk with locus count entirely, since the bound $\exp(-2n(\theta - q)^2)$ degenerates to a constant near 1 as the effective trial count collapses to one. Correlation, not raw count, is what determines whether redundancy actually functions as redundancy. □

Remark 15.2. Financial contagion, agricultural monoculture collapse, and software supply-chain single-points-of-failure are instances of Corollary 15.1 outside any AI context, which is the point: the corollary is about D_R, not about intelligence, artificial or otherwise. The 2021 discovery of a critical vulnerability in a single, near-universally embedded software logging library affected an estimated proportion of enterprise Java systems large enough that the raw count of affected organizations — nominally in the millions — provided no protection whatsoever, because $\bar{\rho}$ across those organizations' exposure was close to 1: they were not independent instances of risk, they were one risk wearing many organizational names, in exactly the sense System A of Sec. 14.1 illustrates. The Irish Potato Famine is the pre-computational version of the same corollary: a food system with a large raw count of individual farms and an enormous total planted area nonetheless had D_R close to 1 with respect to *Phytophthora infestans*, because nearly the entire crop was one genetically near-identical cultivar, and genetic near-identity is exactly the biological form of $\bar{\rho} \to 1$. Neither case required anything resembling artificial intelligence; both are instances of the same structural failure this section formalizes.

This closes Part III's argument. Sec. 13 showed that risk falls exponentially in effective, independent locus count under a stated model, and rises with the ratio of iteration speed to verification speed; Sec. 14 showed that "locus count" has to mean effective, correlation-adjusted diversity rather than raw copies, with a twenty-line calculation separating a ten-times-replicated monoculture from ten genuinely independent institutions; and this section's corollary is the two put together, applied to the limiting case where diversity collapses entirely. None of the three results mention intelligence. All three are about the topology through which recursive modification propagates and gets checked — which is exactly the variable Part IV now asks what AGI does to.

15.1 Monocultures and Path Dependence

The examples above are synchronic: they describe a single failure event, occurring once, against a fixed background of available diversity. A further cost of $\bar{\rho} \to 1$ only becomes visible over a longer

horizon. Once alternative lineages have actually disappeared — not merely gone temporarily unused, but ceased to exist as live, maintained alternatives — future diversity cannot simply be regenerated on demand; it has to be rebuilt from whatever fragments or relatives survive, which is a slower and less certain process than drawing on lineages that had been kept alive in parallel all along. A language that loses its last fluent speakers is not a language whose D_R has temporarily dropped and can be restored by reintroducing it; whatever revival is later possible works from documentation and related languages, which is a fundamentally different and more constrained starting point than the language's own continued use would have been. A software ecosystem that consolidates around one dominant framework and lets competing approaches lapse from active maintenance faces the same asymmetry if the dominant framework later proves to have a structural flaw: rebuilding a genuinely independent alternative from scratch is slower, and initially lower-D_R in its own right, than an alternative that had continued developing in parallel would have been. Agricultural cultivars kept in seed banks after commercial cultivation abandons them are this essay's clearest illustration of the alternative: preserving a lineage in a non-active but recoverable state is cheaper than losing it outright, and cheaper still than the commercial monoculture's own collapse would be without it. The general point strengthens Sec. 14's argument rather than merely restating it: D_R is not a quantity a system can costlessly restore after the fact once it has been allowed to collapse, because the alternatives D_R was measuring the diversity of are themselves subject to atrophy, and atrophied alternatives do not spring back at the moment they turn out to have been needed.

Part IV

Case Studies and Synthesis

16. Science as Recursive Improvement of Improvement

Definition 16.1 (Scientific recursion loop)**.** Let K_t be accumulated knowledge, M_t methods, I_t instruments. Suppose $K_{t+1} = M_t(I_t, K_t)$, and methods and instruments are themselves revised by knowledge: $(M_{t+1}, I_{t+1}) = \Xi(K_{t+1}, M_t, I_t)$.

Science is a Level-3, sympoietic (Sec. 5) recursive system: it is not merely knowledge accumulation but method self-revision, distributed across researchers, institutions, and time. The development of the randomized controlled trial, of blind and double-blind protocols, and of formal statistical inference were not discoveries *within* an existing method; they were revisions *of* method, produced by knowledge (statistical theory, accumulated case studies of bias) feeding back into M_t exactly as Definition 16.1 describes. Its high D_R under ordinary conditions — independent replication, competing research groups, adversarial peer review, many institutions with no shared incentive to agree — is, by Sec. 14, the structural reason this recursive loop has remained admissible for roughly four centuries without central coordination.

Remark 16.2 (Science as an application of Part II, not only an instance of Part I). Replication, peer

review, and adversarial collaboration have been introduced above primarily as sources of D_R (Part III's variable), but they are, first and more basically, repair mechanisms in Part II's sense: they are R_t, the capacity to detect and correct a specific erroneous claim, before they are ever a question of how diverse that capacity is. And per Sec. 8, none of them function without H_t: peer review has nothing to check a claim against without a citable prior literature, and replication is meaningless without an archival record precise enough to specify what is being replicated. Science is therefore not merely an example of recursive continuation (Part I) that happens to also illustrate compression risk (Part III); it is an application of the full apparatus this essay has built, in order — recoverable history enabling repair, repair capacity needing to keep pace with accumulated knowledge, and the diversity of that repair capacity determining whether an individual failure stays local or becomes systemic, which is exactly what the replication crisis below shows going wrong at the third stage while the first two remained intact.

Remark 16.3 (The replication crisis as a D_R failure). Science's own history supplies a case where this mechanism visibly degraded, and it is instructive precisely because the raw locus count m (number of active researchers, journals, and institutions) was, if anything, larger than ever. In the years leading up to psychology's and related fields' replication crisis, a large fraction of published findings relied on closely related statistical practices — small sample sizes, flexible analysis choices made after seeing the data, a shared threshold for statistical significance — that were formally distinct researchers but methodologically correlated choices. In the vocabulary of Definition 14.1, $\bar{\rho}$ across the field's nominal verification channels was high, because the channels were not independently arriving at their methodological standards; they were converging on the same conventions via shared training and shared incentives. The result was exactly Corollary 15.1's prediction: when the shared methodological assumption turned out to be unreliable, it failed simultaneously across a large fraction of the literature, not independently across isolated studies, and the failure looked, retrospectively, far more systemic than the field's large m would have suggested it could be. The field's ongoing response — pre-registration, mandatory replication, adversarial collaboration, greater diversity of statistical approaches — is a direct, if not always self-described, attempt to raise $\bar{\rho}$'s complement back toward the conditions under which Prop. 13.4's bound actually delivers the protection a large m seems to promise.

17. Civilization as Meta-Learning

Definition 17.1 (Civilizational recursion loop)**.** Let $C_t = (T_t, I_t, A_t)$ be tools, institutions, and agents. Suppose $C_t \to (T_{t+1}, I_{t+1})$, $(T_{t+1}, I_{t+1}) \to A_{t+1}$, and $A_{t+1} \to C_{t+1}$.

Schools, libraries, and laboratories are not merely products of civilization; under Definition 17.1 they are the F_t of a civilizational Level-3 system, mechanisms by which civilization modifies its own future capacity to produce. Civilization is species-scale recursive continuation, exhibiting the same sympoietic structure as Sec. 5 at greater scale and slower Δt.

Remark 17.2 (Textual transmission as a historical D_R transition). The transition from scribal to printed textual transmission is a clean historical instance of a civilizational system's D_R changing by orders of magnitude within a single technology's adoption. Under scribal copying, a text's transmission lineage

typically ran through a small number of sequential copies — each new manuscript copied from one, or occasionally two, prior exemplars — so an error introduced at any point propagated forward through every subsequent copy descended from it, with no independent line available for cross-checking. This is Definition 14.1's $\bar{\rho} \to 1$ condition realized structurally, not statistically: sequential, single-source copying is close to the least diverse repair topology available, regardless of how many manuscripts eventually existed in total. Movable-type printing did not merely make copies faster; it made large numbers of *textually identical, independently distributable* copies from a single set type, which — once multiple independent print runs, editions, and geographically dispersed printers entered circulation — allowed textual corruptions to be identified by comparison across genuinely independent lines of transmission in a way scribal culture structurally could not support at scale. The effect on civilizational D_R for the specific function of preserving textual accuracy was closer to the jump from System A to System B in Sec. 14.1 than to a mere increase in raw copy count m.

Civilizational institutions more generally — courts that create precedent reviewable by later courts, patent systems that make prior art searchable and challengeable, contract law that makes agreements independently enforceable rather than dependent on the continued goodwill of the original parties — are, functionally, mechanisms for keeping D_R high for the specific recursive loop of legal and economic continuation, in the same sense peer review keeps D_R high for the scientific loop of Sec. 16. None of these institutions were designed with Definition 14.1 in mind; their persistence, in a Darwinian sense not unlike Sec. 6's treatment of biological recursion, is some evidence that societies lacking them are less likely to sustain civilizational recursion over comparable timescales, whether or not anyone inside those societies ever represented the mechanism explicitly.

17.1 Coordination as Repair Capacity

The institutions above have been described as a D_R mechanism, which is the right description for how they contain error once it occurs. They also do something Part II's variables, not Part III's, are built to describe: a court system, a logistics network, or a standards body is not only diversifying repair pathways, it is a source of R_t itself — the raw capacity to detect and correct faults in the civilizational recursion loop of Definition 17.1, distinct from how diverse that capacity is once it exists. A society can increase R_t by building more courts, more auditors, more logistics capacity, without any of that increase touching D_R at all, if the new capacity is added in a way that is highly correlated with the old (the same training, the same doctrine, the same assumptions) rather than independent of it. Coordination, on this reading, is not merely productive activity that happens to have positive externalities; a meaningful fraction of what coordinating institutions do is better described as manufacturing R_t directly, which makes them subject to Theorem 10.4 and Theorem 11.1 in their own right: a civilization whose coordination capacity fails to scale with the maintenance burden of its own accumulated complexity is in exactly the position Sec. 11.1's imperial-administration example describes, and a civilization that scales coordination capacity without scaling its diversity is in the position Sec. 15 describes. The two failure modes are independent of each other and both available to the same institution at once.

18. AGI as an Extreme Case

18.1 The Claim This Essay Has Actually Earned

It is worth stating precisely what Parts I and II license, because it is narrower than it can sound. The standard argument for an intelligence explosion treats recursive self-improvement as doing most of the work on its own: a system capable of studying and modifying itself acquires a qualitatively new capacity, and that capacity is what carries it from human-level to superhuman capability. Sections 4–6 show that the capacity in question — Level-3, rule-level recursive modification — is not qualitatively new at all. Cells, immune systems, languages, markets, and evolution by natural selection all instantiate it, and Sec. 6 shows this requires no intelligence whatsoever. The defensible conclusion is:

Recursive self-modification is neither unique to intelligence nor sufficient for intelligence.

This is already useful. It removes recursion itself as a marker of anything distinctive, and it shifts the burden of argument: anyone claiming AGI's recursive self-improvement is dangerous *because it is recursive* owes an account of why AGI's instance of an old and common pattern should behave differently from every other instance of it.

18.2 What This Does Not Show

It would be a mistake, and a tempting one, to treat the conclusion above as a refutation of intelligence-explosion arguments generally. It is not. Showing that

$$\text{Rule-level recursion} \not\Rightarrow \text{intelligence, or anything AGI-specific}$$

does not touch a more specific and more carefully stated claim, which is roughly that *high-speed, low-diversity* rule-level recursion produces dynamics unlike anything biological or cultural recursion has produced, precisely because biological and cultural recursion has never operated under those particular conditions. That claim is not refuted by pointing to evolution, language, or markets, because the argument was never that recursion as such is dangerous — the argument, stated carefully, is conditional on rate and concentration, and the honest response to a conditional claim is to examine the conditions, not to point out that the unconditional version of the claim is false. An essay that spent Parts I and II establishing recursion's ubiquity and then treated that as settling the question of AGI risk would be substituting an easier, adjacent victory for the actual argument it needs to have.

18.3 The Question That Remains

The honest reformulation is this: Sections 4–17 establish that Level-3 recursive modification, distributed across many loci ($n \gg 1$) with an iteration period large relative to available external verification, is the ordinary condition of living and cultural systems, and Part II's Repair Dominance and Admissibility results give a structural reason this distribution matters: it is what has historically kept $R_t \geq C(K_t)$ and D_R high enough for recursive modification to remain admissible. AGI is distinguished not by exhibiting Level-3 recursion but by the open possibility of instantiating it under the opposite conditions:

$n \to 1$, $\Delta t \ll \tau$ (Conjecture 13.5), and $D_R \to 1$ if the system's internal repair or oversight channels are themselves centralized or correlated (Conjecture 14.3). Whether any actual AGI system does or will satisfy these conditions, and whether satisfying them is in fact sufficient for the kind of runaway dynamics the intelligence-explosion literature describes, are empirical and architectural questions this essay's formal apparatus does not settle — it only says what would need to be true for the concern to be well-founded, and, just as importantly, what would need to be true for the concern to be misplaced (high D_R, verification periods comparable to or shorter than Δt, in which case the same historical mechanism that stabilized biological and cultural recursion would be available to AGI as well).

Remark 18.1. This reframes the object of concern without dissolving it. The claim is not Risk $\propto$ Intelligence, which Secs. 5 and 6 already show cannot be the whole story. Nor is the claim Risk ≈ 0 because recursion is old. The claim, still marked as a conjecture rather than a proof (Sec. 13), is closer to Risk $\propto N^*/D_R$ — a conjunction of high iteration-to-verification ratio and low repair diversity — of which AGI is a candidate extreme instance, not a refuted one and not a proven one.

18.4 AGI as Compression of Civilizational Recursion

Sections 16 and 17 are, on reflection, the same kind of case examined at two different scales: both are sympoietic Level-3 systems (Sec. 5) with $n \gg 1$ independent loci — researchers and institutions in one case, tools, institutions, and agents in the other — and both derive their long-run admissibility substantially from that scale, per Sec. 17.1 and Remark 16.2. Stated this way, the essay's argument about AGI has a sharper and more specific form than "AGI is a system with low D_R": science and civilization are the two largest-scale, longest-running instances of exactly the recursive process AGI is being asked to perform — accumulating capability, revising the methods and institutions that produce further capability, and doing so recursively across time — and both have only ever done so with $n \gg 1$.

Remark 18.2. The comparison this essay's own apparatus licenses is not AGI versus an abstract standard of safety, but AGI versus the two existing systems that have most successfully performed recursive capability accumulation at civilizational scale:

$$\text{Science: } n \gg 1, \qquad \text{Civilization: } n \gg 1, \qquad \text{AGI (candidate): } n \to 1.$$

The novelty this essay has been arguing for throughout is not that AGI recurses — Sec. 6 already established that recursion needs no intelligence at all — and it is not even that AGI might recurse quickly, since Sec. 13 treats speed and locus count as jointly determining risk rather than either alone. The novelty, stated at the scale this comparison makes vivid, is the possibility of concentrating a process that has, in its two most successful known instances, only ever operated as a distributed, many-institution, many-century undertaking into a small number of loci operating on a computational rather than an institutional or evolutionary timescale. Whether that concentration is achievable without also importing the low-D_R vulnerabilities Sec. 15 describes is exactly Sec. 18.3's open question, restated here at the scale it was always implicitly about.

19. Distinction Preservation as the Deeper Object

The viability manifold $\mathcal{A}$ of Definition 3.1 has, throughout Parts I–III, done the actual explanatory work this essay needed: not capability, not intelligence, but the preservation of whatever structure makes the system's continued activity possible. This is the same object this author's Admissibility Program treats under the name *recoverable distinction*: a system persists not by retaining its states unchanged but by retaining the capacity to reconstruct what matters about them across transformation. A cell preserves distinctions between self and non-self; a scientific field preserves distinctions between supported and unsupported claims; a civilization preserves distinctions encoded in law, language, and institutional memory. Recursive continuation, in the vocabulary developed here, and distinction preservation, in the vocabulary of that separate program, are the same claim viewed from systems theory and from a theory of admissibility respectively. This essay does not import that program's full apparatus; the connection is noted here as the place a future, unified treatment would need to go, not as a claim this essay has independently established.

A second, independent line of support for treating $\mathcal{A}$ as the fundamental object is more useful precisely because it comes from outside this author's own corpus: a body of argument in analytic philosophy of truth, developed with no stake in recursive systems or admissibility theory, converges on structurally the same idea under the heading of *non-propositional truth*.

19.1 Admissibility Beyond Continuation

Before turning to that material, it is worth saying plainly why a theory built to describe recursive systems should be expected to have anything to say about truth at all, since the connection is not obvious on its face and the transition below can otherwise read as an abrupt change of subject rather than a continuation of the argument.

The viability manifold $\mathcal{A}$ has never, in this essay, been defined in terms specific to recursion. Definition 3.1 states a relationship between a state and a standard that state must meet — $x_t \in \mathcal{A}$ — and everything Parts I–III do with that relationship (autopoiesis, repair, compression) is downstream of applying it to states that are also steps in a recursive process. But nothing in the definition itself mentions recursion, and a state can be evaluated against a standard, admissible or not, whether or not it is embedded in any sequence at all. If $\mathcal{A}$ is genuinely the more basic notion this essay has been treating it as — more basic than capability, more basic than improvement — then it should be recognizable in contexts that have nothing to do with recursive systems specifically, wherever the same underlying question arises: does a given particular meet the standard its kind is held to. Truth, on any theory that ties it to success rather than to a purely formal relation between sentences and facts, is exactly this question asked of representations. That is the reason to expect a connection, stated before the connection is drawn, rather than a connection that shows up unannounced and asks to be taken on faith.

19.2 Truth Without Propositions

Contemporary analytic philosophy standardly restricts truth to propositions and propositional attitudes: sentences, beliefs, assertions. Ordinary English usage does not observe this restriction — we speak of

a true friend, true north, a true chair, a true coin — but the standard response, following what Rahlwes and Dickson call *restriction* or the weaker *non-relation*, is to treat these uses as either philosophically irrelevant or simply unrelated to the propositional case, a homonym no more interesting than the two senses of "bank." Rahlwes and Dickson's *In Defense of Nonpropositional Truth* argues against both positions by developing unified theories of truth under which objects, actions, and propositions are truth-bearers in exactly the same sense. Three such theories are relevant here.

19.3 The Non-Deception Theory and the Viability Manifold

> *Non-deception theory* (attributed to early Buddhist and Madhyamaka sources): X is true iff X is non-deceptive, false iff X is deceptive. A true chair supports the weight of the person sitting on it, as its socially established function leads one to expect; a chair that collapses is a false chair.

The structural content of this theory is immediate once stated in this essay's vocabulary. Fix an object or system X of a given kind, and let Φ_X denote the functional profile that kind is expected to satisfy — the behavior a chair, a coin, or a claim about snow is supposed to exhibit. Let $\mathcal{A}_{\Phi_X}$ be the viability manifold determined by that profile: the set of states in which X's behavior continues to satisfy Φ_X's expectations.

> **Correspondence.** X is non-deceptive (true) at t in exactly the sense that $x_t \in \mathcal{A}_{\Phi_X}$ (Definition 3.1); X is deceptive (false) in exactly the sense that $x_t \notin \mathcal{A}_{\Phi_X}$.

A chair that collapses is a chair whose state has exited the viability manifold determined by the chair-function profile. This is not an analogy grafted onto the non-deception theory from outside; it is the same claim, arrived at independently. The non-deception theory's treatment of *graded* truth — salt water is truer than a mirage because it satisfies more of water's socially determined functions, even though it fails others (thirst-quenching) — corresponds to the observation, already available in this essay's own apparatus (Sec. 10), that viability itself is not binary: V_t is a real-valued quantity, and a state can be closer to or further from the boundary of $\mathcal{A}$ rather than simply inside or outside it. Approximate truth, on this reading, is distance from $\partial\mathcal{A}_{\Phi_X}$, not a separate and harder-to-formalize notion bolted onto a binary theory of truth.

19.4 The Trustworthiness Theory and Observer-Independent Admissibility

> *Trustworthiness theory* (derived from the Hebrew Bible): X is true iff X is trustworthy, false iff untrustworthy. A fake doorknob on a theater set is untrustworthy — one *ought not* to trust it — even for the set builder who is never actually deceived by it.

The non-deception theory, taken alone, makes admissibility observer-relative: whether X deceives depends on a particular observer's background knowledge and expectations, which is why Rahlwes and Dickson's stage-door example (true for the set builder who knows better, false for the visitor who doesn't) looks like a straightforward relativism. The trustworthiness theory's normative move — replacing *does it deceive this observer* with *ought it be trusted, independent of who is asking* — is exactly the move already built into Definition 3.1, which fixes $\mathcal{A}$ as a property of the system and its

functional profile, not as a property relative to any given observer's beliefs about that system. This essay's viability manifold was never observer-relative to begin with; the trustworthiness theory shows why that choice, rather than the more psychological non-deception reading, is the one worth preserving if the goal is an objective admissibility claim rather than a claim about who currently happens to be fooled.

19.5 The Teleological Theory and Recursive Continuation

> *Teleological theory*: X is true iff X is an ideal instance of its kind, false iff a defective instance. A true heart is one that pumps blood effectively, because pumping blood is what a heart is for.

This is the closest of the three to this essay's own thesis, because it is not merely a claim about a single state but a claim about a kind's characteristic activity being carried out well over time — which is exactly what Definition 4.1's organizational closure and Theorem 12.1's viable-recursion condition are claims about. A "true" autopoietic system, on this reading, is simply one that realizes its own telos: continued organizational closure, $P_t \to O_t \to P_{t+1}$, carried out ideally rather than defectively. A heart that pumps blood only sporadically, or a cell that only partially regenerates its own components, is a defective instance of its kind in exactly the sense that a state trajectory with $R_t < C(K_t)$ (Theorem 10.4) is a trajectory drifting toward exiting $\mathcal{A}$. The teleological theory's "ideal instance of its kind" and this essay's "trajectory that remains admissible under Theorem 12.1" are not two claims that happen to resemble each other; read carefully, they are the same claim about the same objects, arrived at from opposite directions — one starting from a theory of truth and generalizing outward to objects, the other starting from a theory of recursive systems and finding truth-language already fitting what it needed to say.

19.6 One Structure, Three Parameterizations

The three theories are not competitors offering incompatible accounts of the same cases; they largely disagree only about how the functional profile Φ_X and its associated $\mathcal{A}_{\Phi_X}$ should be fixed.

- **Non-deception:** Φ_X is fixed by a given observer's actual expectations — admissibility relative to *this* observer's model of X's kind.
- **Trustworthiness:** Φ_X is fixed by the socially or normatively correct expectations for X's kind, independent of any particular observer's actual beliefs.
- **Teleological:** Φ_X is fixed by X's own kind-defining function or telos, treated as objective rather than merely socially agreed.

Each is a different answer to *whose expectations, or what standard, defines the viability manifold*, not a different answer to *what it is for X to be true given that manifold*. This essay's own use of $\mathcal{A}$ throughout Parts I–III has consistently taken the teleological parameterization without saying so: an autopoietic system's $\mathcal{A}$ is fixed by its own organizational closure, not by any external observer's

expectations of it, and a scientific community's $\mathcal{A}$ is fixed by what continued inquiry requires, not by what any individual scientist currently believes.

19.7 Truth as the Representational Special Case

The correspondence developed above should not be read as Truth = Admissibility. Admissibility, as this essay uses it, is the broader category: a biological trajectory can be admissible or not, a fiscal trajectory can be admissible or not, a pattern of social coordination can be admissible or not, and in none of these cases is "true" or "false" the natural word for what has succeeded or failed. What non-propositional truth theories pick out is narrower — the special case where the functional profile Φ_X concerns *representation*: where X's job, whatever else it does, includes standing in for, guiding expectations about, or asserting something about some further state of affairs. A chair's function is structural, not representational, and yet the non-deception theory still calls a collapsing chair "false"; this is either evidence that ordinary usage extends "true/false" language somewhat loosely to any functional success or failure, or evidence that even a chair's affordances (Sec. 19.3's doorknob case makes this explicit) carry an implicit representational component — an object that looks graspable is, among other things, asserting something about what will happen if grasped. Either reading is compatible with the more conservative formulation this essay actually needs:

$$\text{Truth} \subseteq \text{Admissibility}, \qquad \text{with equality only under a representational functional profile } \Phi_X.$$

This is the claim worth defending. The stronger identity is not, and collapsing the two would cost this essay nothing it needs while inheriting every objection that has ever been raised against reducing truth to success generally (pragmatist theories of truth face exactly this objection, and this essay has no stake in re-litigating it here).

19.8 What This Does Not Claim

None of the above establishes that non-propositional theories of truth are correct, nor that propositional truth reduces to admissibility, nor that this essay has solved a live problem in philosophy of truth. Rahlwes and Dickson's own text raises serious objections to non-propositional truth theories generally — a change-of-subject argument (perhaps "true chair" and "true proposition" are simply homonyms), worries about whether non-deception theories can preserve ordinary logical entailment, and an epistemic-uselessness argument aimed at showing that only propositional truth serves philosophy's actual (epistemic) interests. This essay takes no position on whether those objections succeed. What can be said without settling that debate is narrower and more defensible: independent of whether "non-propositional truth" is the right name for it, the formal structure those theories describe — a kind-relative functional profile, and a fact of the matter about whether a given instance realizes or fails that profile — is the same structure this essay has called admissibility, and its recurrence in a debate this essay's authors did not construct is some evidence that the structure is doing real work, not merely convenient formal bookkeeping.

19.9 An Aside: Attribution and Inheritance

A version of this same question recurs, entirely outside philosophy of truth, in ordinary disputes about credit and legitimacy: whether a present, visible outcome is a sufficient basis for a claim, or whether the claim is only as good as the causal history that produced it and remains contestable for as long as that history can be re-examined. A priority dispute over who first proposed an idea, a credit dispute over who deserves recognition for a work that someone else made famous, and an inheritance dispute over whether a genealogical connection licenses a present social position are all instances of the same question asked about persons and works rather than about recursive systems: is the current state licensed by its actual generating history, or does it merely coincide with a history that would have licensed it. This is not offered as a further pillar of the essay's argument — nothing in Parts I–III has been shown to extend rigorously from viability manifolds to questions of credit or legitimacy, and the two domains differ in obvious ways a full treatment would need to take seriously. It is offered only as a further instance of the pattern noted throughout this section: that a present state is licensed by its history rather than by itself — the same claim Definition 3.1 makes formal for recursive systems — turns out to be recognizable in registers well outside the ones this essay was built to formalize.

20. Conclusion: The Geometry of Recursive Systems

Theorem 20.1 (Geometry of Recursive Systems, informal capstone)**.** *The long-run admissibility of a recursively self-modifying system is governed less by its capability K than by the relationship among K, its repair capacity R, its repair-path diversity D_R, and its viability manifold $\mathcal{A}$. Autopoiesis, sympoiesis, science, and civilization are instances of systems that have historically remained admissible by keeping R commensurate with $C(K)$ (Sec. 10) and D_R high (Sec. 14). AGI is a candidate instance of a system attempting rule-level recursion under the opposite conditions.*

The essay's argument has moved through four registers, and it is worth retracing them briefly, because the capstone theorem above compresses all of them into a single sentence that can be read too quickly. Part I established that Level-3, rule-level recursive modification — a system's own activity altering the rule governing its future behavior, not merely its state or its parameters — is neither rare nor recent. Cells, immune systems, languages, markets, and evolution by natural selection all instantiate it, and evolution does so without any cognition anywhere in the loop, which was the essay's first and most load-bearing result: whatever eventually turns out to matter about AGI's recursive self-modification, it cannot be the recursion itself, because recursion this old and this common cannot be the exceptional ingredient a genuinely novel danger would require.

Part II supplied the concept that recursion's ubiquity had been quietly missing: admissibility, not improvement, as the property that actually determines whether a recursive trajectory survives. A cancer cell lineage recursively modifies its own production rule and increases its own capability at every step, and is nonetheless a textbook failure by the only standard that was ever going to matter, because its trajectory exits the viability manifold of the organism containing it. The Complexity–Repair Theorem and its Repair Dominance corollary gave this a structural rather than merely definitional grounding: past a threshold set by the convexity of maintenance cost, repair capacity matters more than additional

capability, a result visible in cellular senescence, software technical debt, and imperial administrative overhead alike, none of which share anything but the underlying mathematical shape.

Part III asked when admissible-looking recursion is actually at risk of failing quietly, and derived, rather than merely asserted, an answer under a stated model: risk falls exponentially in effective, independent locus count and rises with the ratio of iteration speed to external verification speed. The word "effective" carries the section's real weight. Raw redundancy is not real redundancy once failure modes correlate, and the difference between ten independent institutions and ten copies of the same institution — quantified concretely via the design-effect correction for correlated channels — is the difference between a system that can catch its own errors and one that only appears to be able to.

Part IV supplied the case studies that make the preceding three parts more than an abstract exercise. Science's replication crisis is what a temporary collapse in effective diversity looks like when it happens to an otherwise well-functioning recursive system, and the field's methodological reforms since are a real-world instance of a system re-raising its own D_R after discovering the cost of having let it fall. The shift from scribal to printed textual transmission is what a civilizational system's diversity transition looks like when a technology changes the topology of verification rather than merely its speed. Neither example was constructed to fit this essay's formalism; both were selected because they already fit it.

Against that background, the question this essay was written to address — whether recursive self-improvement is the novel and singular threshold the contemporary AGI discussion treats it as — has a precise and now fully earned answer. It is not. Recursive self-improvement is not the arrival of a new principle in the history of complex systems; it is the possible acceleration and compression of an old one. Life, evolution, science, and civilization achieved recursive self-modification that remained admissible only by distributing it across many agents, timescales, and independent repair mechanisms. A recursively self-modifying system is not successful because it improves itself; it is successful only insofar as its modifications preserve the admissible conditions under which continuation, repair, and distinction-generation remain possible at all. The question for artificial general intelligence is not whether recursive self-improvement will arrive — by the argument of this essay, in the relevant structural sense, it already has, repeatedly, throughout the history of life and culture. The question, stated as precisely as Part III's formal apparatus allows and no more precisely than that, is whether it remains admissible once compressed into a substrate with few independent loci, a short iteration period relative to any external check, and low repair-path diversity — and that question, this essay has argued throughout, has nothing essential to do with how intelligent the substrate is.

Open Problems

This essay is deliberately left as a research program rather than a closed result, and the following questions are the ones its own formal apparatus makes visible without answering.

1. **The dynamics of H_t.** Sec. 8 treats recoverable history as a standing requirement for repair without developing how H_t itself degrades, is selectively retained, or can be reconstructed from partial traces. A full theory would need to specify H_t's own update rule and ask whether it is subject to the same admissibility constraints as everything built on top of it.

2. **The dynamics of diagnostic information I_t.** Sec. 10.2 bounds repair capacity by diagnostic resolution without modeling how I_t is produced, degraded, or how repair entropy S_R (Sec. 10.3) evolves as a system's component space grows.

3. **Measuring $\bar{\rho}$ in real systems.** Sec. 14.5 names three imperfect proxies for the exchangeable model's central parameter and shows that none is adequate alone. Producing an operational estimator for $\bar{\rho}$ in an actual, current system — the AI ecosystem being the obvious and most urgent case — remains unsolved by anything in this essay.

4. **The interaction of $\mathcal{A}_{\text{obj}}$ and $\mathcal{A}_{\text{proc}}$ under compression.** Sec. 7 distinguishes the two; Remark 13.1 flags that Part III's risk apparatus is agnostic between them; no result in this essay determines whether compression is more dangerous to one than to the other, or whether a system optimized for one kind of admissibility is thereby better or worse protected on the other.

5. **The relation between admissibility and truth.** Sec. 19.7 argues Truth $\subseteq$ Admissibility rather than identity, and gives a representational functional profile as the condition for equality. Whether that condition can be stated more precisely than "concerns representation," and whether it survives the objections Rahlwes and Dickson raise against non-propositional truth generally (Sec. 19.8), is not settled here.

6. **The relation between admissibility and distinguishability.** Proposition 10.10 now formalizes the immediate consequence — distinguishability loss bounds repair capacity downward — but this essay still has not developed a formal theory of what makes a distinction available to a system in the first place, why S_R increases in a given system, or how the cost of maintaining a distinction relates to the cost of maintaining admissibility more broadly beyond the single inequality Prop. 10.10 states. That fuller relationship, if it exists in the tight form Secs. 10 and 14 suggest it might, would connect this essay to a broader question about distinction and its preservation that lies outside this essay's scope.

None of these six problems is required for the essay's central argument to stand; the containment hierarchy of Part I, the admissibility results of Part II, and the compression results of Part III do not depend on any of them being resolved. They are recorded here because a theory that raises questions it did not need to raise, and says so plainly, is more useful to whoever picks it up next than a theory that quietly closes the door behind it.

References

[1] Humberto R. Maturana and Francisco J. Varela. *Autopoiesis and Cognition: The Realization of the Living*. D. Reidel, 1980.

[2] Francisco J. Varela, Humberto R. Maturana, and Ricardo Uribe. "Autopoiesis: The Organization of Living Systems, Its Characterization and a Model." *BioSystems*, 5(4), 1974.

[3] Beth Dempster. *Sympoietic and Autopoietic Systems: A New Distinction for Self-Organizing Systems*. Master's Thesis, University of Waterloo, 2000.

[4] Donna J. Haraway. *Staying with the Trouble: Making Kin in the Chthulucene*. Duke University Press, 2016.

[5] John H. Holland. *Adaptation in Natural and Artificial Systems*. MIT Press, 1992.

[6] John Maynard Smith and Eörs Szathmáry. *The Major Transitions in Evolution*. Oxford University Press, 1995.

[7] F. John Odling-Smee, Kevin N. Laland, and Marcus W. Feldman. *Niche Construction: The Neglected Process in Evolution*. Princeton University Press, 2003.

[8] Wassily Hoeffding. "Probability Inequalities for Sums of Bounded Random Variables." *Journal of the American Statistical Association*, 58(301), 1963.

[9] Bradley Efron and Trevor Hastie. *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. Cambridge University Press, 2016.

[10] Chris Rahlwes and Mark Dickson. "In Defense of Nonpropositional Truth." *The Philosophical Forum*, 2025.

[11] Nassim Nicholas Taleb. *Antifragile: Things That Gain from Disorder*. Random House, 2012.

[12] Scott E. Page. *The Diversity Bonus: How Great Teams Pay Off in the Knowledge Economy*. Princeton University Press, 2018.

[13] Erik Hollnagel. *Safety-I and Safety-II: The Past and Future of Safety Management*. Ashgate, 2015.

[14] W. Ross Ashby. *An Introduction to Cybernetics*. Chapman & Hall, 1956.

[15] Edwin Hutchins. *Cognition in the Wild*. MIT Press, 1995.

[16] Herbert A. Simon. "The Architecture of Complexity." *Proceedings of the American Philosophical Society*, 106(6), 1962, pp. 467–482.

[17] C. S. Holling. "Resilience and Stability of Ecological Systems." *Annual Review of Ecology and Systematics*, 4, 1973, pp. 1–23.

[18] Elinor Ostrom. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, 1990.