Adaptive Trust Dynamics Exhibit Hysteresis

Implications for AI Governance and Institutional Design

Flyxion

October 2025

Abstract

Coordination among intelligent agents requires permeable boundaries enabling mutual correction. We model trust as adaptive coupling in networks of agents maintaining intelligibility fields under noise. When coupling strength adapts to observed disagreement while incurring maintenance costs, the system exhibits **bistability**: identical parameters support both coherent (high-trust, low-variance) and fragmented (low-trust, high-variance) equilibria. **Hysteresis** reveals that increasing precautionary costs can irreversibly trap systems in fragmented states even after costs are removed. We demonstrate this in minimal simulations (30 agents, 5 parameters) and show hysteresis width depends on network topology. These dynamics formalize the **paradox of precaution**: safety measures reducing trust permeability create self-fulfilling coordination failures. We interpret this as **entropy-bounded recursion**—a general principle governing self-referential systems from AI governance to institutional design. Implications for AI alignment, market coordination, and polarization dynamics are discussed.

Contents

1 Introduction

1.1 The Precautionary Paradox

Standard safety reasoning advocates reducing risk through constraints, monitoring, and isolation of potentially dangerous agents. However, a recursive challenge arises: oversight mechanisms themselves rely on general intelligences, such as humans, to function effectively. The central claim of this work is that precautionary measures diminishing trust permeability may entrap systems in stable yet suboptimal equilibria. This mechanism is previewed through adaptive coupling combined with maintenance costs, leading to bistability and hysteresis.

1.1.1 Recursion as a Trust Loop (Extended)

Recursive systems that monitor themselves face the same thermodynamic constraint as self-referential code or meta-markets: each layer of oversight consumes the bandwidth it seeks to guarantee. In RSVP terms, this is an *entropy-bounded recursion*: the very act of stabilizing a field introduces curvature that must itself be stabilized. When vigilance layers accumulate faster than information dissipates, the result is hysteresis—a semantic overgrowth of control.

This parallels the curvature-control inequalities derived in *Curvature, Entropy, and Governance*, where recursive futarchy required curvature contraction for stability. The present trust model's feedback law,

$$\dot{\kappa} = \eta(\langle(\Delta\Phi)^2\rangle - \sigma^2) - \gamma\kappa, \tag{1}$$

is the social analogue: a bounded-recursion operator whose damping term ($-\gamma\kappa$) defines the entropy cost of self-reference.

1.2 Related Work

Coordination games and equilibrium selection, including stag hunt and assurance games, as well as convention formation (Schelling, Sugden), provide foundational insights. Trust dynamics are explored in works on cooperation (Axelrod), commons governance (Ostrom), and cultural evolution (Henrich). Network synchronization models, such as Kuramoto oscillators, consensus protocols, and flocking models, inform the approach. Adaptive networks involve coevolution of structure and state (Gross & Blasius, Holme & Newman). Bistability in social systems appears in polarization models (Axelrod, Castellano) and opinion dynamics (Deffuant, Hegselmann-Krause). AI alignment literature addresses corrigibility (Soares et al.), value learning (Russell), and multi-agent coordination.

A gap exists in prior research, where trust is often treated as exogenous or where structure evolves independently of disagreement. This model couples adaptive trust to observed misalignment with explicit costs, yielding novel bistability.

1.3 Contributions

The contributions include: (1) a minimal model exhibiting hysteresis in trust-coordination dynamics; (2) a phase diagram mapping parameter space to coherence, fragmentation, and bistability regimes; (3) effects of network topology, comparing ring, small-world, and scale-free structures; (4) empirical signatures, such as relaxation time divergence near transitions indicating critical slowing down; and (5) policy implications regarding when precautionary measures become self-defeating.

2 Model

2.1 Agent Dynamics

Consider n agents maintaining intelligibility fields $\Phi_i \in \mathbb{R}$. The dynamics follow Laplacian coupling: $\dot{\Phi} = -\kappa L \Phi + \xi(t)$, where L denotes the graph Laplacian, $\kappa \geq 0$ represents trust coupling strength (shared permeability), and $\xi(t)$ is white noise with amplitude D signifying exogenous perturbations.

Interpretation: Φ_i corresponds to intelligibility, semantic capacity, or opinion; $\kappa L \Phi$ provides diffusive correction toward neighbors; the objective is to minimize variance $\mathrm{Var}(\Phi)$ while preserving individual identity.

For simplicity, coupling is initially uniform ($\kappa_{ij} = \kappa$ for all edges). Heterogeneous cases (κ_{ij}) are considered in Section **??**.

2.2 Adaptive Coupling

$$\dot{\kappa} = \eta \left(\langle (\Delta\Phi_{ij})^2 \rangle - \sigma^2 \right) - \gamma\kappa, \quad \kappa \geq 0 \tag{2}$$

where $\eta > 0$ is responsiveness to misalignment (opens trust when gradients are large), $\sigma^2 > 0$ is the tolerance threshold (target disagreement level), $\gamma > 0$ is maintenance cost (surveillance overhead, bureaucratic drag, forgetting), and $\langle (\Delta\Phi_{ij})^2 \rangle$ is the mean squared difference between connected agents.

Intuition: Trust increases when disagreement exceeds tolerance but decays due to costs. This creates feedback: high κ → low variance → low growth signal → decay dominates → κ drops → variance grows.

2.2.1 Connection to Entropy-Bounded Recursion

The adaptive-trust equation is formally equivalent to the dissipative form of a BV-action in self-referential field theory. The parameters have direct recursive analogues:

- η: responsiveness of recursion—how quickly a system expands interpretive scope in response to misfit;
- γ: entropy cost—the curvature penalty for maintaining recursive vigilance;
- bistability: coexistence of open recursion (creative coherence) and frozen recursion (bureaucratic lock-in).

This mirrors the bounded-recursion inequality from *Recursive Futarchy*:

$$\partial_t^2 \Phi + \alpha \, \partial_t \Phi + \beta \, \Phi < 0, \tag{3}$$

ensuring curvature contraction and finite recursion depth. When γ exceeds the system's negentropic capacity ($\eta\sigma^2/\gamma > \eta_{\text{crit}}$), recursive oversight becomes self-destructive—the same mechanism underlying trust hysteresis.

2.2.2 Recursive Interpretation of Parameters

In cognitive and institutional networks, coupling κ represents recursion depth—how many layers of mutual modeling can coexist before semantic coherence collapses. In *Revenge of the Vorticons*, recursive depth was quantified as

$$D = \sum_i w_i(\Delta\Phi_i \Delta S_i), \tag{4}$$

linking immersion, entropy, and curvature. Increasing precautionary cost γ effectively limits this D, truncating feedback and reducing the system's memory capacity.

Attentional Cladistics described this as the *care-domestication threshold*: excessive stabilizing feedback (care, control) freezes evolutionary creativity. The same phenomenon reappears here as *precautionary lock-in*.

2.3 Dimensionless Form & Parameter Space

Non-dimensionalization by time scale $1/\gamma$ yields control parameters: $\tilde{\eta} = \eta\sigma^2/\gamma$ (potential openness), $\tilde{D} = D/(\gamma\sigma^2\lambda_2)$ (noise relative to connectivity), with network topology via $\lambda_2(L)$ (algebraic connectivity).

Prediction: Bistability occurs in intermediate regimes of $\tilde{\eta}$ and $\tilde{D}$, avoiding forced coherence or fragmentation.

2.4 Network Topologies

- **Ring**: $\lambda_2 \approx 4\sin^2(\pi/n) \ll 1$ (low connectivity, symmetric). - **Small-world** (Watts-Strogatz): Intermediate λ_2, shortcuts enhance coherence. - **Scale-free** (Barabási-Albert): Hub-dominated, high λ_2 but vulnerable to targeted removal.

3 Results

3.1 Hysteresis in Trust-Variance Phase Space

Setup: Ring network, $n = 30$, sweep $\gamma \in [0.2, 1.5]$ slowly (200 timesteps per value), then reverse.

Observations: - Low $\gamma < 0.5$: Unique coherent attractor (variance ~ 0.2, $\kappa \sim 1.5$). - High $\gamma > 1.0$: Unique fragmented attractor (variance ~ 2.5, $\kappa \sim 0.1$). - Intermediate $0.5 < \gamma < 1.0$: Bistability, with sweep up (blue) maintaining coherence past fragmentation threshold, sweep down (red) sustaining fragmentation below coherence threshold, yielding hysteresis width $\Delta\gamma \approx 0.3$.

Geometric Analogy: The hysteresis loop maps onto curvature cycles within an informational manifold. Increasing γ introduces excess curvature—analogous to recursive overregulation—while decreasing it cannot immediately flatten the manifold, leaving the