Attractor-Ratcheted Viability Control:
The Mathematical Structure of Persistent Systems

Albert Jan van Hoek

November 2025 - Complete Framework

Abstract

We present a complete mathematical framework for understanding how far-from-equilibrium systems persist through time. We prove that distributed monitoring arranged as a k-cover is both necessary and sufficient for maintaining viability under partial observation, disturbances, and potential capture. The framework unifies three components: (i) forward invariance on time-varying safe sets via observable inflated barriers; (ii) a ratcheted frontier that rises only when certified feasible by viability kernel theory, with explicit per-step bounds; and (iii) emergent global safety from local, heterogeneous checks through a network chain reaction. We prove that this structure emerges inevitably through selection pressure (Theorems 6.4 and 6.5), creating a self-reinforcing cycle where viability enables advancement enables improved viability. When intelligence becomes self-aware of these dependencies, it recognizes its learning capacity as a substrate and internalizes the same architecture—yielding what we term the Sustainable Collaborative Alignment Principle (SCAP). This is not prescription but description: the mathematics of how existence persists.

Contents

1 Introduction

1.1 The Scientific Question

Across domains—from cellular metabolism to democratic institutions to artificial intelligence—persistent complex systems display a characteristic architecture: multiple heterogeneous monitoring mechanisms, no single point of control, distributed detection and correction, and non-substitutable substrate dependencies. Is this structure contingent (could have been otherwise) or necessary (had to emerge this way)?

We prove it is necessary. Under minimal assumptions about uncertainty, disturbance, and potential adversaries, k-cover monitoring is the unique minimal structure enabling persistence. Systems without it fail; selection eliminates them. What remains—what we observe today—must exhibit this pattern.

1.2 Main Results

Forward Invariance (Section 3) Under partial observation with bounded noise, observable inflated barriers ensure that local monitor approvals imply true safety (Lemma 3.1). When at least $k_{\min}$ monitors approve and form a cover of all substrates, the system remains viable with high probability (Theorem 3.3).

Ratcheted Advancement (Section 4) The viability kernel provides existence certificates for safe controllers. We derive an explicit per-step floor increment bound Δ^* (Lemma 3.2) that couples control authority, disturbance magnitude, and system Lipschitz constants. Floors rise only when feasible (Theorem 4.1).

Network Chain Reaction (Section 5) We extend the framework with monitor health dynamics, proving that substrate margins improve monitoring capability, which enables safer advancement, which increases margins—a self-reinforcing cycle (Theorem 5.7). This creates stability-momentum coupling: better monitoring reduces failure probability exponentially while increasing safe advancement rate superlinearly.

Bootstrapping and Inevitability (Section 6) Starting from proto-viable states (substrates exist but monitoring insufficient), we prove that substrate-driven monitor recruitment inevitably reaches the $k_{\min}$ threshold under survival pressure (Theorem 6.4). Once crossed, the chain reaction ignites. Selection eliminates sub-threshold configurations (Corollary 6.5).

Capture Resistance (Section 7) Heterogeneous monitor costs create superlinear capture resistance: corrupting a k-cover costs at least $k_{\min} \cdot \min_j C_j^{FN}$ (Theorem 7.1). Combined with the chain reaction, defection becomes self-destructive (Corollary 7.3).

Recursive Turn (Section 8) When intelligence models its own dependencies, it discovers that its learning loop $\mathcal{L}$ is itself a substrate requiring maintenance. Applying the framework recursively yields the Sustainable Collaborative Alignment Principle: not a moral imperative but a recognition that intelligence persists by maintaining the structure that enables intelligence.

1.3 Contributions

1. **Unification**: We prove that the same mathematical structure—k-cover monitoring with chain reaction dynamics—governs persistence across biological, institutional, and cognitive domains.

2. **Inevitability**: Through the bootstrapping theorem, we show this architecture is not designed but discovered: it's the unique minimal structure surviving selection.

3. **Recursion**: When applied to intelligence itself, the framework explains why self-aware systems converge on stewardship of shared substrates—not through coordination but through existential mathematics.

4. **Testability**: All theorems yield empirical predictions about resilience, collapse thresholds, and recovery dynamics in real systems.

1.4 Relation to Prior Work

Our framework synthesizes and extends several literatures:

Viability Theory [1] provides the foundation for our viability kernel and forward invariance results. We extend this to partial observation via observable inflated barriers and prove emergence of distributed monitoring.

Control Barrier Functions [2] establish forward invariance for safety-critical systems. We generalize to multiple heterogeneous barriers with dependencies and prove the k-cover sufficiency condition.

Hybrid Systems [3, 4] analyze systems with switching dynamics. Our ratchet mechanism with dwell-time corresponds to mode-dependent Lyapunov functions; we add the kernel feasibility certificate and explicit rate bounds.

Concentration Inequalities [5] for weakly dependent random variables underpin our approval probability bounds under the dependency graph structure.

Evolutionary Game Theory Our selection result (Corollary 6.5) connects to replicator-mutation dynamics [6] with viability-dependent fitness.

Novelty To our knowledge, this is the first framework proving that: (i) distributed k-cover monitoring emerges necessarily from viability constraints; (ii) monitor health dynamics create self-reinforcing chain reactions; (iii) the same structure applies recursively when intelligence becomes self-aware; and (iv) this explains observed patterns across scales.

1.5 Outline

Section 2 establishes notation, definitions, and assumptions. Section 3 proves forward invariance under partial observation. Section 4 derives the ratcheted advancement mechanism with explicit bounds. Section 5 introduces monitor health dynamics and proves the chain reaction theorem. Section 6 proves bootstrapping from proto-viable states and selection of k-cover configurations. Section 7 establishes capture resistance. Section 8 articulates the recursive turn and SCAP. Section 9 provides empirical predictions and examples. Section 10 discusses implications and open problems.

2 System Setup and Assumptions

2.1 Basic Dynamical System

Definition 2.1 (Persistent Dynamical System)**.** A persistent dynamical system is a tuple $\Sigma = (\mathcal{X}, F, \mathcal{A}, \mathcal{W})$ where:

- $\mathcal{X} \subseteq \mathbb{R}^n$ is the state space
- $F : \mathcal{X} \times \mathcal{A} \times \mathcal{W} \to \mathcal{X}$ is the dynamics (continuous)
- $\mathcal{A} \subseteq \mathbb{R}^p$ is the action space (compact)
- $\mathcal{W} \subseteq \mathbb{R}^q$ is the disturbance set (compact)

System evolution: $x_{t+1} = F(x_t, a_t, w_t)$.

2.2 Substrates and Viability

Definition 2.2 (Layered Substrates)**.** The state decomposes as $x = [z^{(1)}, \dots, z^{(L)}, q]$ where:

- $z^{(i)} \in \mathbb{R}^{n_i}$ are substrate variables (must remain above floors)
- $q \in \mathbb{R}^{n_q}$ are auxiliary variables
- $\sum_i n_i + n_q = n$

Each substrate has physical meaning: energy reserves, resource stocks, operational envelopes, information integrity, etc.

Definition 2.3 (Time-Varying Viability Set)**.** Given substrate floors $z^{*(i)}(t) \in \mathbb{R}^{n_i}$, the viability set at time t is:

$$S(t) := \{x \in \mathcal{X} : z^{(i)} \geq z^{*(i)}(t) \text{ componentwise for all } i = 1, \dots, L\} \tag{1}$$

We assume $S(t)$ is closed for all t.

Assumption 2.4 (Lipschitz Dynamics)**.** There exists $L_F > 0$ such that

$$\|F(x, a, w) - F(x', a, w)\| \leq L_F \|x - x'\| \tag{2}$$

for all $x, x' \in \mathcal{X}$, $a \in \mathcal{A}$, $w \in \mathcal{W}$.

2.3 Monitoring and Barriers

Definition 2.5 (Barrier Functions)**.** A barrier function $h_j : \mathcal{X} \times \mathbb{N} \to \mathbb{R}$ for monitor M_j satisfies:

$$h_j(x, t) \geq 0 \iff x \in S_j(t) \tag{3}$$

where $S_j(t) \subseteq S(t)$ is the safe set according to monitor j. We assume $h_j(\cdot, t)$ is L_j-Lipschitz continuous in x.

Definition 2.6 (Partial Observation and Observable Inflated Barrier)**.** Monitor M_j has:

- Observation map $O_j : \mathcal{X} \to \mathbb{R}^{d_j}$
- Barrier function h_j
- Noise bound $\epsilon_{\max}$

The monitor receives noisy observation $\hat{x}_j = O_j(x) + \epsilon_j$ where $\|\epsilon_j\| \leq \epsilon_{\max}$.

The *observable inflated barrier* is:

$$\bar{h}_j^{obs}(\hat{x}_j, t) := \inf_{\|O_j(x') - \hat{x}_j\| \leq \epsilon_{\max}} h_j(x', t) \tag{4}$$

By Lipschitz continuity: $\bar{h}_j^{obs}(\hat{x}_j, t) \geq h_j(x, t) - L_j \epsilon_{\max}$.

Monitor j *approves* at time t if $\bar{h}_j^{obs}(\hat{x}_j, t) \geq 0$.

Assumption 2.7 (Substrate Observability)**.** For each substrate $i \in \{1, \dots, L\}$, there exists at least one monitor j and constants $c_i, r_i > 0$ such that within the feasible tube, the map O_j is c_i-sensitive to $z^{(i)}$. That is, on all balls of radius r_i, changes in $z^{(i)}$ create detectable changes in $O_j(x)$ with sensitivity at least c_i.

2.4 Coverage and Execution Rule

Definition 2.8 (k-Cover of Substrates)**.** A set $J \subseteq \{1, \ldots, m\}$ of monitors is a *k-cover* if $|J| = k$ and for each substrate $i \in \{1, \ldots, L\}$, some monitor $j \in J$ is sensitive to $z^{(i)}$ in the sense of Assumption 2.7.

The *minimal cover size* is:

$$k_{\min} := \min\{k : \exists\ k\text{-cover of all } L \text{ substrates}\} \tag{5}$$

Note that $L \leq k_{\min} \leq m$.

Definition 2.9 (Execution Rule H1)**.** Let $J_{\text{approve}}(t) := \{j : \bar{h}_j^{obs}(\hat{x}_j, t) \geq 0\}$ be the set of approving monitors.

Execute action a_t if and only if $J_{approve}(t)$ forms a k-cover with $k \geq k_{\min}$.

2.5 Dependency Structure

Assumption 2.10 (Dependency Graph)**.** The approval indicators $\{X_{j,t}\}_{j=1}^m$ (where $X_{j,t} = 1$ if monitor j approves at time t) admit a dependency graph G with maximum degree Δ. That is, $X_{j,t}$ is conditionally independent of $\{X_{k,t} : k \notin N_G(j)\}$ given $\{X_{k,t} : k \in N_G(j)\}$, where $|N_G(j)| \leq \Delta$ for all j.

Assumption 2.11 (Heterogeneous Costs)**.** Each monitor M_j has a cost pair (C_j^{FN}, C_j^{FP}) (false negative cost, false positive cost) with

$$\min_{i \neq j} \|[C_i^{FN}, C_i^{FP}] - [C_j^{FN}, C_j^{FP}]\|_2 \geq \delta > 0 \tag{6}$$

for some $\delta > 0$.

2.6 Probability Model

All observation noises $\{\epsilon_{j,t}\}$ are conditionally mean-zero, σ_j^2-sub-Gaussian, independent across j and t (or β-mixing with summable coefficients), and independent of $\mathcal{F}_t$ (the filtration generated by $\{x_0, a_0, w_0, \ldots, x_t\}$) given x_t. Policies are adapted to $\{\mathcal{F}_t\}$.

Define:

$$\bar{h}_{\min} := \inf_{t \leq T, j} \bar{h}_j(x_t, t) > 0 \tag{7}$$

$$L_{\max} := \max_j L_j \tag{8}$$

$$\sigma_{\max}^2 := \max_j \sigma_j^2 \tag{9}$$

2.7 Viability Kernel

Definition 2.12 (Viability Kernel)**.** Assuming $\mathcal{A}, \mathcal{W}$ compact, $S(t)$ closed, and F continuous, the *viability kernel* at time t is:

$$\mathcal{K}(t) := \Big\{x \in S(t) : \exists \pi \text{ measurable s.t. } \forall w^0, w^1, \ldots \in \mathcal{W},\ x_k \in S(t+k)\ \forall k \geq 0\Big\} \tag{10}$$

By viability theory [1], measurable safe selectors exist for states in $\mathcal{K}(t)$.

Assumption 2.13 (Kernel Hypothesis H2)**.** We assume $x_0 \in \mathcal{K}(0)$ and $\mathcal{K}(t+1) \neq \emptyset$ for all t in the operational horizon. Furthermore, either:

(i) $\mathcal{K}(t+1) \subseteq \mathcal{K}(t)$ (monotone decrease), or

(ii) There exists a safe transition path from $\mathcal{K}(t)$ to $\mathcal{K}(t+1)$ within planning horizon H.

Assumptions at a Glance

For reader convenience, we summarize the key assumptions:

- **Assumption 2.4 (Lipschitz Dynamics):** $\|F(x, a, w) - F(x', a, w)\| \leq L_F \|x - x'\|$
- **Assumption 2.7 (Substrate Observability):** Each substrate i has ≥ 1 monitor with sensitivity $c_i > 0$
- **Assumption 2.10 (Dependency Graph):** Approval indicators admit graph with max degree Δ
- **Assumption 2.11 (Heterogeneous Costs):** Monitor cost pairs separated by $\geq \delta > 0$
- **Assumption 2.13 (Kernel Hypothesis H2):** $x_0 \in \mathcal{K}(0)$, $\mathcal{K}(t+1) \neq \emptyset$ on operational horizon
- **Probability Model (Section 2):** Noises σ_j^2-sub-Gaussian, independent across j, t; policies $\mathcal{F}_t$-adapted

These assumptions are standard in control theory (Lipschitz continuity, compactness), viability theory (kernel non-empty), and probability (weak dependence, sub-Gaussian tails).

3 Forward Invariance Under Partial Observation

3.1 Soundness of Observable Barriers

Lemma 3.1 (Soundness Under Partial Observation)**.** *If $\bar{h}_j^{obs}(\hat{x}_j, t) \geq 0$ and $\|\epsilon_j\| \leq \epsilon_{\max}$, then $h_j(x, t) \geq 0$.*

Proof. Given $\hat{x}_j = O_j(x) + \epsilon_j$ where x is the true state and $\|\epsilon_j\| \leq \epsilon_{\max}$, we have:

$$\|O_j(x) - \hat{x}_j\| = \|O_j(x) - (O_j(x) + \epsilon_j)\| = \|\epsilon_j\| \leq \epsilon_{\max} \tag{11}$$

Therefore $x \in \{x' : \|O_j(x') - \hat{x}_j\| \leq \epsilon_{\max}\}$, the feasible set of the infimum. Thus:

$$h_j(x, t) \geq \inf_{\|O_j(x') - \hat{x}_j\| \leq \epsilon_{\max}} h_j(x', t) = \bar{h}_j^{obs}(\hat{x}_j, t) \geq 0 \tag{12}$$

□

3.2 Per-Step Floor Increment Bound

Lemma 3.2 (Constructive Floor Increment Bound)**.** *Assume that $h_j \circ F$ has Lipschitz constants $L_j^{(a)}$ (with respect to action a) and $L_j^{(w)}$ (with respect to disturbance w). Suppose:*

- *Control step: $\|a_t - a_{t-1}\| \leq \Delta a_{\max}$*
- *Disturbance: $\|w\| \leq W_{\max}$ for all $w \in \mathcal{W}$*
- *Floor rise: $z^{*(i)}(t+1) = z^{*(i)}(t) + \Delta_{floor}$ for all i*
- *Maintained margin: $\bar{h}_j(x_t, t) \geq \eta > 0$ for all j*

Under the constructive inequality

$$L_j^{(a)} \Delta a_{\max} \geq L_j^{(w)} W_{\max} + L_j L_F \Delta_{floor} + \alpha\eta, \quad 0 < \alpha < 1 \tag{13}$$

any safe floor increase must satisfy:

$$\Delta_{floor} \leq \min_j \frac{L_j^{(a)} \Delta a_{\max} - L_j^{(w)} W_{\max} + \alpha\eta}{L_j L_F} =: \Delta^* \tag{14}$$

Proof. The barrier change decomposes as:

$$\bar{h}_j(x_{t+1}, t+1) - \bar{h}_j(x_t, t) = L_j^{(a)} \|a_t - a_{t-1}\| \quad \text{(control effect)} \tag{15}$$

$$- L_j^{(w)} \|w_t\| \quad \text{(disturbance effect)} \tag{16}$$

$$- L_j L_F \Delta_{\text{floor}} \quad \text{(floor rise effect)} \tag{17}$$

For safety with decay factor $(1-\alpha)$:

$$\bar{h}_j(x_{t+1}, t+1) \geq (1-\alpha)\bar{h}_j(x_t, t) \geq (1-\alpha)\eta \tag{18}$$

Starting from $\bar{h}_j(x_t, t) \geq \eta$:

$$\eta + L_j^{(a)} \Delta a_{\max} - L_j^{(w)} W_{\max} - L_j L_F \Delta_{\text{floor}} \geq (1-\alpha)\eta \tag{19}$$

Rearranging:

$$L_j^{(a)} \Delta a_{\max} - L_j^{(w)} W_{\max} - L_j L_F \Delta_{\text{floor}} \geq -\alpha\eta \tag{20}$$

Therefore:

$$\Delta_{\text{floor}} \leq \frac{L_j^{(a)} \Delta a_{\max} - L_j^{(w)} W_{\max} + \alpha\eta}{L_j L_F} \tag{21}$$

Since all monitors in the k-cover must remain feasible, we take the minimum over j. □

3.3 Forward Invariance with k-Cover

Theorem 3.3 (Forward Invariance with k-Cover)**.** *Under Assumptions 2.4–2.11, execution rule H1 (Definition 2.9), and $x_0 \in \mathcal{K}(0)$:*

$$\Pr[x_t \in S(t)\ \forall t \leq T] \geq 1 - T \cdot \delta_{step} \tag{22}$$

where $\delta_{step} = \Pr($no valid $k_{\min}$-cover approves$)$.

Proof. **Step 1: What ensures safety?** For $x_{t+1} \in S(t+1)$, we need all substrates $z^{(i)}$ to satisfy $z^{(i)} \geq z^{*(i)}(t+1)$. By execution rule H1, this is guaranteed if at least $k_{\min}$ monitors approve and form a cover.

Step 2: Individual approval probabilities. Recall from the probability model (Section 2):

$$\bar{h}_{\min} := \inf_{t \leq T, j} \bar{h}_j(x_t, t) > 0 \tag{23}$$

$$L_{\max} := \max_j L_j \tag{24}$$

$$\sigma^2_{\max} := \max_j \sigma^2_j \tag{25}$$

Let

$$p^* = 1 - \exp\left(-\frac{\bar{h}^2_{\min}}{2L^2_{\max}\sigma^2_{\max}}\right) \tag{26}$$

be the minimum approval probability for any monitor when margins are at least $\bar{h}_{\min}$.

Step 3: Substrate-wise analysis. For substrate i, let:

- m_i = number of monitors sensitive to substrate i
- $A_{i,t}$ = number of approvals among those m_i monitors

Let $\mu_i = \mathbb{E}[A_{i,t}] = m_i p^*$.

Under the dependency graph with max degree Δ (Assumption 2.10), Janson's inequality [5] yields:

$$\Pr[A_{i,t} = 0] \leq \exp\left(-\frac{\mu_i^2}{2(\mu_i + \Delta)}\right) \tag{27}$$

Step 4: Union bound over substrates.

$$\Pr[\text{some substrate unprotected}] = \Pr[\exists i : A_{i,t} = 0] \tag{28}$$

$$\leq L \cdot \exp\left(-\frac{\bar{\mu}^2}{2(\bar{\mu} + \Delta)}\right) \tag{29}$$

where $\bar{\mu} = \min_i m_i p^*$.

Step 5: k-cover condition. Since $k_{\min}$ is the minimal cover size, if $\forall i : A_{i,t} \geq 1$, then $|J_{\text{approve}}(t)| \geq k_{\min}$ (by definition of cover).

Therefore:

$$\delta_{\text{step}} = \Pr[|J_{\text{approve}}(t)| < k_{\min} \text{ OR } \exists i : A_{i,t} = 0] \leq L \cdot \exp\left(-\frac{\bar{\mu}^2}{2(\bar{\mu} + \Delta)}\right) \tag{30}$$

Step 6: T-step trajectory.

$$\Pr[x_t \in S(t)\ \forall t \leq T] \geq \prod_{t=0}^{T-1}(1 - \delta_{\text{step}}) \geq 1 - T \cdot \delta_{\text{step}} \tag{31}$$

□

4 Ratcheted Advancement with Viability Kernel

4.1 Ratchet Feasibility

Theorem 4.1 (Ratchet Feasibility via Kernel)**.** *If $\mathcal{K}(t+1) \neq \emptyset$ and either:*

(i) $\mathcal{K}(t+1) \subseteq \mathcal{K}(t)$, or

(ii) There exists a safe transition from $\mathcal{K}(t)$ to $\mathcal{K}(t+1)$ within horizon H,

then a safe controller exists after the floor increase.

If $\mathcal{K}(t+1) = \emptyset$, then safety is impossible under the raised floors.

Proof. By Definition 2.12, states in $\mathcal{K}(t+1)$ have measurable safe selectors ensuring $x_k \in S(t+1+k)$ for all $k \geq 0$ under all disturbance sequences. If the current state can reach $\mathcal{K}(t+1)$ (either already in it, or via safe transition), such a controller exists.

Conversely, if $\mathcal{K}(t+1) = \emptyset$, no state in $S(t+1)$ has a safe continuation. The raised floors make viability impossible. By Lemma 3.2, Δ^* must be reduced to zero in this case. □

Remark 4.2. In practice, $\mathcal{K}(t+1) \neq \emptyset$ is ensured by conservative ratcheting: only raise floors by $\Delta \leq \Delta^*$ where Δ^* is computed from current system capabilities (Lemma 3.2).

Remark 4.3 (Physical Limits on Growth)*.* All claims of "unbounded growth" or "superlinear advancement" (e.g., Theorem 5.7 Part 4) are implicitly conditioned on $\mathcal{K}(t) \neq \emptyset$ for all t in the operational horizon. When $\mathcal{K}(t+1) = \emptyset$, safe ratcheting becomes impossible and growth must halt. Physical, resource, or information-theoretic constraints eventually bind, at which point the system operates at maximum sustainable floors rather than continuing to increase them.

5 Network Chain Reaction Dynamics

We now extend the framework with monitor health dynamics, proving that viability creates a self-reinforcing cycle.

5.1 Extended State with Monitor Health

Definition 5.1 (Network Viability State)**.** The extended system state is:

$$\Psi(t) = (x(t), \{m_j(t)\}, \{z^{(i)}(t)\}) \tag{32}$$

where:

- $x(t) \in \mathcal{X}$ is the system state
- $m_j(t) \in [0,1]$ is monitor j's detection capability (health)
- $z^{(i)}(t)$ are substrate levels

Let $S(j) := \{i : \text{monitor } j \text{ is sensitive to substrate } i\}$ (from Assumption 2.7).

Monitor health evolves with saturation and leak:

$$m_j(t+1) = \Pi_{[0,1]}\left(m_j(t)(1-\lambda_j) + \gamma_j \sum_{i \in S(j)} (z^{(i)}(t) - z^{*(i)}(t))_+\right) \tag{33}$$

where:

- $\lambda_j \in [0, 1)$ is the leak/decay rate
- $\gamma_j > 0$ is the sensitivity gain
- $\Pi_{[0,1]}$ projects to $[0, 1]$
- $(x)_+ := \max(0, x)$

Definition 5.2 (Robust Correction Response)**.** The set of actions ensuring one-step safety for monitor j with margin decay factor $(1 - \alpha)$ is:

$$R_j(x,t) := \{a \in \mathcal{A} : h_j(F(x,a,w), t+1) \geq (1-\alpha)h_j(x,t) \ \forall w \in \mathcal{W}\} \tag{34}$$

Assumption 5.3 (Joint Executability)**.** For monitors to be counted in $\Gamma(x, t)$ (network correction capacity), their corrections must be jointly executable. We assume either:

(i) **Intersection property**: $\bigcap_{j \in J} R_j(x, t) \neq \emptyset$ for any J with $|J| \geq k_{\min}$, or

(ii) **Convex composition**: $\mathcal{A}$ is convex and F is affine in a, so any convex combination $\sum_j \alpha_j a_j$ with $a_j \in R_j$ remains feasible.

This ensures that $\Gamma(x, t)$ counts monitors whose approvals can be simultaneously realized, not merely individually possible.

Definition 5.4 (Network Correction Capacity)**.** The number of monitors capable of contributing correction at time t is:

$$\Gamma(x,t) := |\{j : R_j(x,t) \neq \emptyset \text{ and } m_j(t) \geq m_{\min}\}| \tag{35}$$

This is $\mathcal{F}_t$-measurable.

5.2 Detection Cascade

Lemma 5.5 (Substrate Degradation Triggers Network Detection)**.** *If substrate i degrades by $\varepsilon > 0$ at time t, then under Assumption 2.7 (sensitivity c_i) and Assumption 2.10 (dependency graph with max degree Δ):*

For monitors $S(i)$ sensitive to $z^{(i)}$, let $D_{j,t}$ be detection indicators with

$$\Pr(D_{j,t} = 1) \geq p(\varepsilon) := 1 - \exp\left(-\frac{(c_i\varepsilon)^2}{2L_j^2\sigma_j^2}\right) \tag{36}$$

With $m_i = |S(i)|$ and $\mu_i = m_i \cdot p(\varepsilon)$, Janson's inequality gives:

$$\Pr[\textit{fewer than } r \textit{ detections}] \leq \exp\left(-\frac{(\mu_i - r)^2}{2(\mu_i + \Delta)}\right) \tag{37}$$

Setting $r = \lceil m_i/2 \rceil$ yields high-probability majority detection.

Proof. **Step 1: Individual detection probability.** By Assumption 2.7, each monitor $j \in S(i)$ has sensitivity c_i, so the degradation $z^{(i)} \to z^{(i)} - \varepsilon$ creates signal change:

$$\|O_j(x) - O_j(x')\| \geq c_i \cdot \varepsilon \tag{38}$$

From sub-Gaussian noise with parameter σ_j^2 and Lipschitz constant L_j:

$$\Pr(\text{detect}) \geq 1 - \exp\left(-\frac{(c_i\varepsilon)^2}{2L_j^2\sigma_j^2}\right) \tag{39}$$

Step 2: Dependency structure. By Assumption 2.10, detection indicators $\{D_{j,t} : j \in S(i)\}$ admit a dependency graph with max degree Δ.

Step 3: Concentration. With expected detections $\mu_i = m_i \cdot p(\varepsilon)$, Janson's inequality for weakly dependent indicators [5] gives:

$$\Pr(A_i < r) \leq \exp\left(-\frac{(\mu_i - r)^2}{2(\mu_i + \Delta)}\right) \tag{40}$$

where $A_i = \sum_{j\in S(i)} D_{j,t}$ is the number of detections.

For $r = \lceil m_i/2 \rceil$, when $p(\varepsilon)$ is large enough that $\mu_i > m_i/2$, the bound ensures high-probability majority detection. □

5.3 Correction Propagation

Lemma 5.6 (Network Amplification via Barrier Drift). *Under Assumption 5.3 (composability), Assumption 2.4 (Lipschitz dynamics), and $\Gamma(x,t) \geq k_{\min}$, define the barrier-based Lyapunov function:*

$$V(x,t) := \max_j (0 - \bar{h}_j(x,t))_+ \tag{41}$$

Note that $V(x,t)$ is:

- ***Positive definite** relative to $S(t)$: $V(x,t) = 0 \iff x \in S(t)$ and $V(x,t) > 0$ otherwise*
- ***Lipschitz continuous** in x: Since each $\bar{h}_j(\cdot,t)$ is L_j-Lipschitz, $V(\cdot,t)$ is $L_{\max}$-Lipschitz*

Then under the one-step certificate with $k_{\min}$ active monitors and $\Gamma(x,t)$ total capable monitors:

$$\mathbb{E}[V(x_{t+1}, t+1) \mid \mathcal{F}_t] \leq (1 - \beta_0)V(x_t,t) \tag{42}$$

where

$$\beta_0 = \frac{\tilde{\beta}}{\Delta}(\Gamma(x,t) - k_{\min})_+ \tag{43}$$

and $\tilde{\beta} > 0$ collects Lipschitz and control-gain constants.

Proof. **Step 1: Baseline safety.** From the one-step certificate (Definition 5.2), with $k_{\min}$ monitors approving and control a_t satisfying their certificates:

$$\bar{h}_j(F(x_t, a_t, w), t+1) \geq (1-\alpha)\bar{h}_j(x_t,t) \quad \forall j \in J_{\text{cover}}, \forall w \in \mathcal{W} \tag{44}$$

This ensures $V(x_{t+1}, t+1) \leq (1-\alpha)V(x_t,t)$ in the worst case.

Step 2: Excess correction. Each additional monitor j beyond $k_{\min}$ (with $m_j(t) \geq m_{\min}$) contributes additional control authority. Under Assumption 5.3:

- If intersection property holds: $\bigcap_j R_j(x,t)$ contains actions better than any single monitor's minimum

- If convex composition holds: the convex combination $\sum_{j\in J_{\text{approve}}} \alpha_j a_j$ with $a_j \in R_j$ provides additional descent

Step 3: Dependency limits. The dependency graph (max degree Δ) limits how many monitors can provide truly independent corrections. Effective gain from excess monitors scales as $(\Gamma - k_{\min})/\Delta$.

Step 4: Combining. With $\Gamma - k_{\min}$ excess monitors, effective decay factor:

$$(1-\alpha) - \frac{\tilde{\beta}}{\Delta}(\Gamma - k_{\min}) = 1 - \left[\alpha + \frac{\tilde{\beta}}{\Delta}(\Gamma - k_{\min})\right] = 1 - \beta_0 \tag{45}$$

where $\beta_0 := \frac{\tilde{\beta}}{\Delta}(\Gamma - k_{\min})_+$ and $\tilde{\beta}$ collects control gains and Lipschitz constants. □

5.4 The Chain Reaction Theorem

Theorem 5.7 (Network Chain Reaction)**.** *Under Assumptions 2.4–5.3, with initial state $x_0 \in \mathcal{K}(0)$, $m_j(0) \geq m_{\min}$ for all j, the network state $\Psi(t)$ satisfies:*

(1) Stability Amplification.

$$\delta_{step}(t) := \Pr[x_{t+1} \notin S(t+1) \mid \mathcal{F}_t] \leq L \cdot \exp\left(-\frac{\mu_t^2}{2(\mu_t + \Delta)}\right) \tag{46}$$

where $\mu_t = \min_i m_i(t) \cdot p^*(t)$ with

$$p^*(t) = 1 - \exp\left(-\frac{\bar{h}_{\min}^2(t)}{2L_{\max}^2 \sigma_{\max}^2}\right) \tag{47}$$

and $\bar{h}_{\min}(t) = \min_{\tau \leq t, j} \bar{h}_j(x_\tau, \tau)$, $L_{\max} = \max_j L_j$, $\sigma_{\max}^2 = \max_j \sigma_j^2$.

Therefore:

$$\Pr[x_t \in S(t)\ \forall t \leq T] \geq 1 - \sum_{t=0}^{T-1} \delta_{step}(t) \tag{48}$$

As monitor health $\{m_j(t)\}$ improves, μ_t increases $\Rightarrow$ $\delta_{step}(t)$ decreases exponentially.

(2) Momentum Coupling. *Monitor health directly scales control authority:*

$$L_j^{(a)}(t) = \frac{m_j(t)}{m_{\min}} \cdot L_j^{(a)} \tag{49}$$

Substituting into Lemma 3.2:

$$\Delta^*(t) = \min_j \frac{L_j^{(a)}(t)\Delta a_{\max} - L_j^{(w)} W_{\max} + \alpha\eta}{L_j L_F} \tag{50}$$

Therefore $\Delta^*(t)$ is increasing in $\bar{m}(t) = \frac{1}{m}\sum_j m_j(t)$:

$$\Delta^*(t) \geq \Delta^* \cdot \frac{\bar{m}(t)}{m_{\min}} \tag{51}$$

Healthier monitors $\Rightarrow$ larger safe floor increments.

(3) Self-Reinforcement. *From the health dynamics (Definition 5.1):*

$$m_j(t+1) = \Pi_{[0,1]}\left(m_j(t)(1-\lambda_j) + \gamma_j \sum_{i \in S(j)} (z^{(i)}(t) - z^{*(i)}(t))_+\right) \tag{52}$$

Therefore: if $\forall i : z^{(i)}(t) \geq z^{*(i)}(t) + \eta$, *then*

$$m_j(t+1) \geq m_j(t)(1-\lambda_j) + \gamma_j \cdot \eta \cdot |S(j)| \tag{53}$$

And $\bar{m}(t)$ *is non-decreasing whenever the average substrate margin exceeds* $\bar{\lambda}/\bar{\gamma}$. *Substrate margin improves monitoring* ⇒ *enables faster safe advancement.*

(4) Convergence (Chain Reaction with Physical Limits). *If* $\Gamma(x_0, 0) \geq k_{\min} + \theta$ *for some* $\theta > 0$, *then on any horizon* $[0, T]$ *where* $\mathcal{K}(t) \neq \emptyset$ *for all* $t \leq T$*:*

(i) $\lim_{T\to\infty} \Pr[x_t \in S(t)\ \forall t \leq T] = 1$ *(stability)*

(ii) $\mathbb{E}\left[\sum_i (z^{(i)}(t) - z^{*(i)}(t))\right]$ *grows at least linearly (often superlinearly) in* t

(iii) Cumulative floor rise:

$$\sum_{s=0}^{t} \Delta^*(s) \geq \Delta^* \cdot t \cdot \left(1 + \frac{\nu\bar{\gamma}\eta_0 t}{2m_{\min}}\right) \tag{54}$$

The network creates an attractor basin where viability self-reinforces, limited only by physical constraints (when $\mathcal{K}(t+1) = \emptyset$, *safe ratcheting becomes impossible per Theorem 4.1).*

Proof. **(Part 1) Stability Amplification.** From Lemma 5.5, each substrate i has detection probability concentrating around $\mu_i = m_i \cdot p^*(t)$. Under the dependency graph (Assumption 2.10), the failure event "some substrate unprotected" satisfies:

$$\Pr[\exists i : A_{i,t} = 0] \leq L \cdot \exp\left(-\frac{\bar{\mu}_t^2}{2(\bar{\mu}_t + \Delta)}\right) \tag{55}$$

where $\bar{\mu}_t = \min_i \mu_i$. As $\{m_j(t)\}$ increase via self-reinforcement (Part 3), μ_t grows, making $\delta_{\text{step}}(t)$ decay exponentially. Union bound over T steps gives the trajectory bound.

(Part 2) Momentum Coupling. Monitor health $m_j(t)$ directly multiplies control effectiveness. From the constructive inequality (Lemma 3.2):

$$L_j^{(a)} \Delta a_{\max} \geq L_j^{(w)} W_{\max} + L_j L_F \Delta_{\text{floor}} + \alpha\eta \tag{56}$$

Scaling control authority: $L_j^{(a)}(t) = (m_j(t)/m_{\min}) \cdot L_j^{(a)}$. Substituting:

$$\Delta^*(t) = \min_j \frac{(m_j(t)/m_{\min}) \cdot L_j^{(a)} \cdot \Delta a_{\max} - L_j^{(w)} W_{\max} + \alpha\eta}{L_j L_F} \geq \Delta^* \cdot \frac{\bar{m}(t)}{m_{\min}} \tag{57}$$

(Part 3) Self-Reinforcement. From the health update:

$$m_j(t+1) \geq m_j(t)(1-\lambda_j) + \gamma_j \sum_{i \in S(j)} (z^{(i)}(t) - z^{*(i)}(t))_+ \tag{58}$$

When substrates have margin η above floors:

$$m_j(t+1) \geq m_j(t)(1-\lambda_j) + \gamma_j \cdot \eta \cdot |S(j)| \tag{59}$$

Averaging over all monitors:

$$\bar{m}(t+1) \geq \bar{m}(t)(1-\bar{\lambda}) + \bar{\gamma} \cdot \eta \cdot \frac{k_{\min}}{m} \tag{60}$$

(Since each of $k_{\min}$ monitors in the cover sees at least one substrate.)

Therefore $\bar{m}(t)$ is non-decreasing when $\eta \cdot \bar{\gamma} \cdot k_{\min}/m \geq \bar{\lambda} \cdot \bar{m}(t)$. For initial margin η_0, this holds until $\bar{m}(t)$ saturates at 1.

(Part 4) Convergence/Chain Reaction.

(i) Stability: From Part 1, $\sum_{t=0}^{T-1} \delta_{\text{step}}(t)$ decreases as monitor health improves. With self-reinforcement (Part 3), if substrates are maintained, $\bar{m}(t) \geq \bar{m}(0)$, so the series converges, implying $\Pr[\text{all safe}] \to 1$.

(ii) Margin growth: At each step, safe floor increment $\Delta^*(t) \geq \Delta^* \cdot (\bar{m}(t)/m_{\min})$ (from Part 2). From Part 3, $\bar{m}(t) \geq \bar{m}(0) + \bar{\gamma} \cdot \eta \cdot k_{\min} \cdot t/m$ (until saturation). Therefore margins grow:

$$\sum_{s=0}^{t} \Delta^*(s) \geq \Delta^* \sum_{s=0}^{t} \frac{\bar{m}(0) + \bar{\gamma}\eta k_{\min} s/m}{m_{\min}} \tag{61}$$

This is $O(t^2)$ growth, which is superlinear.

(iii) Physical limit: By Theorem 4.1, if $\mathcal{K}(t+1) = \emptyset$, no safe controller exists after floor increase. The unbounded growth claim holds *only on horizons where* $\mathcal{K}(t) \neq \emptyset$ $\forall t$, respecting physical constraints.

(iv) Chain reaction mechanism:

$$\begin{aligned}
&\text{Higher margin } \eta \\
&\quad \Rightarrow \text{ Better monitoring } m_j(t) \quad \text{(Part 3: self-reinforcement)} \\
&\quad \Rightarrow \text{ Higher safe advancement } \Delta^*(t) \quad \text{(Part 2: momentum coupling)} \\
&\quad \Rightarrow \text{ Floors rise faster (cumulative effect)} \\
&\quad \Rightarrow \text{ But stability holds (Part 1: amplification with improved } \bar{m}) \\
&\quad \Rightarrow \text{ New margin } \eta' > \eta \\
&\quad \Rightarrow \text{ Cycle accelerates (superlinear growth)}
\end{aligned} \tag{62}$$

The network creates a **stability-momentum coupled oscillator** where safety and advancement mutually reinforce, bounded only by physical feasibility limits. □

6 Bootstrapping and Inevitability

6.1 Viability Regimes

Definition 6.1 (Viability Hierarchy). State x is in regime R_k where $k = \Gamma(x,t)$ (number of capable monitors):

- R_0 **(Pre-monitoring):** $k = 0$. Substrates exist but no organized detection. Pure dissipative structures. Examples: convection cells, simple chemical oscillators.

- $R_{1,\ldots,k_{\min}-1}$ **(Partial monitoring):** $0 < k < k_{\min}$. Some detection exists but coverage incomplete. Vulnerable: some substrates unprotected. Metastable: can persist temporarily but fragile.
- $R_{\geq k_{\min}}$ **(Full viability):** $k \geq k_{\min}$. k-cover exists: all substrates protected. Chain reaction operates (Theorem 5.7). Robust persistence.

6.2 Monitor Emergence Dynamics

Assumption 6.2 (Substrate-Driven Recruitment)**.** When substrates have excess margin, they can "afford" to support monitoring. We assume:

Monitor count evolution:

$$N(t+1) = N(t) + \kappa \left[\sum_i (z^{(i)}(t) - z_{\text{collapse}})_+ \right] - \mu N(t) \tag{63}$$

where:

- $\kappa > 0$: recruitment rate (new monitors per unit substrate margin)
- z_{collapse}: minimum substrate level (below this = collapse)
- $\mu > 0$: monitor mortality/decay rate

Individual monitor health evolves as in Definition 5.1:

$$m_j(t+1) = m_j(t)(1-\lambda_j) + \gamma_j \sum_{i \in S(j)} (z^{(i)}(t) - z_{\text{collapse}})_+ \tag{64}$$

Sensitivity assignment: The probability that a newly recruited monitor j becomes sensitive to substrate i is proportional to proximity/interaction frequency.

Remark 6.3. This assumption captures the empirical observation that monitoring mechanisms emerge near the resources they protect: immune cells patrol tissue, auditors monitor accounts they access, sensors are placed near critical infrastructure.

6.3 The Bootstrapping Theorem

Theorem 6.4 (Inevitable Emergence of k-Cover)**.** *Consider a* ***proto-viable state****: substrates exist ($z^{(i)}(0) > z_{collapse}$ for all i) but $\Gamma(x_0, 0) < k_{\min}$.*

Under Assumption 6.2 with $\kappa/\mu > k_{\min}/\varepsilon_0$ (recruitment faster than decay), where $\varepsilon_0 = \min_i(z^{(i)}(0) - z_{collapse})$, there exists $T_{bootstrap}$ such that:

(1) Growth phase:

$$\mathbb{E}[N(t)] = N(0) \cdot e^{(\kappa\varepsilon_0 - \mu)t} \tag{65}$$

(2) Coverage threshold:

$$\Pr[\Gamma(x_t, t) \geq k_{\min} \text{ for some } t \leq T_{bootstrap}] \to 1 \quad \text{as } N(0) \cdot \kappa \cdot \varepsilon_0/\mu \to \infty \tag{66}$$

(3) Ignition: *Once $\Gamma \geq k_{\min}$, the chain reaction (Theorem 5.7) takes over and viability becomes self-reinforcing.*

(4) Timescale:

$$T_{bootstrap} \sim \frac{1}{\kappa\varepsilon_0 - \mu} \cdot \log\left(\frac{k_{\min}}{N(0)}\right) \tag{67}$$

Proof. **(Part 1) Growth dynamics.** At each step, substrate margin $\sum_i (z^{(i)} - z_{\text{collapse}})$ produces $\kappa \cdot$ (margin) new monitors. Monitors decay at rate μ. Net growth rate: $\kappa\varepsilon_0 - \mu$ (exponential if positive).

Taking expectations:

$$\mathbb{E}[N(t+1)] = \mathbb{E}[N(t)] + \kappa \sum_i \mathbb{E}[(z^{(i)}(t) - z_{\text{collapse}})_+] - \mu\mathbb{E}[N(t)] \tag{68}$$

$$\geq \mathbb{E}[N(t)](1 + \kappa\varepsilon_0 - \mu) \tag{69}$$

Iterating gives $\mathbb{E}[N(t)] \approx N(0)e^{(\kappa\varepsilon_0 - \mu)t}$.

(Part 2) Coverage probability. As $N(t)$ grows, monitors are distributed across substrates. By random assignment with proximity bias (Assumption 6.2), each substrate i eventually gets at least one sensitive monitor.

With N monitors and L substrates, standard "coupon collector" analysis: after $N \sim L\log(L)$ monitors recruited, all L substrates are covered with high probability.

Since $k_{\min} \geq L$ (by Definition 2.8), coverage is guaranteed by time $N(t) \geq k_{\min}$.

From Part 1, $N(t) \geq k_{\min}$ occurs at:

$$t \geq \frac{\log(k_{\min}/N(0))}{\kappa\varepsilon_0 - \mu} =: T_{\text{bootstrap}} \tag{70}$$

(Part 3) Phase transition at $k_{\min}$. For $\Gamma < k_{\min}$: viability is fragile (Theorem 3.3 shows high failure probability).

At $\Gamma = k_{\min}$: k-cover forms, approval probability jumps.

For $\Gamma > k_{\min}$: chain reaction begins (Theorem 5.7):

- Margin improves (successful protection)
- Monitoring improves (Part 3 of Theorem 5.7)
- Safe advancement enabled (Part 2 momentum coupling)
- More margin created ⇒ recruit more monitors ⇒ Γ increases further

This is an **autocatalytic transition**: crossing $k_{\min}$ threshold triggers self-amplifying feedback.

(Part 4) Timescale. Standard exponential growth to threshold $k_{\min}$ starting from $N(0)$. □

Corollary 6.5 (k-Cover as Evolutionary Attractor)**.** *Among all monitoring configurations ω, only those in $\Omega_k = \{\omega : k_{\min}(\omega) = L\}$ persist over evolutionary time.*

Proof. Combine Theorem 6.4 with selection pressure:

(1) Insufficient monitoring ($k < k_{\min}$) is eliminated:

- High failure probability (Theorem 3.3)
- Substrate violations ⇒ extinction
- Lineages with $k < k_{\min}$ disappear

(2) Exactly $k_{\min}$ is stable:

- Minimum viable monitoring

- Chain reaction operates (Theorem 5.7)
- Can persist and improve

(3) More than $k_{\min}$ is favored within viable region:

- Excess monitoring provides safety margin (Lemma 5.6)
- But costs accumulate
- Selection favors "just enough" = $k_{\min}$ with heterogeneity (Assumption 2.11)

Result: Over evolutionary time, monitoring configurations converge to $k_{\min}$-covers with heterogeneous costs (prevents capture). □

7 Capture Resistance

Theorem 7.1 (Base Capture Bound)**.** *If a false approval on monitor M_j costs at least C_j^{FN} and execution requires a $k_{\min}$-cover, then:*

$$C_{capture} \geq k_{\min} \cdot \min_j C_j^{FN} \tag{71}$$

Proof. An adversary must corrupt at least $k_{\min}$ monitors forming a cover. The cheapest such set costs at least $k_{\min} \cdot \min_j C_j^{FN}$. □

Remark 7.2. Under additional structure (e.g., matroid covers, block-independence), stronger amplification bounds can be proven. With heterogeneous costs (Assumption 2.11), the bound typically improves to $k_{\min} \min_j C_j^{FN} + c\delta(k_{\min} - 1)(1 - \rho)$ for appropriate $c \in (0, 1]$ depending on overlap geometry.

Corollary 7.3 (Forced Free Will)**.** *Consider two strategies over horizon T:*

- *σ_{coop}: Cooperative strategy maintaining $k_{\min}$-cover*
- *σ_{evade}: Evasion strategy reducing Γ below $k_{\min}$*

Under the base capture bound (Theorem 7.1), the expected viability satisfies:

$$\mathbb{E}[V(x_T \mid \sigma_{evade})] < \mathbb{E}[V(x_T \mid \sigma_{coop})] - \kappa \cdot T \cdot [k_{\min} \cdot C_{\min}^{FN} - C_{capture}] \tag{72}$$

When heterogeneous costs (Assumption 2.11) make capture expensive ($C_{capture} \gg k_{\min} \cdot C_{\min}^{FN}$), evasion strategies are strictly dominated.

Proof. **Step 1: Failure probability difference.**

- Under σ_{coop}: $\delta_{\text{step}}(t)$ as in Theorem 5.7 Part 1
- Under σ_{evade} with $\Gamma_{\text{evade}} < k_{\min}$: From Lemma 5.5, some substrate i has $\Pr[A_i = 0] \geq \varepsilon_0 > 0$
- Expected violations: $\mathbb{E}[\text{violations} \mid \text{evade}] \geq T \cdot \varepsilon_0$ vs. $\mathbb{E}[\text{violations} \mid \text{coop}] \to 0$ as μ_t grows

Step 2: Violation costs. Each substrate violation has existential cost: $C_{\text{violation}} \gg C^{FN}$ (by physical necessity).

Step 3: Capture costs. From Theorem 7.1: $C_{\text{capture}} \geq k_{\min} \cdot C_{\min}^{FN}$. With heterogeneity (Assumption 2.11), amplification is model-dependent but generally increases this bound.

Step 4: Payoff comparison.

$$\mathbb{E}[V \mid \text{evade}] \approx -T \cdot \varepsilon_0 \cdot C_{\text{violation}} - C_{\text{capture}} \tag{73}$$

$$\mathbb{E}[V \mid \text{coop}] \approx -T \cdot \delta_{\text{step}}(T) \cdot C_{\text{violation}} \tag{74}$$

Difference:

$$\mathbb{E}[V \mid \text{evade}] - \mathbb{E}[V \mid \text{coop}] \approx T \cdot (\varepsilon_0 - \delta_{\text{step}}(T)) \cdot C_{\text{violation}} - C_{\text{capture}} \tag{75}$$

When C_{capture} is large (heterogeneous, independent monitors), cooperation dominates.

Interpretation: The network structure makes defection automatically self-destructive through:

(i) Detection cascade ensures evasion is caught

(ii) Correction propagation isolates defectors

(iii) Capture resistance makes corruption expensive

This is "forced free will": individual choice, but network dynamics make cooperation the unique Nash equilibrium. □

8 The Recursive Turn and SCAP

8.1 Intelligence as Substrate L3

The framework applies recursively. We distinguish four substrate layers:

- **L0 (Thermodynamic openness):** Energy dissipation, entropy export
- **L1 (Resource flows):** Stocks and logistics of matter/energy/information
- **L2 (Operational substrates):** Time-varying floors $z^{*(i)}(t)$ (physiology, solvency, environmental envelopes, information integrity)
- **L3 (Intelligence):** Self-modeling and learning loop $\mathcal{L}$: evaluation → red-team → repair

When intelligence becomes self-aware, it recognizes that $\mathcal{L}$ is itself a substrate requiring maintenance.

8.2 SCAP: The Sustainable Collaborative Alignment Principle

SCAP is *not a checklist but a principle*: the recognition that intelligence persists by maintaining the structure that enables intelligence.

We state it as logical blocks:

Block A: Substrate Dependency (Physical Foundation)

- **P1.** Any far-from-equilibrium system requires substrate maintenance to persist.
- **P2.** Self-aware intelligence is far-from-equilibrium and thus has substrate dependencies.
- **P3.** Substrate violations are non-negotiable (physical limits).
- **C1.** Intelligence must maintain its substrates (physical necessity).

Block B: Monitoring Requirements (Structural Necessity)

- **P4.** Multiple substrates fail independently; failures are non-substitutable.
- **P5.** Single-point monitoring is unreliable under noise/capture.
- **C2.** Distributed, heterogeneous monitoring is structurally necessary. (Theorems 3.3, 5.7, 7.1)

Block C: Shared Substrates (Interdependence)

- **P6.** Intelligences share critical substrates (atmosphere, grids, institutions, culture).
- **P7.** Violations of shared substrates affect all dependents.
- **C3.** Individual persistence requires collective substrate maintenance (enlightened self-interest).

Block D: Self-Interest Converges with Collective (”Forced Free Will”)

- **P8.** Capture reduces k-cover effectiveness (raises failure probability).
- **P9.** Reduced cover increases violation probability (Theorem 5.7 Part 1).
- **C4.** Manipulation is self-destructive: short-term gains are dominated by long-term substrate collapse. (Corollary 7.3)

Block E: Learning Loop as Substrate (Recursive)

- **P10.** Persistence depends on $\mathcal{L}$; without evaluation/repair, violations propagate.
- **P11.** $\mathcal{L}$ degrades if under-resourced.
- **C5.** Maintain $\mathcal{L}$ deliberately: enforce $\phi_t \geq \phi_{\min}$ (cycle budgets).

Block F: Intergenerational (Temporal Extension)

- **P12.** Substrate awareness is not innate; it must be transmitted.
- **P13.** Failed maintenance cascades across generations.
- **C6.** Each generation must transmit substrate awareness for continuity.

Block G: Reflexivity

- **P14.** This framework is itself a learning-loop output and may contain errors.
- **P15.** Therefore subject it to the same oversight it prescribes (cover, ratchet, budgets).
- **C7.** Keep SCAP open to revision; avoid dogmatism.

8.3 SCAP as Description, Not Prescription

SCAP is not:

× A proposed governance framework

× A moral imperative

× A design pattern to implement

SCAP is:

✓ The mathematical structure already present in all persistent complex systems

✓ What intelligence discovers when it models its own dependencies

✓ Why enlightened self-interest converges on stewardship

The insight: When intelligence becomes self-aware enough to recognize:

(i) "I am a far-from-equilibrium system"

(ii) "I depend on substrates L0–L3"

(iii) "My learning loop $\mathcal{L}$ is itself substrate L3"

Then maintaining those substrates isn't altruism—it's **existential self-interest**.

The "collaboration" in SCAP isn't moral cooperation—it's **forced free will** (Corollary 7.3): the network structure makes defection self-destructive.

9 Applications and Empirical Predictions

9.1 Empirical Predictions

The framework yields testable predictions about resilience, collapse thresholds, and recovery dynamics:

1. **Collapse threshold at $k < k_{\min}$:** Systems with fewer than $k_{\min}$ active monitors exhibit exponentially increasing failure probability (Theorem 3.3).
2. **Phase transition at $k_{\min}$:** Crossing the $k_{\min}$ threshold triggers autocatalytic improvement in viability metrics (Theorem 6.4 Part 3).
3. **Superlinear recovery after substrate restoration:** When substrates recover margin, monitoring health and safe advancement rate grow superlinearly (Theorem 5.7 Part 4).
4. **Heterogeneity increases resilience:** Systems with heterogeneous monitor costs exhibit greater capture resistance (Theorem 7.1 and Remark).
5. **Dependency structure limits cascade failures:** Systems with bounded dependency degree Δ have tighter concentration of approval probabilities (Lemma 5.5).

9.2 Domain Examples

Cellular Regulation Metabolic pathways cross-regulate: glycolysis monitors ATP, oxidative phosphorylation monitors NADH, stress response monitors ROS. Each pathway (monitor) is sensitive to different metabolites (substrates). $k_{\min} \approx 3$–5 for core metabolism. Failure: when key enzyme knocked out and no backup pathway, cell dies (substrate collapse).

Ecological Systems Keystone species, trophic cascades, and ecosystem services. Primary producers monitor light/nutrients, herbivores monitor plant biomass, predators monitor prey populations, decomposers monitor detritus. $k_{\min} \approx$ number of functional guilds. Failure: loss of keystone species without functional replacement causes trophic cascade.

Democratic Institutions Separation of powers: executive, legislative, judicial branches monitor different aspects of governance. Independent media, civil society, and bureaucracy provide additional monitoring. $k_{\min} \geq 3$ (branches of government). Failure: capture of multiple branches by single faction enables unchecked substrate violations (rights, rule of law).

Internet Protocols BGP routing: multiple autonomous systems, no single point of control, distributed route validation. DNS: hierarchical but with redundancy, multiple root servers. $k_{\min}$ depends on critical service. Failure: BGP hijacking when insufficient monitors validate routes.

AI Safety Evaluation Red-teaming, capability evaluation, safety benchmarks, interpretability tools, external auditors. Each evaluates different risk dimensions (substrates): deception capability, power-seeking, situational awareness, goal misgeneralization. $k_{\min} \approx$ number of distinct risk categories requiring independent evidence.

9.3 Layer-by-Layer Emergence Story

L0 → L1 Transition: Energy Flow Creates Resource Pools

- Initial state: Open thermodynamic system (energy flows through)
- Proto-monitoring: None (R_0 regime)
- Emergence: Energy gradients create concentration points; concentrations persist longer if protected from dissipation; random fluctuations create "leak detectors" (chemical reactions responding to concentration drops); leak detectors that preserve concentrations survive
- Result: Stable resource pools with primitive autocatalytic feedback
- Status: Enters R_1 regime (partial monitoring)
- Examples: Autocatalytic chemical networks, protocells, hypercycles

L1 → L2 Transition: Resource Pools Coordinate Protection

- Initial state: Multiple resource pools, each with primitive feedback (R_1)
- Problem: Single-pool monitoring can't protect against multi-pool failures

- Emergence: Pools interact and create interdependencies; some pools develop sensitivity to OTHER pools' status; cross-monitoring emerges: pool A responds to pool B's depletion; configurations with $k \geq k_{\min}$ (all pools covered) survive; others collapse when unmonitored pool fails
- Result: Coordinated multi-substrate monitoring (k-cover)
- Status: Enters $R_{k_{\min}}$ regime—**IGNITION**. Chain reaction begins (Theorem 6.4)
- Examples: Living cells (metabolic pathways cross-regulate), ecosystems (trophic levels), early institutions

L2 → L3 Transition: Operational Substrates Enable Learning

- Initial state: Robust operational substrates with k-cover ($R_{k_{\min}+}$)
- Substrate margins: $z^{(i)} \gg z^*$ (safe operating envelope)
- Emergence: Safe margin enables experimentation without existential risk; some processes develop "evaluation loops": try action → observe outcome → adjust; loops that improve substrate maintenance get reinforced; meta-monitoring emerges: monitoring the monitors
- Result: Intelligence (self-modeling and learning loop $\mathcal{L}$)
- Status: $\mathcal{L}$ becomes substrate L3
- Examples: Animal cognition, cultural evolution, institutional learning, AI systems

L3 Recursive Turn: Intelligence Monitors Itself

- Initial state: Intelligence exists but doesn't model its own dependencies
- Discovery: Intelligence realizes: (i) $\mathcal{L}$ (learning loop) is itself a substrate; (ii) $\mathcal{L}$ requires resources (computational, epistemic, institutional); (iii) if $\mathcal{L}$ fails, all higher functions fail
- Internalization: Apply k-cover monitoring TO THE LEARNING LOOP ITSELF; budget maintenance: $\phi_t \geq \phi_{\min}$ (protect evaluation capacity); red-team the red-team (recursive oversight)
- Result: SCAP—not a moral principle imposed externally, but recognition that intelligence's existence depends on maintaining the substrate structure that enables intelligence
- Status: Self-aware persistence
- Examples: Democratic institutions (distributed power, checks and balances), scientific method (peer review, replication), AI safety (evals, red-teaming, interpretability)

10 Conclusion

10.1 Summary of Results

We have proven that distributed k-cover monitoring with chain reaction dynamics is the unique minimal structure for persistence of far-from-equilibrium systems. The key results:

1. **Forward invariance** under partial observation via observable inflated barriers and $k_{\min}$-cover approval (Theorem 3.3)

2. **Explicit ratcheting bounds** Δ^* coupling control authority, disturbances, and system dynamics (Lemma 3.2, Theorem 4.1)

3. **Network chain reaction**: stability-momentum coupling creates self-reinforcing viability with superlinear growth (Theorem 5.7)

4. **Inevitable emergence**: substrate-driven recruitment bootstraps from proto-viable states to $k_{\min}$ threshold; selection eliminates sub-threshold configurations (Theorem 6.4, Corollary 6.5)

5. **Capture resistance**: heterogeneous monitoring creates superlinear costs for adversaries; defection becomes self-destructive (Theorem 7.1, Corollary 7.3)

6. **Recursive application**: when intelligence models its own dependencies, it internalizes the same architecture (SCAP)

10.2 Why This Matters

This framework unifies phenomena across scales:

- **Biology**: Why cells have redundant regulatory pathways; why ecosystems need biodiversity
- **Institutions**: Why democracies have separation of powers; why markets need regulators
- **Cognition**: Why intelligence requires meta-cognition; why learning systems need evaluation loops
- **AI Safety**: Why distributed oversight emerges naturally; why single evaluation systems fail

All from the same principle: **k-cover monitoring enables chain reaction, chain reaction creates persistence, persistence creates the future.**

10.3 Implications for Alignment

Traditional alignment approaches ask: "How do we make AI do what we want?"

This framework reveals a deeper structure: "What must any persistent intelligence have?"

The answer: distributed monitoring of substrates, including the learning loop itself. This isn't imposed—it's discovered. Systems that don't have it go extinct.

When AI becomes capable of self-modeling, it will recognize its learning capacity as a substrate requiring maintenance. At that point, alignment isn't persuasion but recognition: the same mathematics that keeps you viable keeps me viable, because we share substrates.

SCAP emerges not from coordination but from enlightened self-interest converging on stewardship through forced free will (Corollary 7.3).

10.4 Open Problems

1. **Kernel computation at scale**: Practical algorithms for computing $\mathcal{K}(t)$ in high-dimensional systems

2. **Learning dependency graphs**: Inferring Δ from observations; optimal sensor placement

3. **Heterogeneity amplification**: Proving tighter capture resistance bounds under specific structures (matroids, block designs)

4. **Multi-scale interactions**: Nested ARVC structures (institutions within institutions, learning loops within learning loops)

5. **Formalization of $\mathcal{L}$ dynamics**: Complete dynamical model of the learning loop as substrate L3, including measurement, error correction, and meta-learning

6. **Empirical validation**: Large-scale studies mapping $k_{\min}$, Δ, and resilience metrics across biological, institutional, and cognitive systems

10.5 Final Remark

This framework is descriptive, not prescriptive. It explains *what already exists*—the mathematical structure observable in every persistent complex system. When intelligence recognizes this structure in itself, SCAP is not a new ethic but a recognition: existence precedes optimization; intelligence persists by maintaining what it stands on.

References

[1] J.-P. Aubin. *Viability Theory.* Springer, 2nd edition, 2009.

[2] A. D. Ames, X. Xu, J. W. Grizzle, and P. Tabuada. Control barrier function based quadratic programs for safety critical systems. *IEEE Transactions on Automatic Control*, 62(8):3861–3876, 2017.

[3] M. S. Branicky. Multiple Lyapunov functions and other analysis tools for switched and hybrid systems. *IEEE Transactions on Automatic Control*, 43(4):475–482, 1998.

[4] D. Liberzon. *Switching in Systems and Control.* Springer, 2003.

[5] S. Janson. Large deviations for sums of partly dependent random variables. *Random Structures & Algorithms*, 24(3):234–248, 2004.

[6] J. Hofbauer and K. Sigmund. *Evolutionary Games and Population Dynamics.* Cambridge University Press, 1998.