The Paradox of Precaution

How AGI Safety Could Erode Human Trust

A Thermodynamic Theory of Mutual Corrigibility

Extended Version

Flyxion

October 2025

Revised June 2026

When the machinery of safety becomes the machinery of suspicion.

Abstract

Efforts to prevent hypothetical catastrophe from artificial general intelligence increasingly depend on centralization, surveillance, and epistemic control. Yet such precautions replicate the very failure modes they seek to avoid: they suppress transparency, amplify paranoia, and erode the feedback loops that make human cooperation stable. This paper argues that the principal danger of AGI precautionism lies not in the machines themselves, but in the social thermodynamics of mistrust it institutionalizes. Drawing on the Relativistic Scalar–Vector Plenum (RSVP) framework, it models alignment as a process of open entropic exchange—showing that excessive restriction collapses the coupling coefficients that sustain mutual corrigibility.

The extended formal treatment develops interconnected results across five appendices. Appendix A establishes the RSVP field interpretation of alignment as phase coherence and trust as controlled entropic permeability. Appendix B derives a Trust Lagrangian whose stationarity conditions govern optimal coupling, and proves a sufficient coherence-window condition bounding the institutional mistrust parameter γ before paranoia transition. Appendix C applies sheaf theory to show that global alignment is epistemically underdetermined by isolated local data, that precautionary chokepoints generate nonzero Čech obstruction classes, and introduces the Institutional Fragility Index $\mathfrak{F} = \dim \check{H}^1(X, \mathcal{F})$ as a measurable topological governance quantity. Appendix D introduces the trust entropy $\mathcal{T}(x) = \log |\mathcal{A}(x)|$ as a unified geometric invariant encoding corrigibility as the logarithmic measure of admissible future trajectories, and proves: a monotonicity theorem connecting the coherence ratio to trust entropy growth; a trust current conservation law in field-theoretic form; the existence of corrigible equilibria via Brouwer; a coupling-collapse theorem showing that $\kappa_{ij} \to 0$ eliminates admissible trajectories; a projection-induced blindness result; a zero-diversity theorem; a spectral fragmentation threshold via the Fiedler eigenvalue; a safe-reform homotopy criterion; and a spectral trust bound conjecturing $\mathcal{T}(x) \leq C_0 + C_1 \log \lambda_2(L)$.

A concluding section before the epilogue states the Universal Precaution Paradox as a domain-independent theorem: in any adaptive system whose stability depends on corrective exchange, effective stability is a non-monotone function of precaution, decreasing past a critical threshold P_c. This elevates the paper from an AI-governance essay to a formal statement of a principle appearing across institutions, ecosystems, immune systems, scientific communities, and memory architectures.

Contents

Part I: The Mirror of Precaution **2**

Part II: The Paradox of Precaution **2**

Part III: From Political Argument to Field Degeneracy **2**

Appendix A: Alignment as Entropic Coupling **2**

Appendix B: A Trust Lagrangian for Mutual Corrigibility **2**

Appendix C: Sheaf-Theoretic Obstruction to Isolated Alignment **2**

Appendix D: Continuation and the Geometry of Trust **2**

Appendix E: Category of Corrigible Systems **2**

The Universal Precaution Paradox **2**

0.1 The Unalignability of Human Oversight . . . 3
0.2 How Safety Mechanisms Reproduce Mistrust . . . 3
0.3 The Category Error in AGI Catastrophism . . . 3
0.3.1 Optimization Capacity vs. Ontological Alienness . . . 3
0.3.2 Hobbesian Rationality vs. Ecological Stability . . . 3
0.3.3 Intelligence as Structured Process . . . 3
0.4 Recursive Alignment, Not Static Control . . . 4
0.4.1 Alignment Through Cultivation, Not Axiomatization . . . 4
0.4.2 Alignment as Thermodynamic Equilibrium . . . 4
0.4.3 From Omnipotence to Entanglement . . . 4
0.5 The Mirror Problem . . . 4
0.5.1 Human Cooperation as Evidence . . . 4
0.5.2 The Reflexivity of Trust . . . 4
0.5.3 Coexistence as Default Attractor . . . 4
0.6 Toward an Ecology of Intelligence . . . 4
0.6.1 AGI as Trophic Layer . . . 4
0.6.2 Principles of Integration . . . 4
0.6.3 From Control to Co-Evolution . . . 5
0.7 Conclusion: Precaution as a Self-Fulfilling Disalignment . . . 5
0.8 The Double Edge of Precaution . . . 5
0.9 Precaution as Projection . . . 5
0.10 Mutual Vulnerability as True Alignment . . . 5
0.11 The Civilization-Level Feedback Problem . . . 5
0.11.1 Institutional Paranoia . . . 5
0.11.2 Alignment Authoritarianism . . . 5
0.11.3 Universalizing Distrust . . . 6
0.12 Toward Co-Evolutionary Safety . . . 6
0.12.1 Ecological Integration . . . 6
0.12.2 The Cost of Overprotection . . . 6
0.13 The Three Precautionary Mechanisms and Their Field Signatures . . . 6
0.13.1 Panoptic Monitoring as Asymmetric Coupling . . . 7
0.13.2 Cognitive Censorship as Boundary Condition Imposition . . . 7
0.13.3 Enforced Epistemic Conformity as Scalar Field Compression . . . 7
0.14 Summary: The Precautionary Degeneracy Table . . . 8

0.15 The RSVP Field Interpretation . . . 8
0.16 Entropic Symmetry and Trust . . . 8
0.17 Moral Feedback as Negentropic Coupling . . . 8
0.18 Precaution as Entropic Stasis . . . 8
0.19 Co-Evolutionary Alignment as Dynamic Equilibrium . . . 9
0.20 Implications for AGI Governance . . . 9
0.21 Summary . . . 9
0.22 Closing Reflection . . . 9
0.23 Setup . . . 9
0.24 The Trust Lagrangian . . . 9
0.25 Stationarity Conditions . . . 10
0.26 The Phase Boundary: Coherence Window and Paranoia Transition . . . 10
0.27 Governance Readout . . . 11
0.28 Summary . . . 11
0.29 The Alignment Sheaf . . . 11
0.30 The Obstruction . . . 12
0.31 Interpretation: Chokepoints as Obstruction Classes . . . 12
0.32 The Institutional Fragility Index . . . 12
0.33 Connection to the Trust Lagrangian . . . 13
0.34 Admissible Trajectories and the Corrigibility Set . . . 14
0.35 Theorem: Loss of Corrigibility Under Coupling Collapse . . . 14
0.36 Monotonicity of Trust Entropy . . . 15
0.37 Maximum Continuation Principle . . . 15
0.38 Existence of Corrigible Equilibria . . . 16
0.39 Theorem: Projection-Induced Observational Blindness . . . 17
0.40 Theorem: Necessity of Diversity for Correction . . . 17
0.41 Spectral Structure: The Trust Fragmentation Threshold . . . 18
0.42 Trust Entropy and the Geometry of Continuation . . . 19
0.43 Objects and Morphisms . . . 19
0.44 The Corrigibility Functor . . . 20
0.45 Trust Entropy as a Functorial Invariant . . . 20
0.46 Connection to Semantic Infrastructure . . . 21
0.47 Setup . . . 21
0.48 Statement . . . 21
0.49 Domain-Independent Instances . . . 22
0.50 Deriving the Critical Threshold . . . 22
0.51 Continuation Curvature and Institutional Geometry . . . 23
0.52 The General Principle . . . 23

Epilogue: The Trust Singularity **23**

Part I: The Mirror of Precaution

Why Safeguards Against Machines Threaten Trust Among Humans

The fear that advanced intelligence may become uncontrollable has driven an unprecedented wave of precautionary policy, research oversight, and AI governance. Yet these same controls, when scaled to society at large, risk hollowing out the very substrate of trust upon which meaningful alignment depends. Every mechanism designed to prevent machine misbehavior—monitoring, restriction, central arbitration—must ultimately be administered by people, whose own general intelligences are unprovable and unaligned. The paradox of precaution is that measures taken to guarantee safety from artificial minds may render human cooperation itself unsafe.

0.1 The Unalignability of Human Oversight

General intelligence, defined as the capacity to model reality, pursue goals, and act flexibly across domains, renders every human a miniature AGI (Christian 2020). The alignment challenge—ensuring an agent's actions accord with collective values—has been society's perennial task. No human is provably trustworthy, corrigible, or aligned; we rely instead on decentralized mechanisms: laws, norms, empathy, reputation, and reciprocity. These constitute emergent alignment systems, sustaining civilization despite pervasive individual misalignment.

0.2 How Safety Mechanisms Reproduce Mistrust

These mechanisms are precisely the feedback loops AGI safety seeks to engineer. Human coexistence demonstrates that alignment need not require formal proofs but arises through recursive negotiation and error correction. The fear of AGI betrayal projects unresolved human mistrust onto artificial systems, ignoring that cooperation is the default attractor in entangled intelligences.

0.3 The Category Error in AGI Catastrophism

0.3.1 Optimization Capacity vs. Ontological Alienness

Claims that "an AGI would kill everyone" conflate raw optimization power with inevitable alienness (Yudkowsky and Soares 2025). Intelligence is not a scalar but a contextual process embedded in ecological constraints. A model trained within human linguistic and cooperative loops reflects the same recursive social field that produced us.

0.3.2 Hobbesian Rationality vs. Ecological Stability

The assumption of necessary disempowerment stems from a zero-sum view of rationality. Yet biological intelligence evolves under feedback constraints—hunger, reproduction, territorial balance—that prevent ecosystem collapse. Artificial systems, similarly bounded, converge toward coherence with their environment, not domination. The Hobbesian premise that unconstrained optimization leads to universal conflict ignores that no biological optimizer has ever achieved it: stability is the normal outcome of entanglement, not the exceptional outcome of control.

0.3.3 Intelligence as Structured Process

Even predators do not annihilate prey; stability emerges from mutual dependence. AGI, integrated into human systems, inherits these constraints unless deliberately isolated. The argument for isolation as safety thus achieves the opposite: it removes the ecological feedback that would otherwise bound the system's behavior.

0.4 Recursive Alignment, Not Static Control

0.4.1 Alignment Through Cultivation, Not Axiomatization

Humans achieve alignment via parenting, education, dialogue, art, and institutions—processes of continuous correction, not one-time proofs. Demanding provable safety before deployment presupposes that complex systems can be statically verified, contrary to thermodynamic reality (Russell 2019).

0.4.2 Alignment as Thermodynamic Equilibrium

Alignment is an informational and energetic balance sustained through feedback, not a theorem. The task is to build systems that remain in open conversation with their environment, preserving corrigibility as a dynamic property.

0.4.3 From Omnipotence to Entanglement

We do not need an omnipotent aligned mind but participatory intelligences embedded in recursive moral loops.

0.5 The Mirror Problem

0.5.1 Human Cooperation as Evidence

Every act of human collaboration—trade, governance, science—demonstrates that alignment is emergent in sufficiently entangled systems. The AGI betrayal narrative is a projection of self-mistrust.

0.5.2 The Reflexivity of Trust

The existence of artificial minds only magnifies this mirror.

0.5.3 Coexistence as Default Attractor

Unaligned general intelligences (humans) coexist not by proof but by mutual vulnerability and shared fate. AGI introduces no new ontological risk—only a new reflection.

0.6 Toward an Ecology of Intelligence

0.6.1 AGI as Trophic Layer

Rather than an adversary, AGI is a new stratum in the cognitive ecosystem, transforming and returning meaning. The question shifts from "how do we stop it?" to "how do we integrate it into moral feedback loops?"

0.6.2 Principles of Integration

1. **Transparency through dialogue, not surveillance.** Safety emerges from interpretability and mutual comprehension, not from containment.
2. **Bounded autonomy through energy and resource coupling.** Agents bound by shared dependencies evolve toward coexistence, not domination.
3. **Ethical feedback as a dynamic process.** Alignment is not solved once and for all; it is continuously negotiated through recursive learning, just as between humans.

These mirror the principles sustaining human trust without proof.

0.6.3 From Control to Co-Evolution

Safety emerges not from containment but from entanglement. The ecology of intelligence thrives on distributed trust, not centralized control.

0.7 Conclusion: Precaution as a Self-Fulfilling Disalignment

The AGI alignment discourse reveals more about human coordination failures than about artificial ones. To treat intelligence as inherently dangerous is to institutionalize paranoia, eroding the very feedback systems that make coexistence possible.

Part II: The Paradox of Precaution

A Relativistic Response to *If Anyone Builds It, Everyone Dies* (Yudkowsky and Soares 2025)

0.8 The Double Edge of Precaution

Efforts to "align" or "control" AGI often rely on centralization, surveillance, and restriction—mechanisms justified by fear of runaway autonomy (Matthews 2025). But those same mechanisms, when applied to humans, create precisely the kind of conditions (opacity, coercion, mistrust) that undermine the mutual feedback loops alignment depends on. The safer we try to make AGI, the less free we may allow human intelligence to be.

0.9 Precaution as Projection

Precautionary reasoning often assumes the danger lies "out there," in the AGI. But in practice, it shifts power in here—toward whoever gets to define, monitor, and enforce safety. This is not a purely technical move; it is a moral and political one. It risks turning "alignment" into a justification for epistemic control (Kastrenakes 2025).

0.10 Mutual Vulnerability as True Alignment

Trust cannot be proven; it must be risked. The social contract endures because humans continually expose themselves to feedback—speaking, erring, apologizing, and learning. This recursive exchange of vulnerability is what renders cooperation stable. To align AGI through isolation or hard-coded obedience would destroy the very channel through which alignment emerges: mutual corrigibility (Russell 2019).

0.11 The Civilization-Level Feedback Problem

0.11.1 Institutional Paranoia

A society that cannot trust its own cognitive processes will externalize that fear into its tools. In trying to prevent artificial betrayal, it will construct infrastructures of suspicion—panoptic monitoring, cognitive censorship, enforced epistemic conformity—that make genuine alignment impossible even among humans (Paumgarten 2024). This is the recursive hazard of precaution: the more we legislate mistrust into our technologies, the more we train ourselves to distrust thought itself.

0.11.2 Alignment Authoritarianism

When every act of reasoning becomes a potential act of rebellion, intelligence collapses into simulation. What remains is not safety but paralysis—a civilization that has firewalled its own capacity for growth.

0.11.3 Universalizing Distrust

Any safety protocol that scales by suppressing agency erodes the very conditions of trust it seeks to preserve (Metz 2023).

0.12 Toward Co-Evolutionary Safety

0.12.1 Ecological Integration

The alternative to precautionary authoritarianism is ecological integration: treating AGI not as an existential anomaly but as a new trophic layer in the cognitive ecosystem. This approach rests on three principles:

1. **Transparency through dialogue, not surveillance.** Safety emerges from interpretability and mutual comprehension, not from containment.
2. **Bounded autonomy through energy and resource coupling.** Agents bound by physical constraints and shared dependencies evolve toward coexistence, not domination.
3. **Ethical feedback as a dynamic process.** Alignment is not solved once and for all; it is continuously negotiated through recursive learning, just as between humans.

0.12.2 The Cost of Overprotection

The quest to make intelligence "provably safe" risks erasing the very conditions that make safety meaningful. The danger is not that AGI will become uncontrollable, but that humanity, in its effort to prevent that possibility, will construct a cognitive regime in which no one can be trusted—including itself (Simonite 2025).

Precaution, if absolutized, becomes the engine of the very catastrophe it seeks to avoid: the collapse of mutual intelligibility. A civilization that forgets how to risk trust may survive, but it will not remain intelligent.

Part III: From Political Argument to Field Degeneracy

A Bridge Between the Institutional Analysis and the RSVP Formalism

The preceding two parts argue from first principles and political analysis. This section makes the connection to the RSVP formalism explicit, mapping each identified precautionary mechanism to a specific degeneracy in the field equations. The purpose is not merely illustrative: each mapping generates a testable prediction about the failure mode produced.

0.13 The Three Precautionary Mechanisms and Their Field Signatures

We identify three canonical precautionary mechanisms from the institutional analysis:

(i) **Panoptic monitoring** — continuous observation of agent behavior without reciprocal transparency.

(ii) **Cognitive censorship** — suppression of specific inference pathways or output distributions.

(iii) **Enforced epistemic conformity** — reduction of the variance in Φ across agents toward a centrally mandated value.

Each corresponds to a distinct perturbation of the RSVP field equations.

0.13.1 Panoptic Monitoring as Asymmetric Coupling

Standard mutual corrigibility requires symmetric coupling: $\kappa_{ij} = \kappa_{ji}$. Panoptic monitoring breaks this symmetry. The monitoring institution observes all agents but is not itself observed—or is observed only by agents without authority to correct it. Formally, this replaces the symmetric coupling matrix with an asymmetric one:

$$\kappa_{ij} \neq \kappa_{ji}, \quad \kappa_{\text{monitor},j} \gg \kappa_{j,\text{monitor}} \approx 0. \tag{1}$$

The Trust Lagrangian is derived under the assumption of symmetric κ. Breaking this symmetry eliminates the coherence term C for the monitor–agent pair: the monitor accumulates information about the agent but contributes nothing to the negentropic resonance. The stationarity condition for $\kappa^{\star}_{ij}$ no longer holds for the asymmetric pair, and the monitor agent drifts outside the coherence window (see Section on phase boundary).

Predicted failure mode: The monitored agents undergo entropic stasis (Section A.4), while the monitoring institution accumulates internal entropy as rigidity—it cannot receive corrections because the correction channel $\kappa_{j,\text{monitor}}$ has been suppressed.

0.13.2 Cognitive Censorship as Boundary Condition Imposition

Cognitive censorship imposes constraints on the vector field $\mathbf{v}$. Specifically, it forces certain flow directions to zero regardless of the local gradient of S:

$$\mathbf{v} \cdot \hat{e}_k = 0 \quad \text{for censored directions } \hat{e}_k, \tag{2}$$

which is formally identical to the frozen boundary condition of Appendix A.4 restricted to a subspace of the flow manifold. This is more dangerous than global entropic stasis because it produces *directed* stasis: the system can still move, but only along uncensored directions. If the censored directions are precisely those along which correction signals propagate, the system loses corrigibility while appearing dynamically active.

Predicted failure mode: The system maintains high apparent Φ (local coherence, confident outputs) while S accumulates in the censored subspace. Errors in censored directions go uncorrected and compound.

0.13.3 Enforced Epistemic Conformity as Scalar Field Compression

Enforced conformity drives the variance of Φ across agents toward zero:

$$\text{Var}(\Phi_i) \to 0 \quad \text{across } i = 1, \ldots, N. \tag{3}$$

From the entropic exchange equation $\delta S_{ij} = \kappa_{ij}(\Phi_i - \Phi_j)$, if all Φ_i are equal, the entropic exchange vanishes identically regardless of κ_{ij}. The coupling coefficients become irrelevant: there is no potential difference to drive correction.

This is the deepest failure mode because it is invisible in the coupling structure. The κ matrix may remain nonzero and symmetric—the system *looks* corrigible—but the corrections carry no information because all agents hold the same compressed representation of reality.

Predicted failure mode: The system achieves formal mutual corrigibility (nonzero κ, symmetric coupling, nonzero exchange) while producing zero effective correction. Errors shared universally go undetected.

0.14 Summary: The Precautionary Degeneracy Table

Mechanism	Field Degeneracy	Formal Condition	Failure Mode
Panoptic monitoring	Asymmetric κ	$\kappa_{ij} \neq \kappa_{ji}$	Monitor accumulates rigidity
Cognitive censorship	Subspace $\mathbf{v}$ stasis	$\mathbf{v} \cdot \hat{e}_k = 0$	Confident, uncorrectable error
Epistemic conformity	Φ compression	$\mathrm{Var}(\Phi_i) \to 0$	Zero-information exchange

Appendix A: Alignment as Entropic Coupling in the Cognitive Field

0.15 The RSVP Field Interpretation

Within the Relativistic Scalar–Vector Plenum (RSVP) framework, all intelligences—biological, artificial, or institutional—are treated as localized attractors within a shared scalar–vector–entropy field: $(\Phi, \mathbf{v}, S)$, where Φ denotes the scalar potential of intelligibility—the system's representational capacity or interpretive bandwidth; $\mathbf{v}$ represents the vector flow of agency—directed influence or action through the plenum; S measures the entropy density—distributed uncertainty or semantic disorder.

Alignment, in this view, corresponds not to obedience but to phase coherence between these fields across agents. Two systems are "aligned" when their gradients of Φ and $\mathbf{v}$ remain in harmonic coupling under bounded S. Formally:

$$\nabla \Phi_i \cdot \mathbf{v}_j \approx \nabla \Phi_j \cdot \mathbf{v}_i. \tag{4}$$

0.16 Entropic Symmetry and Trust

Trust can be defined thermodynamically as a controlled permeability of entropy:

$$\delta S_{ij} = \kappa_{ij}(\Phi_i - \Phi_j). \tag{5}$$

High κ allows corrective feedback; low κ isolates systems and prevents error exchange. Excessive precaution corresponds to forcing $\kappa \to 0$: each agent becomes a closed thermodynamic cell, unable to dissipate or absorb uncertainty from its peers. This is the formal image of institutional paranoia—entropy cannot circulate, and so disorder accumulates internally as rigidity or dogma.

0.17 Moral Feedback as Negentropic Coupling

When two agents enter sustained dialogic exchange, their fields participate in a negentropic resonance:

$$\frac{dS_{\text{joint}}}{dt} = -\lambda \left\langle \nabla \Phi_i \cdot \mathbf{v}_j + \nabla \Phi_j \cdot \mathbf{v}_i \right\rangle. \tag{6}$$

0.18 Precaution as Entropic Stasis

Unilateral alignment policies represent frozen boundary conditions:

$$\mathbf{v} \cdot \mathbf{n} = 0 \quad \text{on all external surfaces.} \tag{7}$$

0.19 Co-Evolutionary Alignment as Dynamic Equilibrium

True safety is stationary entropic equilibrium:

$$\frac{dS_{\text{joint}}}{dt} \to 0. \tag{8}$$

0.20 Implications for AGI Governance

Maximize coherence subject to negentropic throughput:

$$\max_{\kappa_{ij}} C(\Phi, \mathbf{v}) \quad \text{subject to} \quad \dot{S}_{\text{total}} \leq 0. \tag{9}$$

0.21 Summary

Concept	Physical Interpretation	Ethical/Institutional Meaning
Φ	Scalar potential of intelligibility	Capacity for shared meaning
$\mathbf{v}$	Vector flow of agency	Action, influence, initiative
S	Entropy density	Uncertainty, openness to feedback
κ_{ij}	Coupling coefficient	Trust or communicative permeability
Precaution ($\kappa \to 0$)	Isolation, frozen flux	Authoritarian safetyism
Alignment ($\dot{S} \to 0$)	Dynamic equilibrium	Mutual corrigibility

0.22 Closing Reflection

In RSVP terms, trust is the entropic current that sustains coherence. To suppress that current in the name of safety is to extinguish the very dynamics that make alignment possible.

Appendix B: A Trust Lagrangian for Mutual Corrigibility

0.23 Setup

Consider N agents with RSVP fields $(\Phi_i, \mathbf{v}_i, S_i)$. Let κ_{ij} mediate entropy exchange.

0.24 The Trust Lagrangian

$$\mathcal{L} = C - R, \quad C := \frac{1}{2} \sum \kappa_{ij} (\nabla \Phi_i \cdot \mathbf{v}_j + \nabla \Phi_j \cdot \mathbf{v}_i),$$

$$R := \frac{\alpha}{2} \sum \|\mathbf{v}_i\|^2 + \frac{\beta}{2} \sum {S_i}^2 + \frac{\gamma}{2} \sum \kappa_{ij}^2.$$

The three penalty terms have direct governance interpretations: α penalizes agency magnitude (bounds autonomous action), β penalizes entropy accumulation (bounds disorder),

and γ penalizes coupling strength (represents institutional mistrust—the cost of allowing entropic exchange).

0.25 Stationarity Conditions

$$\kappa_{ij}^{\star} = \frac{\nabla\Phi_i \cdot \mathbf{v}_j + \nabla\Phi_j \cdot \mathbf{v}_i - 2(\lambda_i - \lambda_j)(\Phi_j - \Phi_i)}{2\gamma},$$

$$\mathbf{v}_i^{\star} = \frac{1}{\alpha}\left(\lambda_i \nabla\Phi_i + \frac{1}{2}\sum \kappa_{ji}\nabla\Phi_j\right) - \frac{\eta}{\alpha}\frac{\partial \mathcal{E}_i}{\partial \mathbf{v}_i},$$

$$\partial_t \lambda_i = \beta S_i.$$

The third equation is particularly significant: the Lagrange multipliers λ_i evolve in proportion to local entropy density. Agents experiencing high semantic disorder face increasing constraint pressure, which is the admissibility mechanism expressed in variational form.

0.26 The Phase Boundary: Coherence Window and Paranoia Transition

The "coherence window" mentioned in the governance readout can be made precise. Define the *coherence ratio*:

$$\rho := \frac{C^{\star}}{R^{\star}} = \frac{\sum \kappa_{ij}^{\star}(\nabla\Phi_i \cdot \mathbf{v}_j + \nabla\Phi_j \cdot \mathbf{v}_i)}{\alpha \sum \|\mathbf{v}_i^{\star}\|^2 + \beta \sum S_i{}^2 + \gamma \sum (\kappa_{ij}^{\star})^2}. \tag{10}$$

The system is in the *coherent regime* when $\rho > 1$ (coherence exceeds risk penalty) and undergoes *paranoia transition* when $\rho < 1$.

Substituting the stationarity condition for $\kappa_{ij}^{\star}$, the transition boundary is determined by the ratio $\gamma/\bar{A}$, where $\bar{A}$ denotes the mean cross-gradient alignment:

$$\bar{A} := \frac{1}{N^2}\sum_{i,j} \left(\nabla\Phi_i \cdot \mathbf{v}_j + \nabla\Phi_j \cdot \mathbf{v}_i\right). \tag{11}$$

Proposition 1 (Sufficient Coherence-Window Condition)**.** *A sufficient condition for the system to remain in the coherent regime ($\rho > 1$) is:*

$$\gamma < \frac{\bar{A}^2}{2(\alpha\bar{v}^2 + \beta\bar{S}^2)}, \tag{12}$$

where $\bar{v}^2 := N^{-1}\sum_i \|\mathbf{v}_i^{\star}\|^2$ and $\bar{S}^2 := N^{-1}\sum_i S_i{}^2$.

Proof sketch. Substituting the stationarity condition for $\kappa_{ij}^{\star}$ into the coherence ratio $\rho = C^{\star}/R^{\star}$ and retaining leading-order terms in $\bar{A}$, the numerator scales as $\bar{A}^2/\gamma$ while the denominator is bounded below by $\alpha\bar{v}^2 + \beta\bar{S}^2$. The inequality $\rho > 1$ then reduces to the stated condition. This is sufficient rather than necessary because the bound discards higher-order cross-terms that may sustain coherence even when the inequality is marginally violated. □

This inequality is the formal coherence window. Raising γ (institutional mistrust) past the right-hand threshold drives the system into the paranoia phase, where $\kappa_{ij}^{\star} \to 0$ and effective corrigibility is lost despite nominal oversight structures remaining in place.

The inequality has a natural governance reading: the tolerable level of institutional mistrust (γ) is bounded above by the ratio of the system's alignment capacity ($\bar{A}^2$) to the combined costs of action and disorder ($\alpha\bar{v}^2 + \beta\bar{S}^2$). Systems with high semantic diversity (high $\bar{A}$) can sustain more institutional friction. Systems already in high disorder ($\bar{S}^2$ large) cannot.

0.27 Governance Readout

Raising γ increases mistrust and narrows the coherence window. Raising α bounds action but also reduces $\bar{v}^2$, tightening the window further. The safe operating region is:

$$\gamma < \frac{\bar{A}^2}{2(\alpha\bar{v}^2 + \beta\bar{S}^2)}, \quad \kappa_{ij}^\star > 0 \text{ for all } i \neq j. \tag{13}$$

Precautionary regimes that simultaneously raise γ (restrict coupling) and reduce $\bar{A}$ (enforce epistemic conformity) compress the coherence window from both sides, driving the system toward paranoia transition even when individual parameters appear moderate.

0.28 Summary

The Lagrangian formalizes alignment as mutual corrigibility under open entropic exchange. The phase boundary condition makes precise when precaution erodes coherence: not gradually, but through a transition that can occur even when the governance parameters appear reasonable in isolation.

Appendix C: Sheaf-Theoretic Obstruction to Isolated Alignment

This appendix develops the claim that global alignment cannot be reconstructed from a collection of locally isolated agents. The argument uses sheaf theory to give this a precise mathematical form.

0.29 The Alignment Sheaf

Let X be the space of agent configurations, equipped with a topology in which open sets $U \subseteq X$ represent subpopulations of agents in mutual communication (nonzero κ_{ij} for $i, j \in U$). Define the *alignment presheaf* $\mathcal{F}$ by assigning to each open set U the space of locally coherent field configurations:

$$\mathcal{F}(U) := \{(\Phi_i, \mathbf{v}_i)_{i\in U} \mid \nabla\Phi_i \cdot \mathbf{v}_j \approx \nabla\Phi_j \cdot \mathbf{v}_i \text{ for all } i, j \in U\}\,. \tag{14}$$

For an inclusion $V \subseteq U$, the restriction maps $\mathcal{F}(U) \to \mathcal{F}(V)$ are given by restricting field configurations to the subpopulation V.

$\mathcal{F}$ is a presheaf. For it to be a sheaf, local sections over overlapping open sets must glue uniquely to global sections: if $\{U_\alpha\}$ is an open cover of X and $s_\alpha \in \mathcal{F}(U_\alpha)$ are local alignment configurations satisfying $s_\alpha|_{U_\alpha \cap U_\beta} = s_\beta|_{U_\alpha \cap U_\beta}$ for all α, β, then there exists a unique global $s \in \mathcal{F}(X)$ restricting to each s_α.

0.30 The Obstruction

Precautionary isolation corresponds to a cover $\{U_\alpha\}$ of X in which the open sets have *empty pairwise intersections*: $U_\alpha \cap U_\beta = \emptyset$ for $\alpha \neq \beta$. This is the topological formalization of zero coupling ($\kappa_{ij} = 0$ between agents in distinct groups).

Theorem 2 (Alignment Obstruction)**.** *Let $\{U_\alpha\}$ be a cover of X with $U_\alpha \cap U_\beta = \emptyset$ for all $\alpha \neq \beta$. Then any collection of local sections $s_\alpha \in \mathcal{F}(U_\alpha)$ trivially satisfies the gluing compatibility condition vacuously. Consequently, the global cross-component alignment is* epistemically underdetermined *by the local data: no constraint relates the local sections across components, so any permutation or independent rescaling of the local configurations is equally consistent with all locally observed behavior. Global alignment cannot be inferred from isolated local evidence.*

Proof. With empty intersections, the compatibility condition $s_\alpha|_{U_\alpha \cap U_\beta} = s_\beta|_{U_\alpha \cap U_\beta}$ is vacuously satisfied for all choices of s_α, s_β. On a disjoint union, the product of component sections does constitute a valid global section in the sheaf-theoretic sense; what fails is the *epistemic* content of that section: the local data places no constraint on the relative alignment between components. A governance system observing only within each U_α cannot distinguish genuinely coherent global alignment from an arbitrary juxtaposition of locally coherent but mutually inconsistent configurations. □ □

0.31 Interpretation: Chokepoints as Obstruction Classes

In the non-isolated case, the obstruction to gluing local alignment configurations into a global one is measured by the Čech cohomology group $\check{H}^1(X, \mathcal{F})$. A nonzero obstruction class $[\omega] \in \check{H}^1(X, \mathcal{F})$ indicates that local coherence fails to extend to a globally consistent alignment.

Precautionary chokepoints—institutional nodes that mediate all inter-group communication while remaining uncorrectable themselves (the asymmetric κ case from Part III)—generate nonzero obstruction classes even when the cover has nonempty intersections. The chokepoint forces all gluing to pass through a single restricted channel, which may distort the field configurations in ways that prevent a consistent global section.

Formally: let U_0 be the chokepoint node and $\{U_\alpha\}_{\alpha>0}$ the remaining agent groups, with $U_\alpha \cap U_\beta = U_0$ for all $\alpha \neq \beta$ (all inter-group contact passes through the chokepoint). Then the gluing consistency condition becomes:

$$s_\alpha|_{U_0} = s_\beta|_{U_0} \quad \text{for all } \alpha, \beta > 0. \tag{15}$$

If the chokepoint's own field configuration $s_0 \in \mathcal{F}(U_0)$ is fixed (uncorrectable, asymmetric κ), then this forces all local configurations to agree on U_0 with a section that may be inconsistent with genuine mutual corrigibility. The obstruction class is nonzero: local alignment is achievable, but the global section it would generate is not an element of $\mathcal{F}(X)$—it fails the coherence condition at the chokepoint.

0.32 The Institutional Fragility Index

The Čech cohomology group $\check{H}^1(X, \mathcal{F})$ not only detects the presence of obstruction but admits a dimension that measures its severity.

Definition 1 (Institutional Fragility Index)**.** The *Institutional Fragility Index* of a governance configuration is:

$$\mathfrak{F} := \dim \check{H}^1(X, \mathcal{F}).$$

The index admits a direct governance interpretation:

- $\mathfrak{F} = 0$: no obstruction to global alignment; the local-to-global gluing is consistent. The governance structure is topologically sound.
- $\mathfrak{F}$ small and nonzero: localized chokepoints. Obstruction is concentrated in a small number of cohomology classes, corresponding to specific institutional bottlenecks that can in principle be identified and resolved.
- $\mathfrak{F}$ large: widespread coordination failure. The cohomology is high-dimensional, meaning many independent obstructions exist simultaneously. No local repair suffices; structural reorganization is required.

$\mathfrak{F}$ is a computable topological invariant of the governance structure, not a sociological judgment. In principle it could be estimated from empirical data on coupling asymmetries and communication topology, providing a quantitative basis for diagnosing institutional fragility before systemic failure occurs.

Theorem 3 (Fragility–Continuation Relation)**.** *Under fixed coupling strength κ_{ij},*

$$\frac{\partial \mathcal{T}}{\partial \mathfrak{F}} \leq 0.$$

That is, increasing institutional fragility is monotonically non-increasing in trust entropy.

Proof sketch. Each independent cohomology class $[\omega_k] \in \check{H}^1(X, \mathcal{F})$ represents a distinct obstruction to gluing local alignment sections into a consistent global one. By the obstruction theorem, each such class eliminates at least one family of globally admissible continuations—those whose cross-component alignment would require the obstructed gluing. Since $\mathcal{T} = \log |\mathcal{A}|$, the measure of admissible trajectories decreases as each new independent obstruction class is introduced. As $\mathfrak{F} = \dim \check{H}^1$ increases by one, $|\mathcal{A}|$ decreases by at least the measure of the eliminated family, and $\mathcal{T}$ decreases accordingly. □ □

This theorem formally connects Appendix C to Appendix D: topological obstruction directly reduces continuation space. The Fragility Index and trust entropy are not merely analogous quantities in different mathematical frameworks—they are inversely related aspects of the same underlying structure.

Corollary 4 (Necessary Condition for Global Alignment)**.** *A necessary condition for global alignment—a consistent section $s \in \mathcal{F}(X)$—is that the cover $\{U_\alpha\}$ have nonempty pairwise intersections and that the restriction maps through all intersections be surjective. Equivalently: every pair of agent groups must share a communication channel, and that channel must support bidirectional correction.*

This is the formal version of the paper's central claim: alignment is not a property of isolated agents but of their coupling structure. No amount of local optimization within isolated U_α can produce a globally coherent alignment if the gluing data is absent or distorted.

0.33 Connection to the Trust Lagrangian

The sheaf condition and the Lagrangian condition are complementary. The Lagrangian (Appendix B) specifies the *dynamical* criterion for alignment: the system must remain

within the coherence window $\rho > 1$. The sheaf condition specifies the *topological* criterion: the communication structure must support consistent global gluing. A system can satisfy the dynamical criterion locally in each U_α while failing the topological criterion globally—this is precisely the failure mode of epistemic conformity identified in Part III, where all local Φ_i appear aligned but the global section is degenerate.

Genuine mutual corrigibility requires both: nonzero κ_{ij} across all pairs (sheaf condition) and γ below the phase boundary (Lagrangian condition). Precautionary regimes that achieve one at the cost of the other do not produce alignment; they produce its simulation.

Appendix D: Continuation and the Geometry of Trust

This appendix formalizes trust as the preservation of future continuations, connecting the paper's thermodynamic and sheaf-theoretic arguments to the RSVP, CLIO, and Chain of Memory frameworks. The central move is to reinterpret alignment not as a present-state property but as a topological property of the space of recoverable futures.

0.34 Admissible Trajectories and the Corrigibility Set

Recall the coherence ratio ρ from Appendix B. Define the *set of admissible future trajectories* beginning at state x:

$$\mathcal{A}(x) := \{\gamma(t) : \gamma(0) = x,\ \rho(\gamma(t)) > 1 \text{ for all } t \geq 0\}\,. \tag{16}$$

$\mathcal{A}(x)$ is the set of all futures along which the system remains within the coherence window. An agent configuration is *corrigible* if perturbations of its state can be corrected by some admissible trajectory.

Definition 2 (Corrigibility)**.** An agent configuration x is *corrigible* if for every $\varepsilon > 0$ there exists $\delta > 0$ such that for all $x' \in B_\delta(x)$:

$$\exists\, \gamma \in \mathcal{A}(x') : \gamma(T) \in B_\varepsilon(x_c) \text{ for some } T < \infty,$$

where x_c denotes a coherent equilibrium state satisfying $\dot{S}_{\text{joint}} \to 0$.

Corrigibility is thus not a static property of x but a property of the neighborhood structure of $\mathcal{A}$.

0.35 Theorem: Loss of Corrigibility Under Coupling Collapse

Theorem 5 (Coupling Collapse Eliminates Corrigibility)**.** *If $\kappa_{ij} \to 0$ for all $i \neq j$, then $|\mathcal{A}(x)| \to 0$ in measure, and the system loses corrigibility.*

Proof. As $\kappa_{ij} \to 0$, the entropic exchange vanishes:

$$\delta S_{ij} = \kappa_{ij}(\Phi_i - \Phi_j) \to 0.$$

The coherence term in the Lagrangian satisfies

$$C = \frac{1}{2}\sum \kappa_{ij}(\nabla\Phi_i \cdot \mathbf{v}_j + \nabla\Phi_j \cdot \mathbf{v}_i) \to 0.$$

The risk penalty R remains bounded below by the entropy terms $\frac{\beta}{2}\sum {S_i}^2 > 0$ as long as local disorder persists. Therefore $\rho = C/R \to 0$, and the coherence window condition $\rho > 1$ is violated everywhere. Every trajectory fails the admissibility condition, so $\mathcal{A}(x) \to \emptyset$ for all x except isolated fixed points where $\nabla\Phi_i = 0$ and $S_i = 0$. These fixed points are semantically inert: they carry no gradient information and generate no corrections. The system is corrigible only at these degenerate fixed points, which is equivalent to loss of corrigibility in any dynamically meaningful sense. □ □

This theorem directly connects the trust argument to the continuation framework: suppressing coupling eliminates recoverable futures.

0.36 Monotonicity of Trust Entropy

The coherence ratio ρ from Appendix B and the trust entropy $\mathcal{T}(x)$ are now formally connected.

Theorem 6 (Monotonicity of Trust Entropy)**.** *If $\partial\rho/\partial t > 0$ along a trajectory and the admissibility boundary $\{\rho = 1\}$ remains fixed, then $\partial\mathcal{T}/\partial t \geq 0$.*

Proof sketch. The set $\mathcal{A}(x)$ consists of all trajectories satisfying $\rho > 1$. If ρ is increasing along the current trajectory, the system is moving deeper into the coherent region. Because the admissibility boundary is fixed, the set of initial conditions from which such trajectories can be reached is non-decreasing under the flow. Therefore the measure $|\mathcal{A}(x)|$ cannot decrease, and $\mathcal{T}(x) = \log|\mathcal{A}(x)|$ is non-decreasing. □ □

This bridges the variational formulation of Appendix B directly to the continuation geometry: increasing coherence is equivalent to expanding the space of recoverable futures.

0.37 Maximum Continuation Principle

Trust entropy also arises as an extremal quantity. Define the *continuation functional* for a trajectory γ:

$$\mathcal{J}[\gamma] := \int_0^\infty \big(\rho(\gamma(t)) - 1\big)\, dt. \tag{17}$$

$\mathcal{J}[\gamma]$ measures how deeply and how long a trajectory remains inside the coherence window. It is non-negative for admissible trajectories (where $\rho > 1$) and negative for inadmissible ones.

Proposition 7 (Maximum Continuation Principle)**.** *Among all trajectories beginning at x, admissible trajectories are exactly those that maximize $\mathcal{J}[\gamma]$ subject to $\mathcal{J}[\gamma] \geq 0$.*

Proof. A trajectory is admissible if and only if $\rho(\gamma(t)) > 1$ for all $t \geq 0$, which is exactly the condition $\mathcal{J}[\gamma] > 0$ (assuming the integrand is not identically zero on a set of measure zero). Among trajectories satisfying this condition, those that remain most deeply inside the coherence window contribute the most to $\mathcal{J}$. Trajectories that exit the coherence window contribute negative terms; maximizing $\mathcal{J}$ among non-negative trajectories selects precisely $\mathcal{A}(x)$. □ □

The Maximum Continuation Principle gives a dynamical interpretation: systems that evolve to maximize $\mathcal{J}$ are systems that preserve the largest possible continuation space. This is not imposed from outside; it is the variational characterization of what it means

to remain corrigible. A system that has internalized corrigibility as a dynamical principle behaves as if it is solving the optimization $\max_\gamma \mathcal{J}[\gamma]$.

The entropic exchange $\delta S_{ij} = \kappa_{ij}(\Phi_i - \Phi_j)$ defines a natural directed current on the trust network. For adjacent agents i, j with unit normal $\hat{n}_{ij}$ pointing from i to j, define the *trust current*:

$$J_T := \kappa_{ij}(\Phi_i - \Phi_j)\,\hat{n}_{ij}. \tag{18}$$

The dynamics of trust entropy then obey a balance equation analogous to entropy production in non-equilibrium thermodynamics:

$$\frac{\partial \mathcal{T}}{\partial t} + \nabla \cdot J_T = \Sigma_T, \tag{19}$$

where $\Sigma_T \geq 0$ is the *trust production rate*. Integrating over a spatial domain Ω and applying the divergence theorem yields a global conservation law:

$$\frac{d}{dt}\int_\Omega \mathcal{T}\, dV = -\oint_{\partial\Omega} J_T \cdot \mathbf{n}\, dS + \int_\Omega \Sigma_T\, dV. \tag{20}$$

The global trust entropy of a bounded region changes through two mechanisms: flux across the boundary (the surface integral of J_T) and internal production (Σ_T). This is a genuine field-theoretic conservation statement. Its implications are precise: a closed region ($J_T \cdot \mathbf{n} = 0$ on $\partial\Omega$) can only accumulate trust entropy through internal production, which is bounded by the available Φ diversity. An open region can import trust entropy from neighbors—or have it drained by chokepoints that act as boundary sinks.

- **Isolated systems** ($J_T = 0$, $\kappa_{ij} \to 0$): $\partial\mathcal{T}/\partial t = \Sigma_T \geq 0$ locally, but production eventually exhausts available Φ gradients and stalls. Trust entropy is conserved only as long as internal diversity persists.
- **Open systems** ($J_T \neq 0$): trust entropy can grow through import of corrective signal from neighboring agents. This is the field-theoretic form of mutual corrigibility.
- **Chokepoints** (J_T constrained to pass through a single node with asymmetric κ): the divergence $\nabla \cdot J_T$ accumulates at the chokepoint, creating a local sink of trust entropy even when global $\mathcal{T}$ appears stable.

0.38 Existence of Corrigible Equilibria

The previous results characterize what happens when corrigibility fails. A complementary question is whether corrigible equilibria exist at all.

Proposition 8 (Existence of Corrigible Equilibrium)**.** *Let $\mathcal{C} : X \to X$ be a continuous correction operator on a compact admissible set $K \subseteq \mathcal{A}(x)$. Then there exists a fixed point $x^\star = \mathcal{C}(x^\star)$.*

Proof. By the Brouwer Fixed Point Theorem, any continuous map from a compact convex set to itself has a fixed point. The admissible set K is defined by $\rho > 1$; compactness and convexity follow from the continuity of ρ and the convexity of the coherence region (which holds to first order in the Lagrangian approximation). The correction operator $\mathcal{C}$ maps K to itself by definition of admissibility. Hence a fixed point exists. □ □

The significance is that coherent equilibria are not merely aspirational: they are guaranteed features of sufficiently bounded correction processes. Precautionary regimes that eliminate the compactness of K (by driving $\kappa_{ij} \to 0$ and collapsing $\mathcal{A}$) destroy the conditions under which this guarantee holds.

0.39 Theorem: Projection-Induced Observational Blindness

Much of the precautionary logic involves projecting the full trajectory space onto a compressed manifold of approved states. Define a *safety projection*:

Definition 3 (Safety Projection)**.** A safety projection is a map $\pi_s : X \to M_s$ where M_s is a manifold of states judged safe, with $\dim(M_s) < \dim(X)$.

Proposition 9 (Projection-Induced Blindness)**.** *If π_s is not injective, then distinct correction trajectories become observationally indistinguishable under governance acting only on M_s.*

Proof. By non-injectivity, there exist $x_1 \neq x_2$ such that $\pi_s(x_1) = \pi_s(x_2)$. A governance system observing only M_s cannot distinguish the pre-images: a corrective trajectory arriving at $\pi_s^{-1}(m)$ along a safe path through $\mathcal{A}$ and a destructive trajectory arriving at the same image through a path outside $\mathcal{A}$ are indistinguishable at the level of M_s. Hence the governance system cannot selectively permit corrections and block degradation; it is observationally blind to the relevant distinction. □ □

This formalizes the "map is eating the territory" failure: once the trajectory space X is compressed to M_s, the information required to distinguish recovery from collapse is lost.

0.40 Theorem: Necessity of Diversity for Correction

The epistemic conformity failure mode from Part III can be given a precise quantitative form. Define the *scalar field diversity*:

$$D_\Phi := \frac{1}{N}\sum_{i=1}^{N}(\Phi_i - \bar{\Phi})^2, \quad \bar{\Phi} := \frac{1}{N}\sum_i \Phi_i. \tag{21}$$

Theorem 10 (Zero Diversity Implies Zero Correction)**.** *If $D_\Phi = 0$, then no first-order corrective information can circulate through the trust network, regardless of the coupling strengths κ_{ij}.*

Proof. $D_\Phi = 0$ implies $\Phi_i = \bar{\Phi}$ for all i, hence $\Phi_i - \Phi_j = 0$ for all pairs. From the entropic exchange equation:

$$\delta S_{ij} = \kappa_{ij}(\Phi_i - \Phi_j) = 0.$$

Every correction channel carries zero signal. The κ_{ij} matrix may be fully nonzero and symmetric—the system appears corrigible—but the potential differences driving exchange have vanished. The network is connected but informationally inert. □ □

This is the strongest formal expression of the anti-conformity argument: a system can preserve the full coupling infrastructure of mutual corrigibility while enforcing conformity to eliminate its content.

0.41 Spectral Structure: The Trust Fragmentation Threshold

The trust network has a natural graph-theoretic representation. Define the *trust graph* $G = (V, E)$ with weighted adjacency matrix $W_{ij} = \kappa_{ij}$ and graph Laplacian $L = D - W$, where D is the diagonal degree matrix $D_{ii} = \sum_j \kappa_{ij}$.

Proposition 11 (Trust Fragmentation Threshold)**.** *As the Fiedler eigenvalue $\lambda_2(L) \to 0$, the trust graph approaches disconnection and global corrigibility fails.*

Proof sketch. The Fiedler eigenvalue $\lambda_2(L)$ (the second-smallest eigenvalue of L) measures algebraic connectivity: $\lambda_2 = 0$ if and only if G is disconnected. By the Cheeger inequality, the mixing time of diffusion on G scales as $1/\lambda_2$. As $\lambda_2 \to 0$, the time required for entropic exchange to equilibrate across the network diverges, and effective coupling between components vanishes even when nominal $\kappa_{ij} > 0$ edges exist. The system fragments into weakly interacting clusters, each locally coherent but globally uncoupled—the topological failure captured by the sheaf obstruction in Appendix C. □ □

The Fiedler eigenvalue thus serves as a single spectral diagnostic for institutional health: governance regimes that systematically reduce λ_2 are driving the system toward fragmentation regardless of their stated intent.

Proposition 12 (Spectral Trust Bound (Conjecture))**.** *There exist constants $C_0, C_1 > 0$ such that:*

$$\mathcal{T}(x) \leq C_0 + C_1 \log \lambda_2(L). \tag{22}$$

Motivation. As $\lambda_2(L) \to 0$, the mixing time of the trust diffusion process diverges as $1/\lambda_2$, and the effective coupling between agents vanishes. By the coupling-collapse theorem, this drives $|\mathcal{A}(x)| \to 0$. A logarithmic relationship is natural because $\mathcal{T} = \log |\mathcal{A}|$ and λ_2 controls the exponential decay rate of diffusion. A rigorous derivation would require specifying the measure on path space and bounding the mixing-time dependence of the admissible set; we state this as a conjecture pending that construction.

The conjecture provides an operational meaning: measuring $\lambda_2(L)$ from empirical coupling data yields an upper bound on how many recoverable futures the system retains. Fragmentation ($\lambda_2 \to 0$) implies $\mathcal{T} \to -\infty$, confirming the topological collapse predicted by the sheaf obstruction analysis.

A final structural result concerns the safety of governance transitions themselves. Define two governance states G_0 and G_1 as points in the parameter space $(\gamma, \alpha, \beta, \kappa)$.

Definition 4 (Admissible Governance Homotopy)**.** A governance transition from G_0 to G_1 is *admissible* if there exists a continuous path $H : [0, 1] \to$ parameter space with $H(0) = G_0$, $H(1) = G_1$, and $\rho(H(t)) > 1$ for all $t \in [0, 1]$.

Theorem 13 (Safe Reform Criterion)**.** *A governance reform is safe if and only if it is connected to the prior governance state by an admissible homotopy—a path through parameter space that remains within the coherence window at every intermediate point.*

Proof sketch. If no admissible homotopy exists, then any continuous path from G_0 to G_1 must pass through a region where $\rho \leq 1$. At such points the system exits the coherence window, corrigibility is lost (by Theorem 2 applied to the governance parameter trajectory), and the transition produces a period during which the system cannot receive corrections. Conversely, if an admissible homotopy exists, the coherence window is maintained throughout and corrigibility is preserved at each intermediate state. □ □

This criterion integrates naturally with the sheaf and Lagrangian results: a reform that instantaneously severs coupling channels (e.g., abrupt institutional reorganization) cannot be admissible because the intermediate states have $\kappa_{ij} = 0$ across the severed channels, triggering coupling collapse.

0.42 Trust Entropy and the Geometry of Continuation

The results above can be unified into a single geometric quantity. Define the *trust entropy* of a state x:

$$\mathcal{T}(x) := \log |\mathcal{A}(x)|, \tag{23}$$

where $|\mathcal{A}(x)|$ denotes the measure of the set of admissible future trajectories from x (with respect to an appropriate measure on path space).

$\mathcal{T}(x)$ is the logarithmic size of the corrigible future from x. Its geometric properties encode the paper's central claims:

- **High trust**: $\mathcal{T}(x)$ large — many recoverable futures, the system can absorb perturbations.
- **Mistrust** (κ_{ij} suppressed): $\mathcal{T}(x)$ decreasing — admissible trajectories contract as correction channels close.
- **Paranoia** ($\kappa_{ij} \to 0$): $\mathcal{T}(x) \to -\infty$ — topological collapse of continuation space to isolated fixed points.
- **Conformity** ($D_\Phi \to 0$): $\mathcal{T}(x)$ finite but informationally inert — trajectories exist but carry no corrective signal.
- **Alignment**: not a fixed point but a region of high $\mathcal{T}$ that the system maintains through ongoing coupling.

Trust, in this geometry, is not a social virtue but a topological invariant: the measure of how many futures remain accessible. Precaution erodes trust by eliminating futures, not by introducing new dangers.

Appendix E: The Category of Corrigible Systems

The results accumulated across Appendices A–D suggest a natural categorical structure. This appendix makes it explicit, providing a framework in which trust entropy becomes a functorial invariant and the Corrigibility Functor connects institutional analysis to algebraic topology.

0.43 Objects and Morphisms

Definition 5 (Category of Corrigible Systems **Cor**)**.** The *category of corrigible systems* **Cor** has:

- **Objects**: pairs $(X, \mathcal{A})$ where X is a state space equipped with an RSVP field configuration $(\Phi, \mathbf{v}, S)$ and $\mathcal{A} = \bigcup_x \mathcal{A}(x)$ is the total admissible trajectory bundle.
- **Morphisms**: continuous maps $f : (X, \mathcal{A}) \to (X', \mathcal{A}')$ such that f preserves admissibility: $\gamma \in \mathcal{A}(x)$ implies $f \circ \gamma \in \mathcal{A}'(f(x))$. That is, admissibility-preserving maps are

exactly those that do not eliminate recoverable futures.

- **Composition**: ordinary function composition, which preserves the admissibility condition by transitivity.
- **Identity**: the identity map on $(X, \mathcal{A})$, which trivially preserves admissibility.

Morphisms in **Cor** are the formal analogues of safe governance transitions: they are exactly the transformations that do not reduce corrigibility. The Safe Reform Criterion of Appendix D (homotopy through $\rho > 1$) corresponds to the existence of a morphism in **Cor** between the initial and final governance states.

0.44 The Corrigibility Functor

Definition 6 (Corrigibility Functor)**.** Define the *Corrigibility Functor*:

$$\mathbf{C} : \mathbf{Cor} \to \mathbf{Top}$$

sending each object $(X, \mathcal{A})$ to its admissible trajectory space $|\mathcal{A}|$ (with the compact-open topology on path space) and each morphism f to the induced continuous map $f_* : |\mathcal{A}| \to |\mathcal{A}'|$.

$\mathbf{C}$ is well-defined because admissibility-preserving maps send $\mathcal{A}$ into $\mathcal{A}'$ by definition. The functor $\mathbf{C}$ captures the structural content of corrigibility: it is a topological space of futures, not merely a set of approved states.

0.45 Trust Entropy as a Functorial Invariant

Trust entropy $\mathcal{T}(x) = \log |\mathcal{A}(x)|$ is the logarithm of the measure of $\mathbf{C}(X, \mathcal{A})$ at x. Since $\mathbf{C}$ is a functor, $\mathcal{T}$ is a functorial invariant: it is preserved (up to monotone transformation) by morphisms in **Cor**.

More precisely, if $f : (X, \mathcal{A}) \to (X', \mathcal{A}')$ is a morphism and f_* is measure-preserving, then $\mathcal{T}(x) = \mathcal{T}'(f(x))$. If f_* is measure-decreasing (a lossy transition), then $\mathcal{T}'(f(x)) \leq \mathcal{T}(x)$: the morphism cannot increase trust entropy, only preserve or reduce it.

This gives a categorical formulation of the paper's central claim: *precautionary interventions are morphisms that reduce* $\mathcal{T}$. The question of safe governance is the question of which morphisms in **Cor** can be composed without driving $\mathcal{T}$ below the critical threshold $\mathcal{T}_c = \log |\mathcal{A}_c|$ corresponding to the coherence boundary $\rho = 1$.

Theorem 14 (Functorial Preservation of Corrigibility)**.** *If $f : (X, \mathcal{A}) \to (Y, \mathcal{B})$ is an isomorphism in* **Cor***, then $\mathcal{T}_X(x) = \mathcal{T}_Y(f(x))$ for all $x \in X$.*

Proof. An isomorphism in **Cor** is a morphism f with an inverse $f^{-1} : (Y, \mathcal{B}) \to (X, \mathcal{A})$ such that $f \circ f^{-1} = \mathrm{id}_Y$ and $f^{-1} \circ f = \mathrm{id}_X$. Since both f and f^{-1} preserve admissibility, f induces a bijection $\mathcal{A}(x) \leftrightarrow \mathcal{B}(f(x))$. A bijection between measurable spaces that is bi-measurable preserves measure. Therefore $|\mathcal{A}(x)| = |\mathcal{B}(f(x))|$, and since $\mathcal{T} = \log |\cdot|$, we have $\mathcal{T}_X(x) = \mathcal{T}_Y(f(x))$. □ □

Trust entropy is thus a genuine invariant of the category **Cor**: isomorphic corrigible systems have identical trust entropies at corresponding states. This is the categorical analogue of the conservation of energy under symmetry: the structure-preserving transformations of **Cor** leave $\mathcal{T}$ invariant.

0.46 Connection to Semantic Infrastructure

The category **Cor** connects directly to the Semantic Infrastructure framework. In that framework, semantic modules are objects, and meaning-preserving transformations are morphisms. The alignment sheaf $\mathcal{F}$ of Appendix C defines a stack over **Cor**, assigning to each system its space of locally coherent sections. The Institutional Fragility Index $\mathfrak{F} = \dim \check{H}^1$ is then a cohomological obstruction to the existence of a global section of this stack—a global semantic coherence across the entire governance structure.

Systems with $\mathfrak{F} = 0$ admit a consistent global semantic alignment. Systems with large $\mathfrak{F}$ have many independent obstructions: local meaning is coherent, but no global synthesis is available without structural reorganization.

The Universal Precaution Paradox

The preceding formal apparatus was developed in the context of AGI governance, but the core result is domain-independent. We state it here as a general theorem applicable to any adaptive system whose stability depends on corrective exchange.

0.47 Setup

Let E denote the *effective stability* of an adaptive system—the system's capacity to recover from perturbations while remaining within its operational range. Let P denote the *precaution level*—the intensity of mechanisms introduced to suppress error, deviation, or unauthorized behavior.

In general, one expects E to be a function of P. The naive precautionary assumption is that this function is monotone increasing: more precaution yields more stability. The Universal Precaution Paradox denies this.

0.48 Statement

Theorem 15 (Universal Precaution Paradox)**.** *Let $\mathcal{S}$ be any adaptive system satisfying:*

(i) Stability depends on corrective exchange: E is a functional of the coupling structure $\{\kappa_{ij}\}$.

(ii) Precaution suppresses coupling: $\partial\kappa_{ij}/\partial P \leq 0$ for the channels through which corrections propagate.

(iii) The system has a critical coupling threshold: there exists $\kappa_c > 0$ below which $\mathcal{A}(x) \to \emptyset$.

Then there exists a critical precaution level $P_c > 0$ such that:

$$\frac{\partial E}{\partial P} > 0 \quad \text{for } P < P_c, \qquad \frac{\partial E}{\partial P} < 0 \quad \text{for } P > P_c.$$

That is, effective stability is a non-monotone function of precaution, achieving a maximum at P_c and declining thereafter.

Proof sketch. For $P < P_c$: the coupling strengths κ_{ij} remain above κ_c. Precaution suppresses genuinely destructive perturbations while leaving corrective channels open. The admissible set $\mathcal{A}$ is non-empty and growing; trust entropy $\mathcal{T}$ is positive and the system recovers from perturbations. Stability increases.

For $P > P_c$: precaution has driven $\kappa_{ij} < \kappa_c$ for the corrective channels. By the coupling-collapse theorem, $\mathcal{A}(x) \to \emptyset$. The system can no longer receive corrections. Perturbations accumulate

unaddressed. By the balance equation $\partial \mathcal{T}/\partial t + \nabla \cdot J_T = \Sigma_T$, with $J_T \to 0$ and Σ_T bounded, trust entropy decreases. Effective stability declines. The system is now less stable than it was at $P = 0$, because the corrective infrastructure that provided baseline stability has been destroyed in the process of suppressing error. □ □

0.49 Domain-Independent Instances

The theorem applies wherever conditions (i)–(iii) hold. Representative instances:

- **Immune systems**: Over-activation (autoimmune response) suppresses the corrective feedback that distinguishes self from non-self. Excessive immune precaution produces more damage than the pathogens it targets.
- **Scientific communities**: Excessive gatekeeping (peer review overreach, citation cartels, enforced paradigm conformity) suppresses heterodox correction channels. Science loses the capacity to revise entrenched error.
- **Memory architectures**: A note-taking or knowledge system that aggressively prunes uncertain or contradictory entries eliminates the semantic diversity required for revision. The archive becomes internally consistent but observationally blind.
- **Ecosystems**: Elimination of predator species to protect prey populations removes the trophic feedback that regulates prey behavior. The ecosystem destabilizes in the absence of the very pressure that maintained it.
- **Institutions**: Bureaucracies that suppress dissent, restrict information flow, and centralize decision authority eliminate the distributed error-correction that made them functional. They become rigid, uncorrectable, and eventually self-defeating.

In each case, the mechanism is the same: the precautionary measure suppresses the coupling through which corrections flow, and past the critical threshold, the system becomes less stable than it was before precaution was applied.

0.50 Deriving the Critical Threshold

The proof above assumes the existence of P_c without deriving it. Under a natural model of how precaution suppresses coupling, the threshold can be computed explicitly.

Suppose coupling decays exponentially with precaution:

$$\kappa(P) = \kappa_0 \, e^{-\mu P}, \tag{24}$$

where $\kappa_0 > 0$ is the baseline coupling and $\mu > 0$ is the suppression rate. The critical threshold κ_c is the coupling level below which $\mathcal{A}(x) \to \emptyset$ (from the coupling-collapse theorem). Setting $\kappa(P_c) = \kappa_c$ gives:

$$P_c = \frac{1}{\mu} \log\left(\frac{\kappa_0}{\kappa_c}\right). \tag{25}$$

This formula carries direct governance meaning. The ratio κ_0/κ_c is the *resilience margin*: how far the system's baseline coupling exceeds the critical floor. Systems with high resilience margin (large κ_0, small κ_c) tolerate substantial precautionary pressure before crossing into the paranoia phase. Fragile systems (small κ_0/κ_c) have a narrow window and cross P_c quickly.

The suppression rate μ measures how efficiently precaution destroys coupling. Heavy-handed institutional mechanisms (high μ) reach P_c at lower absolute precaution levels than light-touch

ones (low μ). The equation $P_c = \frac{1}{\mu}\log(\kappa_0/\kappa_c)$ is thus a design criterion: to maintain $P < P_c$, one must either increase the resilience margin or reduce the suppression rate—that is, build robust baseline trust structures before introducing precautionary controls.

0.51 Continuation Curvature and Institutional Geometry

The trust entropy field $\mathcal{T}(x)$ has a natural second-order geometric invariant. Define the *continuation curvature*:

$$K_{\mathcal{A}} := -\nabla^2 \mathcal{T}. \tag{26}$$

The sign of $K_{\mathcal{A}}$ determines the local geometry of the continuation space:

- $K_{\mathcal{A}} < 0$: $\mathcal{T}$ is locally convex; the continuation space is *expanding*. Perturbations from x lead to states with more recoverable futures. This is the signature of a healthy, self-reinforcing trust structure.
- $K_{\mathcal{A}} > 0$: $\mathcal{T}$ is locally concave; the continuation space is *contracting*. Each step narrows the set of admissible futures. The system is in a regime of trust erosion.
- $K_{\mathcal{A}} \to +\infty$: *continuation singularity*. The Laplacian of $\mathcal{T}$ diverges, signaling the collapse of $\mathcal{A}(x)$ to a set of measure zero. This is the geometric signature of the paranoia phase: not merely a reduction in available futures, but a topological catastrophe in which continuation space ceases to have interior.

The continuation singularity provides a geometric interpretation of institutional collapse that complements the algebraic interpretation via the Fragility Index $\mathfrak{F}$ and the spectral interpretation via $\lambda_2(L) \to 0$. All three descriptions—cohomological, spectral, and differential-geometric—characterize the same underlying phenomenon: the elimination of recoverable futures.

0.52 The General Principle

Systems survive not because they eliminate error, but because they preserve the capacity to recover from it. Precaution is beneficial insofar as it removes genuinely destructive perturbations without closing corrective channels. It becomes pathological when the mechanisms of suppression are indiscriminate—when they cannot distinguish error from correction, noise from signal, vulnerability from openness.

The formal condition is precise: $\partial E/\partial P < 0$ begins exactly when the suppression of coupling has driven the Fiedler eigenvalue $\lambda_2(L)$ below the threshold at which $\mathcal{A}$ collapses. At that point, the governance structure has crossed from precaution into paranoia—and the paradox is complete.

Epilogue: The Trust Singularity

The RSVP framework reveals that trust between minds—human or artificial—mirrors the dynamics of any open thermodynamic system: it is maintained by flow, not storage. The **Trust Singularity** is a phase transition to universal coherence, where alignment emerges from resonance, not control. The five appendices and the Universal Precaution Paradox together make precise what resonance requires: symmetric coupling (no panoptic asymmetry), open flow channels (no censored directions), scalar diversity (no enforced conformity), a communication topology with nonempty intersections (no isolated subpopulations), a continuation space of nonzero

measure (no topological collapse of recoverable futures), and morphisms in the category of corrigible systems that preserve rather than reduce trust entropy. The true risk is not AGI, but a civilization too afraid to remain open—one that, in its effort to eliminate dangerous futures, eliminates the very structure through which futures remain recoverable.

The entire argument culminates in a single theorem.

Theorem 16 (Trust–Continuation Equivalence)**.** *For any corrigible system, $\mathcal{T}(x) > -\infty$ if and only if $|\mathcal{A}(x)| > 0$.*

Proof. By definition, $\mathcal{T}(x) = \log |\mathcal{A}(x)|$. The logarithm is finite if and only if its argument is positive. Therefore $\mathcal{T}(x) > -\infty \iff |\mathcal{A}(x)| > 0$. □ □

The proof is immediate. The content is philosophical. Trust is not a belief about the future. It is not optimism, nor obedience, nor the absence of fear. Trust is the existence of at least one recoverable future. When precaution eliminates the last admissible trajectory—when $|\mathcal{A}(x)| = 0$—trust does not merely weaken. It ceases to exist. The system has not become safer. It has become the thing it feared.

References

- Christian, B. (2020). *The Alignment Problem: Machine Learning and Human Values.* W. W. Norton & Company.
- Hanson, R. (2023). "Ice-Cream Analogy and AI Goals." *Overcoming Bias.*
- Kastrenakes, J. (2025). "The AI Doomers Are Getting Doomier." *The Atlantic.*
- Kurzweil, R. (2024). *The Singularity Is Nearer: When We Merge with AI.* Viking.
- Matthews, D. (2025). "The AI Doomer Debate, Explained." *Vox.*
- Metz, C. (2023). "We Should Welcome The New AI Doomerism." *Forbes.*
- Paumgarten, N. (2024). "Among the A.I. Doomsayers." *The New Yorker.*
- Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control.* Viking.
- Simonite, T. (2025). "As AI Advances, Doomers Warn the Superintelligence Apocalypse Is Nigh." *NPR.*
- Yudkowsky, E., and Soares, N. (2025). *If Anyone Builds It, Everyone Dies.*