The Paradox of Precaution

How AGI Safety Could Erode Human Trust

A Thermodynamic Theory of Mutual Corrigibility

Flyxion

October 2025

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.

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.

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.

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 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's 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.

Appendix A: Alignment as Entropic Coupling in the Cognitive Field

0.13 The RSVP Field Interpretation

Within the Relativistic Scalar–Vector Plenum (RSVP)$framework, allintelligences^{-}biological, artificial, orin$
$(\Phi, \mathbf{v}, S)$, 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:

$$\nabla\Phi_i \cdot \mathbf{v}_j \approx \nabla\Phi_j \cdot \mathbf{v}_i. \tag{1}$$

0.14 Entropic Symmetry and Trust

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

$$\delta S_{ij} = \kappa_{ij}(\Phi_i - \Phi_j). \tag{2}$$

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.

0.15 Moral Feedback as Negentropic Coupling

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

$$\frac{dS_{\text{joint}}}{dt} = -\lambda \langle \nabla\Phi_i \cdot \mathbf{v}_j + \nabla\Phi_j \cdot \mathbf{v}_i \rangle. \tag{3}$$

0.16 Precaution as Entropic Stasis

Unilateral alignment policies represent frozen boundary conditions:

$$\mathbf{v} \cdot \mathbf{n} = 0 \quad \text{on all external surfaces.} \tag{4}$$

0.17 Co-Evolutionary Alignment as Dynamic Equilibrium

True safety is stationary entropic equilibrium:

$$\frac{dS_{\text{joint}}}{dt} \to 0. \tag{5}$$

0.18 Implications for AGI Governance

Maximize coherence subject to negentropic throughput:

$$\max_{\kappa_{ij}} C(\Phi, \mathbf{v}) \quad \text{subject to} \quad \dot{S}_{\text{total}} \leq 0. \tag{6}$$

0.19 Summary

0.20 Closing Reflection

In RSVP $terms, trustistheentropiccurrentthatsustainscoherence.Tosuppressthatcurrentinthenameofsafe$

Appendix B: A Trust Lagrangian for Mutual Corrigibility

0.21 Setup

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

0.22 The Trust Lagrangian

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

0.23 Stationarity Conditions

$$\kappa_{ij}^\star = \frac{\nabla\Phi_i \cdot \mathbf{v}_j + \nabla\Phi_j \cdot \mathbf{v}_i - 2(\lambda_i - \lambda_j)(\Phi_j - \Phi_i)}{2\gamma},$$
$$\mathbf{v}_i^\star = \frac{1}{\alpha}\left(\lambda_i \nabla\Phi_i + \frac{1}{2}\sum \kappa_{ji}\nabla\Phi_j\right) - \frac{\eta}{\alpha}\frac{\partial \mathcal{E}_i}{\partial \mathbf{v}_i},$$
$$\partial_t \lambda_i = \beta S_i.$$

0.24 Governance Readout

Raising γ increases mistrust; raising α bounds action. Maintain coherence window to avoid paranoia phase.

0.25 Summary

The Lagrangian formalizes alignment as mutual corrigibility under open entropic exchange—precaution erodes this at scale.

Epilogue: The Trust Singularity

The RSVP $frameworkrevealsthattrustbetweenminds^{-}humanorartificial^{-}mirrorscosmicevolution.The*$
$*TrustSingularity**isaphasetransitiontouniversalcoherence,wherealignmentemergesfromresonance,notc$

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.* Publisher.