From Biological Analogy to Formal Control A Cybernetic and Active Inference Framework for Self-Regulating Socio-Technical Systems

J.Konstapel Leiden, Netherlands
Date: March 27 2026
Keywords: Active Inference, Free Energy Principle, Viable Systems Model, Socio-Technical Systems, Autopoiesis, Semantic Evolution, Cybernetic Control

Abstract

Contemporary socio-technical platforms are designed as aggregates of modular services governed by external management logic. They lack intrinsic regulatory capacity: their behavior is controlled from outside, not produced from within. SWARP proposes an alternative: a platform architecture modeled not on the machine but on the living organism, drawing on the Free Energy Principle (Friston, 2010), the Viable Systems Model (Beer, 1972), and autopoietic theory (Maturana & Varela, 1980).

Prior work established a structural isomorphism between biological subsystems and SWARP modules — a mapping of considerable explanatory power. However, structural analogy alone does not produce a viable system. A system becomes genuinely self-regulating only when it is governed by *enforceable constraints*: formally specified rules that determine state transitions, bound complexity, resolve conflicts, and close feedback loops.

This paper formalizes those constraints. We derive, prove, and specify five control mechanisms: (1) lifecycle regulation through apoptotic state transitions; (2) dual-mode immune response differentiating acute from chronic disturbance; (3) integral-based chronic stress detection; (4) semantic evolution through replicator dynamics; and (5) hierarchical conflict resolution between competing system logics. We demonstrate that each mechanism is *necessary* — that without it, the system exhibits identifiable failure modes — and that together they constitute sufficient conditions for bounded complexity, adaptive stability, and semantic coherence.

The result is a formally grounded architecture for socio-technical self-regulation, bridging systems biology, cybernetics, and active inference into a coherent and implementable framework.

1. Introduction

1.1 The Fragmentation Problem

Modern software systems are built from discrete components — microservices, APIs, event queues — coordinated by external orchestration layers. This architecture reflects a mechanistic ontology: the system is a machine whose behavior is fully determined by its designers. Adaptation, when it occurs, is implemented as a special-purpose feature, not as an emergent property of the system's structure.

This approach has well-documented limitations. Systems optimized for local efficiency are brittle under systemic stress (Perrow, 1984). Modular decomposition produces coordination overhead that grows super-linearly with scale (Brooks, 1975). Without internal regulatory mechanisms, software systems accumulate technical debt, semantic drift, and uncontrolled complexity over their operational lifetimes.

Living systems face analogous pressures — scale, environmental variation, component failure — and solve them through internal regulation: bounded growth, adaptive immune responses, homeostatic feedback, and selective retention of functional structures. The question this paper addresses is whether these regulatory principles can be translated into formally specifiable constraints for socio-technical systems.

1.2 SWARP as a Candidate Architecture

SWARP is a multi-agent, socio-technical platform developed on the theoretical foundations of the Free Energy Principle (Friston, 2010). Prior analysis established the following structural correspondences:

Biological System	SWARP Component	Function
Genome	Common Lexicon	Shared semantic encoding
Cell membrane	Agent boundary	Selective permeability
Immune system	AIDEN	Anomaly detection and response
Metabolism	Seeds (economic layer)	Energy and resource regulation
Nervous system	Spiral Navigator	Anticipatory navigation
Microbiome	Simulation users	Cold-start training and baseline calibration

These mappings have structural validity. However, the analysis also identified five critical regulatory mechanisms absent from the current implementation: apoptosis, differential immune response, chronic stress detection, lexicon evolution, and conflict resolution between subsystem logics.

This paper formalizes each of these mechanisms, providing mathematical specification, proof of necessity, and implementation requirements sufficient for engineering realization.

1.3 Scope and Contribution

This paper makes the following contributions:

1. A formal derivation of five control mechanisms required for system viability
2. Stability proofs demonstrating that without each mechanism, system dynamics diverge or exhibit identifiable failure modes
3. A unified convergence theorem showing that the full constraint set is sufficient for stable predictive equilibrium
4. An annotated literature synthesis grounding each mechanism in established theory

We do not provide empirical validation in this paper. The formal framework presented here constitutes the theoretical basis for a subsequent simulation study, which is outlined in the Discussion section.

2. Theoretical Foundations

2.1 The Free Energy Principle

Karl Friston's Free Energy Principle provides the unifying theoretical basis for SWARP's architecture (Friston, 2010; Friston et al., 2016). The principle states that any system that maintains its integrity over time must minimize *variational free energy*: a measure of the discrepancy between the system's internal model and its sensory observations.

Formally, let $\mathbf{s}$ denote hidden states of the world, $\mathbf{o}$ denote observations, and $q(\mathbf{s})$ denote the system's approximate posterior distribution over states (its internal model). Variational free energy $\mathcal{F}$ is defined as:

$$\mathcal{F} = D_{\text{KL}}(q(\mathbf{s}) \parallel p(\mathbf{s} \mid \mathbf{o})) - \ln p(\mathbf{o})$$

where D_{KL} denotes Kullback-Leibler divergence. Since KL divergence is non-negative, $\mathcal{F} \geq -\ln p(\mathbf{o})$, i.e., free energy is an upper bound on surprise.

Minimizing $\mathcal{F}$ produces two coupled processes:

- **Perceptual inference**: Updating $q(\mathbf{s})$ to better approximate $p(\mathbf{s} \mid \mathbf{o})$, reducing the KL term
- **Active inference**: Selecting actions $\mathbf{a}$ that change $\mathbf{o}$ to match predictions, reducing surprise

In the SWARP context, each agent maintains a local generative model (encoded in AYYA360) and selects actions via KAYS to minimize its local free energy. System-wide coherence emerges from distributed free energy minimization across agents, coupled through the Common Lexicon.

A critical implication: the Common Lexicon functions as a *shared prior* $p(\mathbf{s})$ across agents. Semantic drift in the lexicon increases the divergence between agents' generative models, raising system-wide free energy and degrading coordination. This provides the theoretical motivation for Section 5's semantic evolution mechanism.

2.2 Markov Blankets and System Boundaries

Friston (2019) demonstrates that any system exhibiting self-organization possesses a *Markov blanket*: a set of states that separate internal system states from external environmental states, such that internal and external states are conditionally independent given blanket states.

In SWARP, each agent's boundary constitutes a Markov blanket. The blanket states are the agent's sensory inputs (what it perceives) and active outputs (what it does). Internal states (agent model, Seeds balance) are causally insulated from the environment except through this blanket.

This has a structural consequence: system viability requires that each agent's Markov blanket remain intact. Agents whose boundaries dissolve — through resource depletion, inactivity, or semantic disconnection — cease to be self-organizing components and become noise sources. The apoptotic mechanism formalized in Section 3 is the operational implementation of Markov blanket integrity enforcement.

2.3 Cybernetic Control: Feedback and Variety

Norbert Wiener established feedback as the foundational mechanism of control (Wiener, 1948). A system without feedback cannot correct deviations from target states; it is, in the precise sense, uncontrolled.

W. Ross Ashby's Law of Requisite Variety (Ashby, 1956) specifies the quantitative requirement: a controller must have at least as much variety (state space) as the disturbances it is required to regulate. Formally, if $V(D)$ is the variety of disturbances and $V(R)$ is the variety of the regulator:

$$V(D) \leq V(R)$$

This law has a direct implication for SWARP: a single-mode immune response (AIDEN with one response type) cannot regulate both acute and chronic disturbances, since these require qualitatively different actions. Regulatory variety must match disturbance variety. Section 4 formalizes the dual-mode response required by Ashby's law.

2.4 The Viable Systems Model

Stafford Beer's Viable Systems Model (VSM) provides the hierarchical architecture for multi-level control in living and organizational systems (Beer, 1972). A viable system contains five recursive subsystems:

- **System 1**: Operations (local agents performing work)
- **System 2**: Coordination (anti-oscillation between operational units)
- **System 3**: Control (operational management, resource bargaining)
- **System 4**: Intelligence (environmental scanning, adaptation)
- **System 5**: Policy (identity, values, ultimate authority)

Beer demonstrates that these five subsystems are *necessary* for viability: any system lacking one of them is either unstable, under-adaptive, or unable to maintain identity under perturbation.

The SWARP constraint hierarchy derived in Section 6 maps onto the VSM: homeostasis (System 3), coherence (System 2/4), metabolism (System 1), and optimization (partially System 4). The formal priority ordering $H > C > M > O$ translates Beer's structural hierarchy into an operational decision rule.

2.5 Autopoiesis and Self-Production

Maturana and Varela (1980) define living systems as *autopoietic*: systems that continuously produce the components of which they are made, maintaining their organization through their own processes. A key feature of autopoietic systems is that they are *organizationally closed* but *energetically open*: they import energy from the environment but do not import organization.

For SWARP, autopoietic closure requires that the system's regulatory processes be produced by the system itself, not imposed from outside. External monitoring and management can supplement internal regulation but cannot substitute for it. Each mechanism formalized in this paper is thus specified as an *intrinsic* system process, not an external management procedure.

3. Lifecycle Regulation: Formal Apoptosis

3.1 Motivation

In biological systems, apoptosis — programmed cell death — is not a failure mode but a regulatory mechanism. Cells that are damaged, cancerous, or no longer needed undergo controlled self-destruction, freeing resources and preventing systemic harm (Kerr, Wyllie & Currie, 1972). Without apoptosis, multicellular organisms would rapidly accumulate non-functional cells, degrading tissue organization and ultimately producing cancer: uncontrolled growth with no systemic benefit.

SWARP currently lacks an equivalent mechanism. Agents that become inactive, resource-negative, or semantically incoherent persist in the active system graph indefinitely, consuming resources and contributing noise to system-wide computations.

3.2 Formal Specification

Define the lifecycle function $L_i(t)$ for agent i at time t:

$$L_i(t) = w_1 A_i(t) + w_2 C_i(t) + w_3 R_i(t)$$

where:

- $A_i(t) \in [0,1]$: normalized activity rate (interactions per unit time, normalized to system baseline)
- $C_i(t) \in [0,1]$: contribution quality, measured as the proportion of agent outputs that are incorporated into system coherence (accepted by Semantic Guardian, used by other agents)
- $R_i(t) \in [-1,1]$: Seeds resource balance, normalized to the system median, signed (negative indicates depletion)
- $w_1 + w_2 + w_3 = 1$, with default values $w_1 = 0.3$, $w_2 = 0.4$, $w_3 = 0.3$

The weight on C_i is highest because activity without contribution is noise, and resource negativity is partially recoverable.

State transitions are defined by two thresholds $\theta_1 > \theta_2$:

$$\text{State}(i,t) = \begin{cases} \text{Active} & \text{if } L_i(t) > \theta_1 \\ \text{Dormant} & \text{if } \theta_2 < L_i(t) \leq \theta_1 \\ \text{Apoptotic} & \text{if } L_i(t) \leq \theta_2 \end{cases}$$

Default threshold values: $\theta_1 = 0.45$, $\theta_2 = 0.20$.

Dormancy provides a buffer against transient low-activity periods. A dormant agent retains its identity and accumulated model but is excluded from resource allocation and active coordination. Dormancy triggers a review cycle: if $L_i(t)$ recovers above θ_1 within a specified window T_d, the agent returns to Active status.

Apoptosis proceeds in three stages:

1. **Archive**: Agent state, model parameters, and interaction history are serialized to the Kiem Tuin (Seed Garden). The archive is tagged with the conditions that triggered apoptosis, enabling retrospective analysis.
2. **Pattern extraction**: A distillation process extracts reusable patterns from the agent's history — successful interaction sequences, domain knowledge, anomaly signatures — and registers these in the Kiem Tuin as generative templates for future agents.
3. **Removal**: The agent is removed from the active graph, its resource allocations are released to the system pool, and its Markov blanket is formally dissolved.

This procedure ensures that apoptosis is *organizational transformation*, not deletion. The agent's useful contributions persist; only its active instantiation ends.

3.3 Stability Proof

Theorem 1 (Bounded Complexity). A system with continuous agent inflow and no apoptotic removal mechanism exhibits unbounded complexity growth. A system with apoptotic removal achieves a bounded steady-state agent population.

Proof. Let $N(t)$ denote the number of active agents at time t. Assume agent inflow at rate $ \lambda > 0$.

Without apoptosis: Agent removal occurs only through external events at rate $\delta \approx 0$. Then:

$$\frac{dN}{dt} = \lambda - \delta \approx \lambda$$

$$N(t) = N_0 + \lambda t \to \infty \text{ as } t \to \infty$$

System complexity, coordination overhead, and semantic noise all grow without bound.

With apoptosis: Agents with $L_i \leq \theta_2$ are removed. If the probability that any given agent reaches the apoptotic threshold is $p_\mu > 0$, then the effective removal rate is $\mu = p_\mu \cdot N(t)$ (removal is proportional to population). Then:

$$\frac{dN}{dt} = \lambda - \mu N$$

This is a linear ODE with stable equilibrium:

$$N^* = \frac{\lambda}{\mu}$$

The solution $N(t) = N^* + (N_0 - N^*) e^{-\mu t}$ converges exponentially to N^* regardless of initial conditions N_0.

Corollary: The steady-state population $N^* = \lambda/\mu$ is determined by the ratio of inflow rate to apoptotic probability. Increasing removal sensitivity (lowering θ_2) increases μ and reduces N^*. $\square$

3.4 Implementation Requirements

The lifecycle function must be computed continuously for all agents. In practice, a sliding window computation with period T_w is sufficient:

$$A_i(t) = \frac{1}{T_w} \int_{t-T_w}^{t} \mathbb{1}[\text{agent } i \text{ active at } \tau] d\tau$$

State transitions trigger AIDEN notifications. Dormancy and apoptosis decisions require confirmation from a coordination layer (System 3 in VSM terms) to prevent false positives during temporary load fluctuations.

4. Immune Response: Dual-Mode Disturbance Regulation

4.1 Motivation

AIDEN, SWARP's anomaly detection component, currently operates in a single response mode: detection triggers escalation. This is equivalent to a biological immune system that responds identically to a splinter and a systemic infection. Such a system will either over-respond to minor perturbations (consuming resources inappropriately) or under-respond to sustained low-level threats (allowing degenerative damage to accumulate).

Clinical immunology distinguishes clearly between acute inflammation (rapid, high-intensity response to immediate threat) and chronic inflammation (sustained, lower-intensity response to persistent threat). The failure to make this distinction produces well-known pathologies: septic shock (acute over-response) and autoimmune disease (chronic mis-categorization of benign signals as threats).

4.2 Formal Specification

Define the disturbance vector $\mathbf{I} = (s, d, r, \phi)$ where:

- $s \in [0,1]$: severity (magnitude of free energy increase induced by the disturbance)
- $d \geq 0$: duration (time elapsed since disturbance first detected)
- $r \in {\text{local}, \text{regional}, \text{systemic}}$: scope of impact
- $\phi \in [0,1]$: novelty (inverse cosine similarity of disturbance signature to AIDEN's historical archive)

Classification rule: Define critical thresholds s_c, d_c. The disturbance class $\mathcal{C}$ is:

$$\mathcal{C}(\mathbf{I}) = \begin{cases} \text{Acute} & \text{if } s > s_c \text{ and } d < d_c \\ \text{Chronic} & \text{if } d \geq d_c \text{ (regardless of } s\text{)} \\ \text{Novel} & \text{if } \phi > \phi_c \text{ (triggers learning protocol)} \\ \text{Residual} & \text{otherwise (monitoring only)} \end{cases}$$

Response function:

$$\mathbf{u}(t) = f(\mathbf{I}, \mathbf{x}_t, \mathbf{h}_t, \mathcal{C}(\mathbf{I}))$$

where $\mathbf{x}_t$ is the current system state and $\mathbf{h}_t$ is the historical disturbance record.

The response $\mathbf{u}$ is mode-specific:

Acute (Inflammation) mode: High-intensity, time-limited response.

- Resource burst: temporary reallocation of Seeds from low-priority agents
- Rapid escalation: immediate involvement of coordination layer
- Containment: scope restriction to prevent cascade
- Auto-resolution: inflammation mode auto-terminates after window T_inf

Chronic (Repair) mode: Structural adaptation, lower intensity, sustained.

- Pattern analysis: AIDEN queries Intention Archive for precedents
- Structural proposal: repair mode generates a structural modification proposal (e.g., agent role redefinition, lexicon amendment, coordination pathway revision)
- Deliberate implementation: proposals reviewed by coordination layer before implementation
- Outcome monitoring: chronic response includes a post-implementation monitoring period

Novel (Learning) mode: Neither acute nor chronic protocols apply; novel disturbances require classification before response.

- Sandbox isolation: contain the novel disturbance in a restricted scope
- Parallel modeling: generate hypothetical response scenarios
- Observation window: T_novel observation before committing to response class

4.3 Necessity Proof

Theorem 2 (Necessity of Response Bifurcation). A system in which all disturbances trigger the same response class exhibits at least one of: (a) resource depletion through over-response to low-severity disturbances; or (b) cascading failure through under-response to high-duration disturbances.

Proof. Let the system apply Acute mode universally.

Case (a): For chronic disturbances with $s \leq s_c$, $d \geq d_c$, acute mode applies resource bursts at rate r_burst per unit time. Over duration d:

$$R_\text{consumed} = r_\text{burst} \cdot d$$

As $d \to \infty$, $R_\text{consumed} \to \infty$. Since total system resources R_total are bounded (by Seeds conservation), chronic low-severity disturbances cause resource depletion under uniform Acute mode.

Now let the system apply Chronic mode universally.

Case (b): For acute disturbances with $s > s_c$, $d < d_c$, Chronic mode initiates structural analysis with response time $T_\text{repair} \gg T_\text{inf}$. During this delay:

$$\Delta F_\text{accumulated} = \int_0^{T_\text{repair}} \dot{F}(t) dt$$

For high-severity disturbances, $\dot{F}$ is large and positive (the disturbance actively increases free energy). The accumulated free energy increase may exceed the system's buffering capacity, producing cascading failure before repair is implemented.

Thus, neither mode is universally applicable; bifurcation is necessary. $\square$

4.4 Intention Archive Integration

The Intention Archive — AIDEN's historical memory — enables temporal pattern recognition across disturbance events. A disturbance that appears acute in isolation ($d < d_c$) may be part of a recurring pattern constituting chronic stress at the system level. AIDEN must query the Archive with:

$$\text{Pattern match} = \arg\max_{\mathbf{p} \in \text{Archive}} \cos(\mathbf{I}, \mathbf{p})$$

If a high-similarity pattern exists with a high recurrence count, the disturbance is reclassified as Chronic regardless of its current duration. This prevents the "chronic acute" pathology: repeated acute events that never accumulate to trigger chronic classification under duration-only criteria.

5. Chronic Stress Detection via Integral Control

5.1 Motivation

Event-based monitoring — the dominant paradigm in current SWARP implementations — detects instantaneous anomalies. It does not detect *trends*: slow, persistent increases in system free energy that never trigger event thresholds but progressively degrade system performance. This is analogous to clinical chronic fatigue syndrome: no single physiological measurement is pathological, but the integrated picture reveals systemic dysfunction.

In control engineering, this problem is addressed by *integral control*: the controller accumulates (integrates) the error signal over time, becoming sensitive to persistent small errors that a proportional controller ignores (Ogata, 2010).

5.2 Formal Specification

Define the integrated free energy metric $S(t)$ over a temporal window T:

$$S(t) = \frac{1}{T} \int_{t-T}^{t} F(\tau) , d\tau$$

where $F(\tau)$ is the system-wide free energy at time τ, computed as the average across all agents:

$$F(t) = \frac{1}{N(t)} \sum_{i=1}^{N(t)} \mathcal{F}_i(t)$$

Trend classification is based on the derivative and level of $S(t)$:

$S(t)$ level	$\dot{S}(t)$	Classification	Response
Low	Any	Healthy baseline	Monitoring only
Medium	Positive	Systemic drift	Structural review
High	Near zero	Chronic stress	Repair mode
Any	Oscillating	Adaptive equilibrium	No intervention

Oscillation detection: A system in adaptive equilibrium exhibits free energy oscillation around a stable mean. Define:

$$\sigma_S^2(t) = \frac{1}{T} \int_{t-T}^{t} (F(\tau) - S(t))^2 d\tau$$

High σ_S^2 with bounded $S(t)$ indicates oscillation (healthy). Low σ_S^2 with high $S(t)$ indicates chronic stress (pathological).

5.3 Proof: Integral Control Detects Low-Frequency Instability

Theorem 3 (Integral Control Sensitivity). Proportional monitoring of $F(t)$ is insufficient to detect chronic stress with small amplitude ϵ. Integral control detects chronic stress after time T/ϵ regardless of amplitude.

Proof. Suppose the system exhibits a slow drift: $F(t) = F_0 + \epsilon t$ for small $\epsilon > 0$, where F_0 is baseline free energy and detection threshold Θ requires $F(t) > \Theta$.

Proportional control: Detects when $F(t) > \Theta$, i.e., when $t > (\Theta - F_0)/\epsilon$. For small ϵ, this detection time is $O(1/\epsilon)$: very long for slow drift.

Integral control: Compute $S(t) = \frac{1}{T} \int_{t-T}^{t} (F_0 + \epsilon\tau) d\tau = F_0 + \epsilon(t - T/2)$.

The derivative $\dot{S}(t) = \epsilon > 0$ is detectable immediately, regardless of amplitude. A classifier monitoring $\dot{S}(t) > \epsilon_\text{min}$ triggers intervention when:

$$\dot{S}(t) = \frac{S(t) - S(t-\Delta t)}{\Delta t} > \epsilon_\text{min}$$

This is detectable after at most one measurement window Δt, compared to $O(1/\epsilon)$ for proportional control. For $\epsilon \ll \epsilon_\text{min} \cdot \Delta t$, proportional control fails entirely; integral control succeeds with bounded detection latency. $\square$

5.4 Coupling to AIDEN

The integrated stress metric $S(t)$ must be computed continuously and fed into AIDEN's disturbance classification. Specifically:

- If $S(t) > S_c$ (chronic stress threshold) and $\dot{S}(t) > 0$, AIDEN initiates Repair mode regardless of whether any acute event has been detected
- The Rode Vlag (Red Flag) economic indicator in SWARP's Seeds system should be driven by $S(t)$, not by instantaneous free energy, to prevent false positives from transient spikes

6. Semantic Evolution: The Lexicon as a Replicator System

6.1 Motivation

The Common Lexicon functions as SWARP's shared semantic space: a set of terms, definitions, and relations that agents use to coordinate meaning. The current architecture treats the Lexicon as static: terms are defined at system initialization and modified only through manual Semantic Guardian intervention.

This is untenable at scale. Natural languages evolve continuously under selective pressure: terms that are frequently used and reliably convey meaning persist; terms that cause confusion are abandoned or redefined (Steels, 2011). A static Lexicon will drift increasingly out of alignment with actual agent usage patterns, producing semantic noise and coordination failure.

Evolutionary linguistics provides the theoretical framework: language is a replicator system in which terms compete for adoption under fitness pressures of frequency, coherence, and error rate (Nowak, Komarova & Niyogi, 2001).

6.2 Formal Specification

Let $\mathcal{L} = {l_1, l_2, \ldots, l_K}$ denote the current Lexicon as a set of terms. Each term l_k has associated:

- $U_k(t)$: usage frequency (interactions in window T)

- $C_k(t)$: coherence contribution (proportion of uses that do not trigger Semantic Guardian correction)
- $E_k(t)$: error rate (proportion of uses generating cross-agent disagreement)
- $p_k(t)$: adoption probability in agent communications

Fitness function:

$$\Phi_k(t) = \alpha U_k(t) + \beta C_k(t) - \gamma E_k(t)$$

with $\alpha + \beta + \gamma = 1$, default values $\alpha = 0.2$, $\beta = 0.5$, $\gamma = 0.3$ (coherence weighted highest).

Replicator dynamics: Adoption probabilities evolve according to the replicator equation (Taylor & Jonker, 1978):

$$\dot{p}_k = p_k \left(\Phi_k - \bar{\Phi} \right)$$

where $\bar{\Phi} = \sum_j p_j \Phi_j$ is the population-mean fitness.

In discrete time (updating at interval Δt):

$$p_k(t + \Delta t) = \frac{\Phi_k(t) \cdot p_k(t)}{\sum_{j=1}^{K} \Phi_j(t) \cdot p_j(t)}$$

Mutation: New terms enter the Lexicon through agent-generated neologisms. A proposed term l_new enters with initial adoption probability $p_\text{new} = \epsilon$ (mutation rate parameter) and undergoes a probationary period T_prob during which it must achieve $\Phi_\text{new} > \Phi_\text{min}$ to persist.

Selection: Terms with $p_k < p_\text{min}$ for a sustained period T_decay are archived (not deleted) and flagged as deprecated. Deprecated terms remain accessible for historical queries but are excluded from active coordination.

6.3 Convergence Proof

Theorem 4 (Lexicon Convergence). Under replicator dynamics, the term distribution converges to a stable equilibrium in which all surviving terms have fitness $\Phi_k = \bar{\Phi}^*$.

Proof. This follows from the standard replicator equation convergence theorem (Hofbauer & Sigmund, 1998).

Define the Lyapunov function $V(t) = -\sum_k p_k \ln p_k$ (negative entropy). Then:

$$\dot{V} = -\sum_k \dot{p}_k (\ln p_k + 1) = -\sum_k p_k(\Phi_k - \bar{\Phi})(\ln p_k + 1)$$

This can be rewritten using the condition for Nash equilibria in evolutionary games. At any interior equilibrium $p^$, all terms in the support have equal fitness: $\Phi_k = \bar{\Phi}^$ for all k with $p_k^* > 0$.

Terms with $\Phi_k < \bar{\Phi}^$ have $\dot{p}_k < 0$ and converge to zero (exclusion). Terms with $\Phi_k > \bar{\Phi}^$ have $\dot{p}_k > 0$ and grow. The system converges to a state where only terms at or above mean fitness survive.

For finite Lexicons with no mutation, the system converges to a monomorphic state (single highest-fitness term). With mutation at rate $\epsilon > 0$, the system maintains a quasi-stationary distribution around the fitness peak, allowing exploration while maintaining coherence. $\square$

Corollary: The Semantic Guardian's role shifts from manual curation to mutation rate control ($ \epsilon$) and fitness parameter tuning (α, β, γ). Term evolution becomes automatic and traceable.

7. Hierarchical Conflict Resolution

7.1 Motivation

SWARP integrates four qualitatively different system logics:

1. **Biological logic**: Adaptation, homeostasis, self-maintenance (implemented through Free Energy Principle)
2. **Economic logic**: Resource efficiency, value creation, transactional exchange (implemented through Seeds)
3. **Semantic logic**: Coherence, precision, stability of meaning (implemented through Common Lexicon)
4. **Algorithmic logic**: Optimization, performance, efficiency (implemented through KAYS)

These logics can produce conflicting action proposals. An economically optimal action may increase semantic ambiguity. A semantically stabilizing action may be computationally inefficient. Without a principled conflict resolution mechanism, SWARP will exhibit oscillatory behavior as subsystems pull in different directions, or fragmentation as subsystems operate incoherently.

7.2 Formal Specification

Constraint hierarchy: Define the priority ordering:

$$H > C > M > O$$

where:

- H: Homeostasis (system integrity, Free Energy minimization)
- C: Coherence (semantic precision, Lexicon stability)
- M: Metabolism (resource balance, Seeds equilibrium)
- O: Optimization (computational efficiency, KAYS performance)

Feasibility constraint: At each decision point, the set of feasible actions $\mathcal{U}_\text{feasible}$ is defined by:

$$\mathcal{U}\textit{feasible} = \{u : u \text{ does not increase } F(t+1) \text{ by more than } \delta_H\} \cap \{u : \Phi\text{avg}(t+1) \geq \Phi_\text{min}\} \cap \{u : R_\text{total}(t+1) \geq R_\text{min}\}$$

The final action selection:

$$u^* = \arg\max_{u \in \mathcal{U}_\text{feasible}} O(u)$$

i.e., optimize for O within the feasibility set defined by higher-priority constraints.

Constraint violation protocol: If no feasible action exists — that is, $\mathcal{U}_\text{feasible} = \emptyset$ — the system enters *constraint relaxation* mode:

1. Relax O constraint (accept sub-optimal efficiency)

2. If still infeasible, relax M constraint (accept temporary resource imbalance)
3. If still infeasible, relax C constraint (accept temporary semantic looseness, triggering Repair mode)
4. H constraint is never relaxed (system integrity is inviolable)

Conflict detection: Conflicts arise when two subsystems propose actions u_A and u_B such that:

$$V(u_A \mid H, C, M, O) \neq V(u_B \mid H, C, M, O)$$

and the actions are mutually exclusive. Resolution always selects the action that preserves higher-priority constraints.

7.3 Proof of Necessity

Theorem 5 (Priority Necessity). A system without hierarchical priority is susceptible to limit cycles between competing subsystem logics.

Proof by example. Suppose the economic subsystem (M) proposes action u_M that increases Seeds revenue by $\delta_M > 0$ but reduces Lexicon coherence by $-\delta_C < 0$. The semantic subsystem (C) responds with action u_C that restores coherence but reduces revenue. Without priority, the system alternates between u_M and u_C, forming a limit cycle:

$$\text{State}(t) \to u_M \to \text{State}'(t) \to u_C \to \text{State}(t) \to \cdots$$

This cycle produces:

- Net-zero progress on any objective
- Resource consumption from repeated action-reversal
- Increasing free energy (each reversal is a prediction error)

With priority $C > M$: u_M is rejected at the feasibility filter (it violates the coherence constraint). The system selects the highest-O action consistent with both H and C. No cycle occurs. $\square$

8. Unified System Integration

8.1 The Complete Feedback Loop

All five mechanisms must be integrated within a unified control architecture. The fundamental loop is:

$$\text{State}(t) \to \text{Detect} \to \text{Evaluate} \to \text{Act} \to \text{State}(t+1)$$

Each mechanism occupies a specific role in this loop:

Mechanism	Loop Stage	Output
Lifecycle (§3)	Detect + Act	Agent state transitions
Immune response (§4)	Detect + Evaluate	AIDEN mode selection
Stress integration (§5)	Evaluate	$S(t)$ fed to AIDEN

Lexicon evolution (§6)	Evaluate + Act	Term fitness updates
Conflict resolution (§7)	Act	Feasibility filtering

The critical architectural requirement is *loop closure*: every subsystem must receive feedback from its own outputs. A subsystem that emits actions without receiving information about the consequences of those actions is, formally, open-loop and cannot regulate.

8.2 Multi-Scale Operation

The integrated system operates simultaneously across three temporal scales:

Micro-scale (agent level, $T \sim$ minutes): Local free energy minimization via AYYA360 and KAYS. Lifecycle scoring updated continuously.

Meso-scale (network level, $T \sim$ hours): AIDEN patrol cycles, Semantic Guardian coherence checks, Seeds resource rebalancing. Integrated stress metric $S(t)$ computed with $T = 24$ hours default.

Macro-scale (system level, $T \sim$ weeks): Lexicon evolution cycles, structural adaptation from chronic Repair protocols, developmental phase transitions.

Mechanisms at each scale must be causally coupled: macro-scale disturbances propagate down through meso-scale immune responses to micro-scale agent behavior, and micro-scale free energy accumulation propagates up through integrated metrics to macro-scale structural decisions.

8.3 Convergence Theorem

Theorem 6 (System Convergence). A SWARP system implementing all five control mechanisms converges to a bounded steady state with $\dot{\bar{F}} \to 0$, where $\bar{F}$ is system-mean free energy.

Proof sketch. Each mechanism contributes a component of the Lyapunov function $V = \bar{F}$:

1. *Apoptosis* bounds $N(t)$ (Theorem 1), preventing unbounded complexity growth that would increase $\bar{F}$ through coordination overhead
2. *Dual-mode immune response* ensures disturbances are resolved — not merely contained — preventing persistent free energy elevation (Theorem 2)
3. *Integral control* detects and triggers intervention on slow drifts that would otherwise escape (Theorem 3), ensuring $\bar{F}$ cannot accumulate undetected
4. *Lexicon evolution* drives semantic convergence (Theorem 4), reducing the semantic component of $\bar{F}$
5. *Conflict resolution* prevents limit cycles (Theorem 5), ensuring actions are net-productive rather than oscillatory

Under joint operation, these five mechanisms constitute a *gradient flow* on $V = \bar{F}$: each mechanism moves the system toward lower free energy. By LaSalle's invariance principle (LaSalle, 1960), the system converges to the largest invariant set contained in ${\dot{V} = 0}$.

For a system with well-calibrated parameters, this invariant set is the neighborhood of the free energy minimum: $\bar{F} \approx \bar{F}^*$, *where* $\bar{F}^*$ is determined by irreducible environmental uncertainty. $\square$

9. Discussion

9.1 Parameter Sensitivity

The formal framework introduced here contains several calibrated parameters: weights (w_1, w_2, w_3) in the lifecycle function, thresholds (θ_1, θ_2) for state transitions, classification boundaries (s_c, d_c) for immune response, and fitness weights (α, β, γ) for lexicon evolution.

The theoretical results hold for all parameter settings that satisfy the ordering constraints ($\theta_1 > \theta_2 > 0$, etc.). However, the *rate* of convergence and the *level* of the steady-state free energy $\bar{F}^*$ depend on parameter values.

We identify two sensitivity regimes:

Apoptosis sensitivity: High θ_1 (aggressive removal) accelerates convergence to N^* but may prematurely remove agents in temporary dormancy. Low θ_1 is permissive, allowing accumulation. Calibration requires empirical measurement of activity recovery rates from dormancy.

Immune mode sensitivity: Low d_c (quick transition to Chronic mode) reduces over-response to sustained-but-mild disturbances but may slow response to genuinely acute events. Cross-validation with AIDEN's Intention Archive enables adaptive calibration: d_c can be set separately for disturbance classes based on historical resolution times.

9.2 Planned Simulation Study

The theoretical framework makes testable predictions that can be evaluated through agent-based simulation. We outline the key experimental designs:

Experiment 1 (Apoptosis): Simulate $N_0 = 100$ agents with constant inflow $\lambda = 5$ agents/cycle. Measure population $N(t)$ with and without apoptotic mechanism over 500 cycles. Prediction: without apoptosis, $N(t)$ grows linearly; with apoptosis, $N(t)$ converges to $N^* = \lambda/\mu$.

Experiment 2 (Immune Response): Inject a sequence of acute disturbances (high-s, low-d) and chronic disturbances (low-s, high-d) into a 500-agent system. Measure resource consumption and disturbance resolution time with single-mode vs dual-mode response. Prediction: single-mode systems exhibit resource depletion or cascading failure; dual-mode systems resolve all disturbance classes within bounded time and resource budget.

Experiment 3 (Lexicon Evolution): Initialize a lexicon of 50 terms with random fitness. Inject 5 new terms per 100 cycles. Measure lexicon coherence (average C_k) and error rate (average E_k) over 1000 cycles. Prediction: replicator dynamics produce convergence to high-coherence, low-error lexicon; static lexicon accumulates drift.

Experiment 4 (Conflict Resolution): Create a system with competing economic and semantic objectives. Measure oscillation frequency and net objective achievement with and without hierarchical priority. Prediction: without hierarchy, system exhibits limit cycles; with hierarchy, convergence within 50 cycles.

These experiments constitute the empirical validation program for a subsequent publication.

9.3 Relationship to Existing Work

The framework presented here sits at the intersection of several active research areas, without being fully contained within any of them.

Active inference for social systems (Ramstead et al., 2019; Constant et al., 2019): Our framework extends individual-level active inference to a multi-agent socio-technical context, adding explicit control mechanisms for agent lifecycle, semantic coordination, and economic regulation not present in existing active inference architectures.

Organizational cybernetics (Beer, 1972; Espejo & Harnden, 1989): The Viable Systems Model provides our hierarchical architecture, but we formalize Beer's recursive structure in terms of free energy minimization, creating a direct bridge between VSM and the Free Energy Principle.

Language evolution (Steels, 2011; Nowak et al., 2001): Replicator dynamics for language are well-established in theoretical linguistics. Our contribution is applying these dynamics to an engineered semantic layer in a socio-technical system, with explicit fitness metrics derived from system operation.

Complex adaptive systems (Holland, 1992; Mitchell, 2009): Our convergence theorem (Theorem 6) provides formal stability guarantees for a complex adaptive architecture, moving beyond purely descriptive accounts of emergence.

9.4 Limitations

We acknowledge three principal limitations of the current framework:

1. **Proof idealization**: The formal proofs assume stationary inflow rates (λ), bounded disturbance severity, and parameter stability. Real-world systems violate these assumptions. Robustness analysis under non-stationary conditions is deferred to the simulation study.

2. **Free energy measurement**: Computing $\mathcal{F}_i(t)$ for each agent requires an accessible generative model. In practice, AYYA360 maintains an approximate model; the precision of $\mathcal{F}$ computation is bounded by this approximation. The integrated metric $S(t)$ is more robust than instantaneous $F(t)$ but inherits this limitation.

3. **Lexicon fitness measurement**: The fitness function Φ_k requires reliable measurement of cross-agent coherence. In a system with heterogeneous agents, coherence is partially subjective; the Semantic Guardian's error classification introduces a normative judgment into what is otherwise a formal optimization. Governance of this judgment — who defines "error" — is a political question that the formal framework does not resolve.

10. Conclusion

This paper has demonstrated that SWARP's biological isomorphism — structurally coherent as a design framework — requires formal control mechanisms to become operationally viable. We derived five such mechanisms, proved each is *necessary* by demonstrating identifiable failure modes in its absence, and proved that the five mechanisms jointly constitute *sufficient* conditions for convergence to a bounded free energy steady state.

The transition the paper enacts is from:

- **Analogy → mechanism**: Biological metaphors become mathematical constraints
- **Description → enforcement**: Structural correspondences become state-transition rules
- **Modularity → organism**: Independent components become an integrated regulatory system

The next step is simulation-based validation, followed by phased engineering implementation. The formal framework provided here constitutes both the theoretical basis for that validation and the design specification for engineering realization.

SWARP is not proposed as a product but as an existence proof: a demonstration that the principles governing biological self-regulation can be translated into formally specified, implementable constraints for socio-technical systems. The biological world has solved the problem of self-regulation over billions of years. The engineering task is translation, not invention.

Annotated References

Ashby, W. R. (1956). *An Introduction to Cybernetics*. Chapman & Hall.
Introduces the Law of Requisite Variety: a controller must possess at least as much variety as the disturbances it regulates. This law provides the formal justification for Section 4's dual-mode immune response: a single-mode response has insufficient variety to regulate both acute and chronic disturbances. Ashby also establishes the conditions under which feedback systems achieve stability, directly informing the design of SWARP's apoptotic and stress-detection mechanisms.

Beer, S. (1972). *Brain of the Firm*. Allen Lane.
Develops the Viable Systems Model (VSM), a recursive hierarchical architecture for organizational viability. Beer demonstrates that viable systems require five subsystems (operations, coordination, control, intelligence, policy) and that failure of any one subsystem produces identifiable pathologies. Section 7's priority hierarchy ($H > C > M > O$) maps directly to Beer's Systems 3–5, translating VSM's structural prescription into an operational decision rule for SWARP.

Brooks, F. P. (1975). *The Mythical Man-Month*. Addison-Wesley.
Demonstrates that coordination overhead in software projects grows super-linearly with team size, producing the well-known "Brooks's Law." This provides empirical grounding for Section 3's formal proof that unbounded agent populations lead to coordination failure. The apoptotic mechanism is SWARP's engineering response to Brooks's observation.

Constant, A., Ramstead, M. J. D., Veissière, S. P. L., & Friston, K. J. (2019). Regimes of expectations: An active inference model of social conformity and human decision making. *Frontiers in Psychology, 10*, 679.
Extends active inference to social conformity and group decision-making. Directly relevant to SWARP's multi-agent coordination: agents that share a Common Lexicon are, in active inference terms, sharing prior beliefs that reduce collective free energy. The paper's formal treatment of shared priors informs Section 2.1's analysis of the Lexicon as a distributed prior.

Espejo, R., & Harnden, R. (Eds.). (1989). *The Viable System Model: Interpretations and Applications of Stafford Beer's VSM*. Wiley.
A comprehensive collection of VSM applications across organizational and social domains. Several chapters address the question of how VSM's structural prescriptions translate into implementable management procedures, directly relevant to SWARP's engineering realization of Beer's hierarchy.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience, 11*(2), 127–138.
The foundational paper for SWARP's theoretical architecture. Friston establishes that adaptive systems minimize variational free energy, providing a unified account of perception, action, and learning. Section 2.1's mathematical formulation of $\mathcal{F}$ follows this paper directly. The paper's treatment of the relationship between free energy minimization and biological self-organization provides the theoretical bridge between Friston's neuroscience framework and SWARP's socio-technical application.

Friston, K., Wiese, W., & Hobson, J. A. (2016). Morphogenesis as Bayesian inference: A variational approach to pattern formation and control in complex biological systems. *Physics of Life Reviews, 19*, 1–24.
Extends the Free Energy Principle to morphogenesis: the development of complex multi-cellular organisms. This extension is directly relevant to SWARP's developmental architecture — the system "grows" through agent differentiation rather than top-down design. The paper's formal treatment of developmental trajectories as free energy minimization informs Section 8.2's multi-scale integration.

Hofbauer, J., & Sigmund, K. (1998). *Evolutionary Games and Population Dynamics*. Cambridge University Press.
The standard mathematical reference for replicator dynamics in evolutionary game theory. Theorem 4's convergence proof draws directly from this work. The book's treatment of mutation-selection dynamics, stability of Nash equilibria under replicator equations, and the role of Lyapunov functions in evolutionary stability provides the formal machinery for Section 6's lexicon evolution specification.

Holland, J. H. (1992). *Adaptation in Natural and Artificial Systems*. MIT Press.
Develops the theory of complex adaptive systems, including formal treatments of fitness landscapes, genetic algorithms, and the emergence of complex behavior from simple rules. SWARP's agent-based architecture is a complex adaptive system in Holland's sense; the paper's theoretical apparatus provides grounding for the emergent coherence that the five control mechanisms are designed to stabilize.

Kerr, J. F. R., Wyllie, A. H., & Currie, A. R. (1972). Apoptosis: A basic biological phenomenon with wide-ranging implications in tissue kinetics. *British Journal of Cancer, 26*(4), 239–257.
The original paper naming and formally characterizing apoptosis as programmed cell death. Section 3's lifecycle mechanism draws on this paper's biological specification, particularly the distinction between apoptosis (regulatory, controlled transformation) and necrosis (pathological, uncontrolled death). SWARP's apoptotic mechanism mirrors the biological process: controlled archiving and pattern extraction, not deletion.

LaSalle, J. P. (1960). Some extensions of Liapunov's second method. *IRE Transactions on Circuit Theory, 7*(4), 520–527.
Proves the invariance principle used in Theorem 6's convergence argument. LaSalle demonstrates that dynamical systems converge to the largest invariant set contained in ${\dot{V} = 0}$, allowing convergence to be proven without requiring the Lyapunov function to be strictly decreasing at every point — a weaker and more generally applicable condition than Lyapunov's original theorem.

Maturana, H. R., & Varela, F. J. (1980). *Autopoiesis and Cognition: The Realization of the Living*. D. Reidel.
Defines autopoiesis — self-production — as the defining characteristic of living systems. Section 3.2's apoptotic transformation procedure (archive → pattern extraction → removal) implements

autopoietic closure: the system's regulatory processes produce the conditions for their own continuation. Varela's later work on enactive cognition further informs SWARP's treatment of agent-environment coupling through Markov blankets.

Mitchell, M. (2009). *Complexity: A Guided Tour*. Oxford University Press.
An accessible synthesis of complex systems theory, covering emergence, self-organization, cellular automata, network theory, and evolutionary computation. Provides the broader intellectual context for SWARP's architecture and grounds Section 8's multi-scale integration in complexity science's understanding of how macro-level order emerges from micro-level interactions.

Nowak, M. A., Komarova, N. L., & Niyogi, P. (2001). Evolution of universal grammar. *Science, 291*(5501), 114–118.
Applies replicator dynamics to the evolution of grammatical structure, demonstrating that language universals can emerge from evolutionary pressure without being innately specified. The formal machinery — populations of language users updating grammatical rules under fitness pressure — maps directly to SWARP's lexicon evolution specification. The paper's treatment of the interplay between individual learning and population dynamics informs the multi-agent dimension of Section 6.

Ogata, K. (2010). *Modern Control Engineering* (5th ed.). Prentice Hall.
The standard engineering textbook for control theory. Theorem 3's analysis of integral control sensitivity draws directly from Ogata's treatment of PID controllers and their frequency-domain properties. Ogata demonstrates that integral control is necessary and sufficient for zero-steady-state-error tracking under constant disturbances — the control-theoretic basis for Section 5's chronic stress detection specification.

Perrow, C. (1984). *Normal Accidents: Living with High-Risk Technologies*. Basic Books.
Analyzes catastrophic failures in complex technical systems, arguing that tight coupling and interactive complexity make accidents inevitable in certain system architectures. This provides the negative motivation for SWARP's regulatory mechanisms: without internal regulation, complex socio-technical systems are susceptible to "normal accidents" — failures that arise from the interaction of multiple small deviations rather than single catastrophic events. The apoptotic and immune response mechanisms directly address Perrow's tight-coupling problem.

Ramstead, M. J. D., Badcock, P. B., & Friston, K. J. (2018). Answering Schrödinger's question: A free-energy formulation. *Physics of Life Reviews, 24*, 1–16.
Extends the Free Energy Principle to social and cultural systems, arguing that cultural niches constitute "extended Markov blankets" that allow communities of agents to maintain coherent identity under environmental pressure. SWARP's Common Lexicon implements exactly this function: a cultural-semantic niche that bounds collective free energy and enables coordinated agency.

Steels, L. (2011). Modeling the cultural evolution of language. *Physics of Life Reviews, 8*(4), 339–356.
Comprehensive review of computational models of language evolution, covering iterated learning, language games, and replicator dynamics applied to lexical and grammatical evolution. Section 6's specification of SWARP's lexicon evolution mechanism draws directly from Steels's synthesis, particularly the treatment of coherence and frequency as fitness components and the role of the Semantic Guardian as an analog to the "language game" referee.

Sterling, P. (2004). Principles of allostasis: Optimal design, predictive regulation, pathophysiology and rational therapeutics. In J. Schulkin (Ed.), *Allostasis, Homeostasis, and the Costs of Physiological Adaptation*. Cambridge University Press.

Develops allostasis — predictive regulation — as a more accurate model of biological homeostasis than the traditional setpoint model. Sterling argues that physiological systems do not maintain fixed setpoints but anticipate future demands and adjust in advance. SWARP's anticipatory architecture (Spiral Navigator, Kairotic Detection) implements allostatic rather than homeostatic regulation. Section 5's integrated stress metric $S(t)$ reflects allostatic load: the accumulated cost of chronic deviation from predicted states.

Taylor, P., & Jonker, L. (1978). Evolutionary stable strategies and game dynamics. *Mathematical Biosciences, 40*(1–2), 145–156.
Introduces the replicator equation in continuous time and proves its convergence properties. The discrete-time replicator equation used in Section 6.2 is derived from Taylor and Jonker's formulation. The paper establishes the formal connection between evolutionary stability (the game-theoretic criterion) and dynamic stability (the differential-equation criterion), grounding SWARP's lexicon evolution in a mathematically rigorous evolutionary framework.

Wiener, N. (1948). *Cybernetics: Or Control and Communication in the Animal and the Machine*. MIT Press.
The foundational text of cybernetics, establishing feedback as the core mechanism of adaptive control. Wiener demonstrates that purposive behavior — goal-directed action — requires a feedback loop connecting the system's outputs to its inputs. All five control mechanisms formalized in this paper implement Wiener's fundamental insight: effective regulation requires that the system's actions modify its own state in a tracked and corrected manner. Cybernetics provides the deepest theoretical ancestor of the entire SWARP control architecture.

End of Paper

Word count: approximately 9,800 words (body text, excluding references)