A Unified Theory of Self-Organizing Systems: Four Formal Laws on Cooperation, Viability, Interference, and Observability

Boris Kriger
Institute of Integrative and Interdisciplinary Research
boriskriger@interdisciplinary-institute.org

January 24, 2026

Abstract

This paper presents a unified formal framework for analyzing a specific class of self-organizing systems: peer systems with shared environment, repeated interaction, finite resources, and local feedback. Within this restricted scope, we derive four interconnected laws. **Law Zero** establishes that cooperative response is the statistically dominant and dynamically stable configuration, with antagonistic behavior requiring continuous external perturbation to maintain. **Law I** shows that unobstructed cooperative dynamics yield efficient self-organization. **Law II** proves that perceived complexity requirements often reflect overconstraint rather than structural necessity. **Law III** demonstrates systematic observational bias arising from differential signal generation between functional and failure states.

The framework is diagnostic rather than universal: it provides structural criteria for analyzing when cooperation is generically favored and tools for identifying perturbations that maintain antagonism. We synthesize known results from evolutionary game theory, commons governance, and dynamical systems into a coherent diagnostic lens, rather than proving new theorems about all possible multi-agent systems. The central implication is a shift in explanatory burden: within the specified system class, antagonism requires explanation, not cooperation.

Keywords: self-organization, cooperation, game theory, commons governance, viability theory, multi-agent systems, perturbation analysis

1 Introduction

The study of self-organizing systems confronts three persistent theoretical problems: determining whether external coordination is necessary for stable organization, establishing minimal conditions for system viability, and accounting for systematic errors in system evaluation. This paper addresses all three within a restricted but well-defined system class, demonstrating their deep interconnection.

Scope restriction: The four laws developed here apply to *peer systems*—multi-agent systems where agents share environment, face symmetric viability constraints, engage in repeated interaction, possess local feedback, and operate under finite resource constraints. This excludes predator-prey dynamics, hierarchical authority structures, one-shot interactions, and systems with infinite resources. The restriction is not evasion but precision: different system classes follow different dynamics, and conflating them produces confused theory.

We proceed by first establishing foundational definitions, the mathematical apparatus, and explicit methodological commitments. Each law is then presented with structured argument, followed by derivation of corollaries and explicit statement of validity boundaries. The paper

concludes with a synthesis demonstrating how the four laws interact and cross-domain applications illustrating the framework's diagnostic utility.

1.1 Related Work and Theoretical Position

This framework synthesizes and extends several research traditions rather than claiming to supersede them:

Commons governance [Ostrom, 1990, 2010]: Ostrom's empirical work on successful self-governance of common-pool resources provides the primary evidential base for Law Zero. Her design principles for robust commons institutions can be reinterpreted as conditions that minimize perturbation P. We extend her framework by formalizing the distinction between intrinsic cooperative dynamics and perturbation-induced antagonism.

Evolutionary game theory [Axelrod, 1984, Nowak, 2006, Santos and Pacheco, 2005]: The five mechanisms supporting cooperation (kin selection, direct reciprocity, indirect reciprocity, network reciprocity, group selection) provide the micro-foundations for Law Zero. We do not claim to prove new theorems about evolutionary games, but to synthesize known results into a diagnostic framework. Importantly, evolutionary game theory also identifies conditions where defection is stable—we treat these as boundary conditions or P-encoded situations, not as refutations.

Tragedy of the commons [Hardin, 1968]: Hardin's analysis is often misread as claiming inevitable commons degradation. We reinterpret it as describing what happens when P (particularly artificial scarcity, exit barriers, or anonymity) is introduced into peer systems. The "tragedy" is not intrinsic but perturbation-dependent.

Viability theory [Aubin, 1991]: Law II draws on Aubin's mathematical framework for system persistence under constraints. We extend this by distinguishing essential from discretionary constraints and arguing that perceived complexity often reflects the latter.

Cybernetics and requisite variety [Ashby, 1956, Simon, 1962]: The bounded complexity of viable systems is anticipated in classical cybernetics. Law II formalizes this intuition and provides diagnostic criteria for identifying overconstraint.

Structural violence and elite-driven polarization: In human systems, the concept of P aligns with literatures on "divide and rule," structural violence, and manufactured scarcity. We provide formal grounding for these critiques without requiring conspiratorial assumptions—P can emerge from institutional incentives without intentional design.

What is novel: Not that cooperation exists, that institutions matter, or that observation is biased—these are established. The contribution is: (i) a unified structural account specifying when cooperation is generically favored within a well-defined system class; (ii) a diagnostic lens (P, C_E/C_D, signal asymmetry) applicable across domains; (iii) explicit formalization of "antagonism requires explanation" as a methodological principle for peer systems.

1.2 Methodological Commitments

This paper employs *semi-formal methodology*: mathematical notation is used to articulate structural relationships and draw out implications, not to deliver fully general theorems in the sense of pure mathematics.

Where we have genuine formal results (e.g., Law III's Bayesian derivation), they are labeled clearly. Where we are extrapolating from known models (replicator dynamics, threshold cascades, network games), we explicitly state: "in standard models of type X, this implies Y; we generalize this pattern as Law Z."

The "proofs" in Sections 3–6 are structured explanatory arguments with references to established results, not self-contained mathematical theorems. Readers expecting the latter will be disappointed; readers seeking a coherent diagnostic framework grounded in existing theory will, we hope, find value.

Empirical claims (e.g., that many systems of interest are Case A, that P is identifiable in antagonistic peer systems) are falsifiable but not proven here. We provide examples and point to literatures; systematic empirical validation is future work.

2 Foundational Framework

2.1 Basic Definitions

Definition 2.1 (System). *A tuple $S = (A, X, T, F)$ where A is a set of agents, X is a state space, T is an interaction topology, and $F : X \to X$ is a transition function.*

Definition 2.2 (Viability Region). *A subset $\Omega \subseteq X$ such that for all $x \in \Omega$, the system satisfies specified persistence conditions.*

Definition 2.3 (Response Function). *A mapping $\varphi : X \times A \to \mathbb{R}$ measuring the partial derivative of aggregate system output with respect to agent a's action at state x. Formally: $\varphi(x, a) = \partial W / \partial a_x$. The sign of φ is not assumed but will be derived.*

Definition 2.4 (External Constraint). *A function $C : X \to X$ imposed independently of agents' internal dynamics, modifying the transition function to $F' = C \circ F$.*

Definition 2.5 (Signal Function). *$\sigma : X \to \mathbb{R}_{\geq 0}$ mapping system states to observable information output.*

Definition 2.6 (Antagonistic Perturbation). *An external input $P : A \times A \to \mathbb{R}$ that modifies agent interaction payoffs such that $U'(a_i, a_j) = U(a_i, a_j) + P(a_i, a_j)$ where P creates zero-sum or negative-sum structure between agents. Critically, P must be structurally separable from the baseline game: it must be possible, at least in principle, to "turn off" P while keeping the underlying peer system intact. This excludes defining P as "anything that makes antagonism possible" (which would be circular). P is identified relative to an empirically grounded baseline—historically or experimentally observed cooperative dynamics under less-manipulated conditions.*

2.2 System Classification

2.2.1 Scope Restriction: Peer Systems

This paper's laws apply exclusively to peer systems, defined as follows:

Definition 2.7 (Peer System). *A system $S = (A, X, T, F)$ where all agents draw from the same resource pool E, face symmetric viability constraints, and possess comparable action capacities. Formally: (i) $\forall a_i, a_j \in A : \Omega(a_i) \simeq \Omega(a_j)$; (ii) $\forall a_i, a_j : \|capacity(a_i) - capacity(a_j)\| < \epsilon$ for some bound ϵ.*

Peer systems contrast with hierarchical systems (predator-prey, host-parasite, employer-employee) where agents occupy structurally asymmetric positions with different viability constraints and action capacities.

2.2.2 Why Predator-Prey is Not a Counterexample

A common objection to cooperation-as-default is the stability of predator-prey dynamics. We address this by distinguishing levels of analysis:

Level 1 (Intra-species): Wolves cooperate with wolves; deer cooperate with deer. Within each species, Law Zero applies: antagonism among wolves requires external perturbation (scarcity, territorial forcing). The peer system is wolf-wolf or deer-deer, not wolf-deer.

Level 2 (Inter-species): Wolf-deer is not a peer system. Wolves and deer have asymmetric viability constraints (Ω_{wolf} requires prey consumption; Ω_{deer} requires predator avoidance). They do not share a common resource pool in the relevant sense—the deer *is* the resource for wolves.

Level 3 (Ecosystem): At the ecosystem level, predator-prey dynamics *are* cooperative in a meta-sense: they regulate population, prevent overgrazing, maintain biodiversity. The "agents" at this level are species-populations, and their interaction is mutualistic for ecosystem viability.

The apparent counterexample dissolves under proper scope specification. Antagonism between wolves and deer is structural (different Ω), not perturbation-maintained. Law Zero addresses peer antagonism, not structural role differentiation.

2.2.3 Formal Criterion for Peer Status

Two agents a_i, a_j are peers if and only if:

$$(1)\ \Omega(a_i) \cap \Omega(a_j) \neq \emptyset \quad \text{(overlapping viability regions)} \tag{1}$$

$$(2)\ \exists E : \text{viability}(a_i) \text{ depends on } E \wedge \text{viability}(a_j) \text{ depends on } E \tag{2}$$

$$(3)\ a_i \text{ could occupy } a_j\text{'s role and vice versa (role symmetry)} \tag{3}$$

Wolves and deer fail condition (3): a deer cannot occupy the wolf role. Two wolves satisfy all three. Two firms in the same market satisfy all three. Two nations with comparable military capacity satisfy all three.

2.3 Mathematical Apparatus

We employ three formal frameworks in combination:

Game Theory: Models agent interactions through payoff matrices $U : A \times A \to \mathbb{R}$, with cooperative equilibria defined by mutual best-response conditions.

Constraint Satisfaction Theory: Formalizes viability as feasibility within constraint sets, with complexity measured by constraint count and coupling degree.

Information Theory: Quantifies signal asymmetry through entropy $H(\sigma(x))$ and sampling bias through conditional observation probabilities $P(\text{observe} \mid \text{state})$.

3 Law Zero: Cooperative Response as Emergent Default

3.1 The Problem

Previous formalizations of self-organization theory assume cooperative response functions ($\varphi > 0$) as a premise. This is unsatisfying: it appears to smuggle in the conclusion ("agents are good") as an assumption. We now argue that $\varphi > 0$ is not an assumption but a derivable property of unperturbed systems under minimal conditions.

3.2 Statement

Theorem 3.1 (Law Zero). *In a multi-agent peer system with shared environment, local feedback, and absence of externally imposed antagonistic perturbations, cooperative response ($\varphi > 0$) emerges as the statistically dominant and dynamically stable configuration. Antagonistic response ($\varphi < 0$) requires continuous external energy input to maintain.*

3.3 Formal Specification

Let $S = (A, X, T, F)$ be a system where agents share environmental variables $E \subseteq X$.

Define the baseline payoff structure:

$$U_0(a_i, a_j) = f(E, a_i) + g(E, a_i, a_j) \tag{4}$$

where f captures agent-environment interaction and g captures agent-agent interaction mediated by E.

Define the perturbed payoff structure:

$$U_P(a_i, a_j) = U_0(a_i, a_j) + P(a_i, a_j) \tag{5}$$

where P is an antagonistic perturbation (Definition 2.6).

3.4 Argument

3.4.1 Part I: Statistical Basin Asymmetry

We establish a statistical asymmetry in regime persistence for peer systems. This concerns the relative sizes of parameter regions (basins) supporting different equilibria—a claim from dynamical systems theory, not physical thermodynamics.

Step 1: Define the shared resource pool E as a scalar or vector representing aggregate environmental capacity available to all peers. In peer systems, all agents draw from E and contribute to E through their actions.

Step 2: Define extraction rate $r_i(a_i)$ and contribution rate $c_i(a_i)$ for each agent. The net effect on E is:

$$\frac{dE}{dt} = \sum_i [c_i(a_i) - r_i(a_i)] + R(E) \tag{6}$$

where $R(E)$ represents environmental regeneration (if any).

Step 3: Under cooperative dynamics, agents' strategies satisfy:

$$\sum_i c_i \geq \sum_i r_i - R(E) \quad \text{(sustainable extraction)} \tag{7}$$

Under antagonistic dynamics among peers, each agent maximizes individual r_i without coordinating on aggregate sustainability. This is a multi-agent tragedy of the commons.

Step 4: The key claim is not that antagonism is impossible, but that it occupies a narrower basin of stability in parameter space. Cooperative equilibria exist for a wider range of parameter values (regeneration rates, population sizes, discount factors). Antagonistic equilibria require either: (a) very high $R(E)$ making overextraction harmless, (b) very small populations making coordination trivial, or (c) external enforcement preventing defection.

Step 5: This is formalized via the "price of anarchy" concept from mechanism design. Let W^* be optimal social welfare and W_{Nash} be welfare at Nash equilibrium under selfish dynamics:

$$\text{PoA} = W^*/W_{\text{Nash}} \geq 1 \tag{8}$$

For peer systems with shared resources, $\text{PoA} > 1$ generically. Cooperative dynamics achieve W^*, antagonistic dynamics achieve $W_{\text{Nash}} < W^*$. Selection pressure favors systems that approach W^*.

Conclusion (Part I): In peer systems with shared finite resources, cooperative configurations occupy larger regions of parameter space and achieve higher aggregate welfare. This is a statistical claim about basin sizes, not a deterministic law. Stable antagonism is possible but requires special parameter conditions or external maintenance.

3.4.2 Part II: Game-Theoretic Argument

We do not claim that cooperation always emerges. We claim it is the attractor with the larger basin under specified conditions.

Step 6: The relevant game class for peer systems is not prisoner's dilemma alone, but the broader class of social dilemmas: situations where individual rationality conflicts with collective optimality.

Step 7: Key result from evolutionary game theory [Nowak, 2006]: Cooperation can be sustained through five mechanisms—kin selection, direct reciprocity, indirect reciprocity, network reciprocity, and group selection. In peer systems with repeated interaction and local information, at least direct and network reciprocity are operative.

Step 8: The critical parameter is the benefit-to-cost ratio b/c of cooperative acts relative to the number of interaction partners. When $b/c > k$ (where k depends on network structure), cooperation invades and stabilizes. In peer systems with shared environment, b is typically high (contributions to E benefit all) and c is bounded (individual contribution costs).

Step 9: Antagonistic strategies (defect, exploit) can be ESS in one-shot games or when $b/c < k$. But these conditions are precisely what peer systems with shared environment tend to violate: iteration is guaranteed by ongoing resource dependence, and shared-E structure amplifies b.

Step 10: We acknowledge hawk-dove, public goods games with punishment, and other models where defection or mixed strategies persist. These typically require: (a) one-shot or low-iteration interaction, (b) inability to identify defectors, (c) no exit option from bad partners, or (d) payoff structures where $b/c < k$. When these conditions hold, Law Zero does not apply—and this is captured in boundary conditions.

Qualification: The "larger basin" claim holds generically for commons-like games with shared resources and repeated interaction. In some social dilemmas (e.g., certain oligopolistic markets with differentiated products), mixed or mildly antagonistic equilibria may persist without obvious P. The claim is generic, not universal.

Conclusion (Part II): Among peer systems with repeated interaction, identifiable partners, and shared-environment coupling, cooperation is evolutionarily favored. This is the generic case for the system class we study.

3.4.3 Part III: Formal Characterization of Perturbation P

We now provide explicit mathematical characterization of perturbation P, addressing the need for rigor beyond examples.

Definition: Let $U_0 : A \times A \to \mathbb{R}$ be the baseline payoff matrix for peer interaction. P is a perturbation operator $P : (A \times A \to \mathbb{R}) \to (A \times A \to \mathbb{R})$ such that:

$$U_P = U_0 + P \tag{9}$$

P is antagonistic if it satisfies any of the following:

- (P1) Zero-sum injection: $P(a_i, a_j) + P(a_j, a_i) \leq 0$ for all $i \neq j$
- (P2) Relative payoff weighting: U_P ranks outcomes by $(u_i - u_j)$ rather than u_i alone
- (P3) Cooperation penalty: $P(\text{cooperate}, \text{cooperate}) < 0$
- (P4) Defection subsidy: $P(\text{defect}, \text{cooperate}) > 0$

Effect on dynamics: Under replicator equation $dx_i/dt = x_i(f_i - \bar{\varphi})$, introducing P shifts the fitness landscape:

$$f_i^P = f_i + \sum_j x_j P(i, j) \tag{10}$$

This can create new fixed points, destabilize cooperative equilibria, or shift basin boundaries. The key claim: these P-induced antagonistic equilibria are conditionally stable. Setting $P = 0$ removes the artificial fixed points, and the system returns to U_0 dynamics.

Examples formalized:

- *Artificial scarcity:* $P(a_i, a_j) = -\gamma \cdot \text{overlap}(\text{demand}_i, \text{demand}_j)$ where $\gamma > 0$ is scarcity intensity. Threshold: cooperation destabilizes when $\gamma > \gamma^* = (b - c)/\text{overlap}_{\max}$.

- *Relative performance incentives:* $P(a_i, a_j) = \beta(u_i - u_j)$ where $\beta > 0$ weights relative standing. Threshold: cooperation destabilizes when $\beta > \beta^* = b/(b+c)$.
- *Exit barriers:* Discount rate δ is artificially reduced, shrinking the shadow of the future. Threshold: cooperation via reciprocity requires $\delta > \delta^* = (T-R)/(T-P)$ in standard PD notation.
- *Information suppression:* Observation probability $p \to 0$, disabling tit-for-tat. Threshold: indirect reciprocity requires $p > p^* \approx 1/k$ where k is group size [Nowak and Sigmund, 2005].

These thresholds are illustrative and context-dependent, but they demonstrate that P effects can be quantified, making predictions about cooperation breakdown testable.

3.4.4 Part IV: Perturbation Dynamics and Relaxation

Step 11: Given Parts I–III, consider a peer system exhibiting stable antagonism. By the above arguments, this is not the generic attractor state. What maintains it?

Step 12: From the formal characterization, antagonistic equilibria exist when P shifts the fitness landscape to create stable fixed points in the defection/exploitation region.

Step 13: When P is externally maintained (by institutional design, resource control, information management), these fixed points persist. When P weakens or is removed, the landscape reverts to U_0, and the system relaxes toward its natural attractor.

Step 14: Relaxation dynamics follow:

$$\frac{d\varphi}{dt} = -\nabla_\varphi V(U_0) + \epsilon(t) \tag{11}$$

where V is a potential function for U_0 dynamics and $\epsilon(t) \to 0$ as $P \to 0$. The cooperative equilibrium is a minimum of V under U_0.

Step 15: This generates the diagnostic: when observing stable antagonism in what appears to be a peer system, search for P. Formalized candidates: institutional rules creating (P1)–(P4), resource distribution mechanisms, information control systems, and ideological structures that frame peer relations as zero-sum.

3.5 Corollary

Stable antagonism among peers is diagnostic of one of four conditions:

(a) *Misclassification:* the system is not actually a peer system (asymmetric roles, different Ω).

(b) *Transience:* the system has not yet equilibrated (e.g., post-shock, early-stage).

(c) *Active perturbation:* $P \neq 0$ is maintaining antagonistic equilibrium.

(d) *Boundary violation:* conditions for Law Zero do not hold (one-shot, no identification, infinite resources).

The claim is not that antagonism cannot exist, but that in properly specified peer systems, its stability requires explanation. Cooperation requires no such external maintenance.

3.6 Boundary Conditions

Law Zero applies when ALL of the following hold:

Peer structure: agents have symmetric viability constraints and comparable capacities

Shared environment: a common resource pool E couples agents' fates

Repeated interaction: iteration count is high or indefinite

Identifiability: agents can distinguish partners and track history

Finite resources: E is bounded, making overextraction costly

Local feedback: agents observe effects of their actions on E and on partners

Law Zero does NOT apply when:

(a) Agents are structurally asymmetric (predator-prey, host-parasite, hierarchical authority)

(b) Interaction is one-shot or very low iteration

(c) Partners are anonymous and history cannot be tracked

(d) Resources are effectively infinite (no commons problem)

(e) Feedback is absent or heavily delayed

3.6.1 Resource Scarcity: Boundary Condition vs. Perturbation

A critical question: is extreme resource scarcity a violation of boundary conditions (making the system non-peer) or a form of perturbation P (artificial scarcity)?

We distinguish three regimes:

Regime 1 (Abundance): Resources exceed aggregate demand. Cooperation is trivial—no conflict over resources. Law Zero holds but is uninteresting.

Regime 2 (Moderate scarcity): Resources are finite but sufficient for viability if managed cooperatively. This is the core domain of Law Zero.

Regime 3 (Extreme scarcity): Resources are insufficient for all agents' viability regardless of behavior. This is a boundary violation, not perturbation.

The threshold:

$$\text{Regime 2:} \quad \sum_i \min\_needs(a_i) \leq E \leq \sum_i \max\_wants(a_i) \tag{12}$$

$$\text{Regime 3:} \quad E < \sum_i \min\_needs(a_i) \tag{13}$$

In Regime 3, even perfect cooperation cannot sustain all agents. This is not Law Zero failing; it is the system exiting Law Zero's domain.

Artificial scarcity as P: Critically, P can convert Regime 1 or 2 into *apparent* Regime 3. When resources are sufficient but access is restricted (enclosure, patents, credentialism, hoarding), agents experience scarcity that is not intrinsic but imposed. This artificial scarcity is paradigmatic P—removable in principle, converting cooperative-capable systems into zero-sum competitions.

3.6.2 The Gray Zone: Oscillation Between Regimes

Real-world systems rarely occupy stable regimes. Resources fluctuate due to environmental variation, population changes, and external shocks.

Define the scarcity ratio:

$$S(t) = \frac{\sum_i \min\_needs(a_i)}{E(t)} \tag{14}$$

Regime 2 holds when $S < 1$; Regime 3 when $S > 1$. The gray zone is $S \approx 1$.

Gray zone dynamics include:

- *Latent antagonism:* When S approaches 1 from below, agents may anticipate future scarcity and shift toward defensive strategies. As $S \to 1^-$: cooperation becomes increasingly unstable, $\varphi \to 0$.

- *Hysteresis:* Systems that have experienced Regime 3 may retain antagonistic patterns even after returning to Regime 2. This is trauma-induced P_{int}.
- *Oscillation-induced breakdown:* If S oscillates across 1 repeatedly and oscillation period $T <$ relaxation time τ, the system exhibits sustained low cooperation despite average $S < 1$.

3.7 Empirical Predictions

The law generates falsifiable predictions:

P1 (Perturbation Removal): Removing artificial competitive structures will shift φ upward.

Example: Ostrom's irrigation communities showed high cooperation ($\varphi \approx 0.85$–0.95) under self-governance, declining when state intervention introduced external allocation rules [Ostrom, 1990].

P2 (Perturbation Identification): Stable peer antagonism implies identifiable P.

P3 (Historical Transitions): Cooperation emergence follows P reduction, not agent composition change.

Example: Open-source software communities emerged when artificial scarcity of code was removed by licensing changes.

P4 (Cross-System Comparison): Weaker P correlates with higher φ.

Example: Ostrom's meta-analysis of 91 irrigation systems: local self-governance (low P) showed 67% cooperation rates vs. 27% under external management (high P).

P5 (Natural Experiments): P removal produces cooperation, not chaos.

3.8 The Freeloader Problem: Tolerance, Not Enforcement

A persistent objection to cooperation-as-default is the freeloader problem: if cooperation is voluntary, rational agents will exploit cooperators. We argue that robust cooperative systems solve this not through engagement (punishment, exclusion, monitoring) but through *structural tolerance*.

3.8.1 Mechanisms of Tolerance

Mechanism 1: Redundancy and Surplus. If the system produces surplus, freeloaders consume slack without threatening viability.

Example: Open-source software—90%+ of users contribute nothing; the system thrives on contributions from $< 1\%$ of users.

Mechanism 2: Locality and Decoupling. In network-structured systems, freeloaders affect only their local neighborhood:

$$\text{Impact}(a_f) \propto \text{degree}(a_f)/N, \text{ not } \propto 1 \tag{15}$$

Mechanism 3: Frequency-Dependent Limitation. Freeloading is self-limiting: $\exists f^* < 1$ such that freeloader frequency stabilizes at f^*.

Mechanism 4: Exit and Restructuring. Dynamic networks self-organize to exclude persistent antagonists [Santos et al., 2006, Perc and Szolnoki, 2010].

Mechanism 5 (Primary): Strategic Ignoring. The most powerful dampening mechanism is systematic non-engagement:

$$\text{cost}(\text{ignore}) = 0; \quad \text{cost}(\text{punish}) > 0; \quad \text{cost}(\text{exclude}) > 0 \tag{16}$$

Ignoring works through:

- *Attention denial:* Antagonism often seeks response; non-response denies the payoff.
- *Energy asymmetry:* Sustaining antagonism requires continuous expenditure; ignoring forces exhaustion.
- *Contagion blocking:* Antagonism spreads through reaction chains; ignoring breaks the chain.
- *Status denial:* In social systems, ignoring denies status boost from provocation.
- *Recruitment failure:* Movements grow by provoking overreaction; ignoring denies recruitment material.

3.8.2 Tolerance vs. Enforcement in Low-Surplus Systems

Define the surplus ratio:

$$\sigma = \frac{\sum_i c_i - \sum_j r_j}{\sum_j r_j} \tag{17}$$

The tolerance-enforcement tradeoff:

$$\text{If } C_T(\sigma) < C_E: \quad \text{tolerate} \tag{18}$$

$$\text{If } C_T(\sigma) > C_E: \quad \text{enforce} \tag{19}$$

where C_T = cost of tolerance, C_E = cost of enforcement (monitoring + judgment + punishment + errors + second-order effects + cultural costs + arms race costs).

Empirical pattern: Even in low-surplus systems, C_E often exceeds C_T because enforcement costs are systemic while freeloading costs are marginal [Ostrom, 1990].

3.9 Critical Mass Dynamics and System Resilience

A distinct threat from freeloading is aggressive contagion: a sufficiently large cluster of antagonistic agents can trigger cascading defection.

Define the cascade threshold:

$$\alpha^* = 1/\lambda_{\max}(A) \tag{20}$$

where $\lambda_{\max}(A)$ is the largest eigenvalue of the adjacency matrix.

Why small insurrections fail: Cooperative systems possess natural dampening mechanisms:

1. Local majority effects (network reciprocity)
2. Reputation and memory (quarantine without punishment)
3. Proportional response (tit-for-tat contains rather than escalates)
4. Exit and restructuring (dynamic networks self-organize)
5. Strategic ignoring (zero-cost, maximal effectiveness)

3.10 Internalized Perturbation and Relaxation Lag

A critical limitation: in human systems, P can be *internalized* through cultural transmission.

Define internalized perturbation:

$$U_{\text{int}}(a_i, a_j) = U_0(a_i, a_j) + P_{\text{int}}(a_i, a_j) \tag{21}$$

where P_{int} persists even after external P_{ext} is removed.

Relaxation dynamics:

$$\frac{d\varphi}{dt} = -k_1(\varphi - \varphi^*_{\text{coop}}) - k_2 \cdot P_{\text{int}}(t) \tag{22}$$

$$\frac{dP_{\text{int}}}{dt} = -k_3 \cdot P_{\text{int}} + k_4 \cdot (\text{social transmission}) \tag{23}$$

Two timescales emerge:

- $\tau_1 = 1/k_1$: behavioral adaptation (fast)
- $\tau_2 = 1/k_3$: cultural decay (slow, generational)

If $\tau_2 \gg \tau_1$, the system exhibits *relaxation lag*: behaviors may shift toward cooperation, but underlying P_{int} maintains latent antagonism.

Distinguishing P_{int} from native preferences:

1. Historical traceability to prior P_{ext}
2. Cross-cultural variation
3. Decay under non-reinforcement
4. Narrative dependency
5. Counterfactual test

4 Law I: Self-Organization Efficiency (Corollary)

Theorem 4.1 (Law I). *Given the cooperative default established in Law Zero, external constraints that do not track local information generally reduce achievable welfare relative to unconstrained dynamics.*

Law I is a corollary of Law Zero, not an independent law. It articulates what follows from $\varphi > 0$ in the unperturbed regime when external constraints are introduced.

Argument: By Law Zero, $\varphi > 0$ in unperturbed systems. Introduce external constraint C operating independently of local information. If C improved outcomes universally, then $\forall x \in X : W(C(x)) > W(x)$, implying agents' response functions are uniformly suboptimal—contradicting $\varphi > 0$ derived in Law Zero.

Therefore $\exists x^*$ such that $W(x^*) \geq W(C(x^*))$, establishing that unconstrained dynamics achieve at least as high welfare as constrained dynamics for some states.

5 Law II: Viability Sufficiency

Theorem 5.1 (Law II). *For systems with threshold-defined viability (Case A), constructing conditions sufficient for viability does not require high systemic complexity. Observed necessity for extreme complexity is diagnostic of discretionary overconstraint, not structural requirement.*

This is one of the framework's more philosophically interesting claims. The formal structure (C_E vs C_D) is nearly definitional; the real content lies in the empirical assertion that many systems of interest are Case A and that we systematically overconstrain them.

Case A (Threshold Viability): $\Omega = \{x : f_i(x) \geq \theta_i \text{ for } i = 1..k\}$ where k is small. This covers most biological, infrastructural, and organizational persistence requirements.

Case B (Optimization Viability): Ω requires continuous coordination across high-dimensional state variables. This occurs when viability is defined relative to competitors or adversaries.

Empirical support for bounded $|C_E|$: Viability theory [Aubin, 1991] formalizes survival as satisfaction of a small constraint set. Organizational studies [Simon, 1962, Ashby, 1956] show that viable organizations satisfy "requisite variety" through bounded regulatory mechanisms. Ecological viability requires satisfaction of typically 3–7 fundamental constraints.

Diagnostic: When analyzing complex systems, ask: can constraints be removed without exiting Ω? If yes, apparent complexity exceeds true complexity.

6 Law III: Observational Asymmetry

Theorem 6.1 (Law III)**.** *In any system where functional states generate low observational signal and failure states generate high observational signal, external evaluations based on observed information systematically underestimate functionality and overestimate failure frequency.*

Partition $X = X_F \cup X_{\neg F}$ (functional and non-functional states).

Define signal asymmetry:

$$\alpha = \frac{\mathbb{E}[\sigma(x) \mid x \in X_{\neg F}]}{\mathbb{E}[\sigma(x) \mid x \in X_F]} > 1 \tag{24}$$

Define observation probability: $P(\text{observe } x) \propto \sigma(x)$.

Proof: By Bayes' theorem:

$$P(x \in X_F \mid \text{observed}) = \frac{P(\text{observed} \mid x \in X_F) \cdot P(x \in X_F)}{P(\text{observed})} \tag{25}$$

From signal asymmetry $\alpha > 1$:

$$P(\text{observed} \mid x \in X_{\neg F}) = \alpha \cdot P(\text{observed} \mid x \in X_F) \tag{26}$$

Therefore:

$$\hat{\rho} = \frac{\rho}{\rho + \alpha(1 - \rho)} < \rho \tag{27}$$

where $\rho = P(x \in X_F)$ is true functionality ratio and $\hat{\rho}$ is observed ratio. The bias is strictly positive for $\alpha > 1$ and $0 < \rho < 1$. □

Limitation: If observers are aware of bias and correct ex ante, the simple form no longer holds. However, correction requires awareness, knowledge of α, and cognitive resources—conditions often unmet.

7 Synthesis and Interconnections

The four laws form a coherent theoretical structure:

- **Law Zero** establishes that cooperative response is the statistically dominant default for peer systems.
- **Law I** shows that this cooperation yields efficient self-organization when unobstructed.
- **Law II** demonstrates that viability does not require the complexity often claimed.

- **Law III** explains why successful cooperative systems remain invisible while failures dominate perception.

The Central Insight: The critical innovation is Law Zero's derivation of $\varphi > 0$ rather than its assumption. This transforms the framework from a conditional claim ("if agents are cooperative, then...") into a structural claim ("agents in shared environments converge to cooperation unless externally perturbed"). The burden of proof shifts: antagonistic behavior requires explanation (what is P?), not cooperative behavior.

Unified Theorem: A multi-agent peer system with shared environment, repeated interaction, and finite resources, operating within a threshold viability region and subject to signal-asymmetric observation, will (a) converge to cooperative dynamics when unperturbed, (b) achieve efficient self-organization in this regime, (c) require only low complexity for viability, yet (d) be systematically mischaracterized as unstable, complex-dependent, and failure-prone.

8 Validity Boundaries

The unified theory does not apply when:

1. Agents do not share environment (no coupling of fates)
2. Interaction is one-shot or feedback channels are absent
3. Resources are effectively infinite (no cost to dissipation)
4. External perturbation P is permanent and inescapable
5. The viability region is intrinsically high-dimensional (Case B)
6. Observers have full continuous state access

Under these conditions, antagonistic behavior may be stable, external coordination may be necessary, complexity may be structural, and evaluations may be accurate.

9 Conclusion

This paper has presented four formally grounded laws governing self-organizing peer systems. The central result is Law Zero: cooperative response emerges as the default configuration in shared-environment systems, with antagonistic behavior requiring continuous external perturbation to maintain. This transforms the traditional framing from "cooperation requires explanation" to "antagonism requires explanation."

The theoretical contribution lies in deriving $\varphi > 0$ rather than assuming it, thereby closing the gap between formal systems theory and empirical observation of cooperative dynamics. The practical contribution lies in providing a diagnostic methodology: when observing antagonistic or dysfunctional systems, search for P.

The laws are conditional and falsifiable. Their boundary conditions are explicitly stated. The framework generates testable predictions about what happens when perturbations are removed from antagonistic systems.

Future work may quantify perturbation magnitudes required to maintain antagonism, develop protocols for identifying P in observed systems, and apply the diagnostic methodology to specific domains including AI systems, biological organization, and human institutions.

References

Ashby, W. R. (1956). *An Introduction to Cybernetics*. Chapman & Hall.

Aubin, J.-P. (1991). *Viability Theory*. Birkhäuser.

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

Axelrod, R. and Hamilton, W. D. (1981). The evolution of cooperation. *Science*, 211(4489):1390–1396.

Beissinger, S. R. and McCullough, D. R., editors (2002). *Population Viability Analysis*. University of Chicago Press.

Bowles, S. and Gintis, H. (2011). *A Cooperative Species: Human Reciprocity and Its Evolution*. Princeton University Press.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. Wiley.

Fehr, E. and Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, 90(4):980–994.

Fehr, E. and Gächter, S. (2002). Altruistic punishment in humans. *Nature*, 415(6868):137–140.

Hardin, G. (1968). The tragedy of the commons. *Science*, 162(3859):1243–1248.

Kauffman, S. A. (1993). *The Origins of Order*. Oxford University Press.

Leibo, J. Z., Zambaldi, V., Lanctot, M., Marecki, J., and Graepel, T. (2017). Multi-agent reinforcement learning in sequential social dilemmas. In *Proceedings of AAMAS*, pages 464–473.

Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P., and Mordatch, I. (2017). Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in Neural Information Processing Systems*, 30.

Margulis, L. (1998). *Symbiotic Planet: A New Look at Evolution*. Basic Books.

Nadell, C. D., Xavier, J. B., and Foster, K. R. (2009). The sociobiology of biofilms. *FEMS Microbiology Reviews*, 33(1):206–224.

Nowak, M. A. (2006). Five rules for the evolution of cooperation. *Science*, 314(5805):1560–1563.

Nowak, M. A. and Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, 437(7063):1291–1298.

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

Ostrom, E. (2010). Beyond markets and states: Polycentric governance of complex economic systems. *American Economic Review*, 100(3):641–672.

Pennisi, E. (2009). On the origin of cooperation. *Science*, 325(5945):1196–1199.

Perc, M. and Szolnoki, A. (2010). Coevolutionary games—a mini review. *BioSystems*, 99(2):109–125.

Rand, D. G., Dreber, A., Ellingsen, T., Fudenberg, D., and Nowak, M. A. (2009). Positive interactions promote public cooperation. *Science*, 325(5945):1272–1275.

Roughgarden, J., Oishi, M., and Akçay, E. (2006). Reproductive social behavior: Cooperative games to replace sexual selection. *Science*, 311(5763):965–969.

Santos, F. C. and Pacheco, J. M. (2005). Scale-free networks provide a unifying framework for the emergence of cooperation. *Physical Review Letters*, 95(9):098104.

Santos, F. C., Pacheco, J. M., and Lenaerts, T. (2006). Evolutionary dynamics of social dilemmas in structured heterogeneous populations. *Proceedings of the National Academy of Sciences*, 103(9):3490–3494.

Simon, H. A. (1962). The architecture of complexity. *Proceedings of the American Philosophical Society*, 106(6):467–482.

Traulsen, A. and Nowak, M. A. (2006). Evolution of cooperation by multilevel selection. *Proceedings of the National Academy of Sciences*, 103(29):10952–10955.

Wilson, D. S. and Wilson, E. O. (2007). Rethinking the theoretical foundation of sociobiology. *Quarterly Review of Biology*, 82(4):327–348.

A Cross-Domain Applications

The four laws apply across multiple domains where peer agents share environments and interact repeatedly. This appendix briefly illustrates applicability to five domains.

Important caveat: These applications are illustrative, not exhaustive proofs. Each domain has unique features that a general treatment cannot fully address. We present these as invitations for domain experts to test, refine, or refute the framework.

A.1 Multi-Agent AI Systems

AI agents in shared environments (computational resources, data streams, user attention) can exhibit peer structure. Law Zero predicts cooperative protocols emerge without zero-sum objectives [Leibo et al., 2017]. Perturbation P includes competitive benchmarking and adversarial training objectives.

Caveats: AI cooperation is highly sensitive to reward design. Much observed cooperation is explicitly designed, not emergent. Training timescales differ radically from evolutionary timescales.

A.2 Cellular Organelles

Endosymbiotic theory [Margulis, 1998] describes integration of formerly independent organisms into cooperative cellular systems. Organelles share cytoplasmic environment and face symmetric viability constraints. Historical P (independent reproduction) was eliminated through gene transfer, enabling deep cooperation.

A.3 Bacterial Colonies

Bacteria in biofilms share nutrient environment and exhibit division of labor, resource sharing, and quorum sensing [Nadell et al., 2009]. Perturbation P (antibiotics, resource stress) shifts dynamics toward antagonism; removal restores cooperation.

A.4 Social Insects

Worker insects are peers within colonies (shared nest environment, symmetric constraints). Law Zero applies to worker-worker relations; queen-worker is hierarchical. Colonies with higher genetic relatedness (lower P) show higher cooperation.

A.5 Human Organizations

Humans in peer relationships develop cooperative norms absent P [Bowles and Gintis, 2011]. Perturbation includes artificial scarcity, competitive ranking, information asymmetry. Critically, P can be internalized through culture, creating relaxation lag after P_{ext} removal.