The Geometry of Spherepop

A Calculus of Mutual Corrigibility in Intelligence Ecologies

With Applications to AGI Safety and the Paradox of Precaution

Flyxion

October 2025

When the machinery of safety becomes the machinery of suspicion.

Abstract

The Spherepop Calculus (SPC)$providesaformalgeometryofintelligenceasnested, merging, andprobabilistical$
$--centralizedmonitoring, epistemicrestriction, andhard-codedconstraints---collapsethecouplingcoeffic$
$boundedautonomy, dialogictransparency, andrecursivefeedback.TheprincipalriskofAGIprecautionismisnot$

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 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.

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 Geometry of Spherepop

A Formal Calculus of Intelligence Ecologies

The Spherepop Calculus (SPC)$formalizesintelligenceasgeometricoperationsonnestedspheresofagency.Itsco$

$$\begin{aligned} t, u ::= x \mid a \mid \mathrm{Sphere}(x : A.\ t) \mid \mathrm{Pop}(t, u) \mid \mathrm{Merge}(t, u) \\ \mid \mathrm{Nest}(t, u) \mid \mathrm{Choice}(p, t, u) \end{aligned}$$

with reduction rule β: $\mathrm{Pop}(\mathrm{Sphere}(x : A.\ t),\ u) \to t[u/x]$.

Spherepop Literals and Operators

The SPC $DSLprovidessurfacesyntaxforauthoringgeometricscenes:$

```
@scene {
  sphere f(type: Pi x:A.B, body: pop k with x)
  sphere k(type: Pi x:A.B, value: <primitive_k>)
  sphere a(type: A, value: a0)

  pop f with a
  choose 0.4: burst f(a) | burst k(a)
}
```

This lowers deterministically to SPC *core* :

$$\begin{aligned}
&f = \text{Sphere}(x : A.\ \text{Pop}(k,\ x))\\
&k = \text{<primitive\_k>} : \Pi x : A.B\\
&a = a_0 : A\\
&\ \text{Pop}(f,\ a)\\
&\ \text{Choice}(0.4,\ \text{Pop}(f,\ a),\ \text{Pop}(k,\ a))
\end{aligned}$$

Operators in SPC

- `link a -> b` — scheduling edge (meta).
- `link a` $\nabla b - - - differential flow$:`Pop(`$\nabla,\ \ (a,b)$`)`.
- `link a` $\otimes b - - - parallelmerge$:`Merge(a,b)`.
- `link a` $\oplus b - - - sharedscope$:Σ`-pair or Merge`.
- `link a` $\circ b - - - composition$:`Pop(a,b)`.

Typing and Reduction

Typing follows dependent Π/Σ rules with Merge requiring type equality and Choice requiring branch agreement. Evaluation is β-reduction plus stochastic sampling of Choice.

Spherepop and the Paradox of Precaution

Precautionary architectures correspond to frozen boundary conditions: $\mathbf{v}\cdot\mathbf{n} = 0$. In SPC , $thisisolatesspheres, p$
0. The result is entropic stasis: institutional paranoia and loss of mutual corrigibility.

By contrast, open SPC $scenes---richin$Pop,Merge, andChoice$---sustainnegentropicflux, modelingtrusta$

Appendix A: Alignment as Entropic Coupling in the Cognitive Field

0.8 The RSVP Field Interpretation

Within the Relativistic Scalar–Vector Plenum (RSVP)$framework, allintelligences - - -$
$biological, artificial, orinstitutional---aretreatedaslocalizedattractorswithinasharedscalar-$
$-vector - -entropyfield$:$(\Phi,\ \mathbf{v},\ S)$, $where\Phi denotesthescalarpotentialofintelligibility -$
$--thesystem'srepresentationalcapacityorinterpretivebandwidth;\mathbf{v}representsthevectorflowofagency-$
$--directedinfluenceoractionthroughtheplenum;Smeasurestheentropydensity---distributeduncertaint$

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:

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

0.9 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.

0.10 Moral Feedback as Negentropic Coupling

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

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

0.11 Precaution as Entropic Stasis

Unilateral alignment policies represent frozen boundary conditions:

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

0.12 Co-Evolutionary Alignment as Dynamic Equilibrium

True safety is stationary entropic equilibrium:

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

0.13 Implications for AGI Governance

Maximize coherence subject to negentropic throughput:

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

0.14 Summary

0.15 Closing Reflection

In RSVP $terms, trustistheentropiccurrentthatsustainscoherence. Tosuppressthatcurrentinthenameofsaf$

Appendix B: The Spherepop Calculus Core

0.16 Syntax

$$t, u ::= x \mid a \mid \mathrm{Sphere}(x : A.\ t) \mid \mathrm{Pop}(t, u) \mid \mathrm{Merge}(t, u) \mid \mathrm{Choice}(p, t, u)$$

0.17 Reduction

$$\mathrm{Pop}(\mathrm{Sphere}(x : A.\ t),\ u) \to t[u/x]$$

0.18 Typing (selected rules)

$$\frac{\Gamma \vdash A : \mathrm{Type} \quad \Gamma, x : A \vdash t : B}{\Gamma \vdash \mathrm{Sphere}(x : A.\ t) : \Pi x : A.B}$$
$$\frac{\Gamma \vdash t : \Pi x : A.B \quad \Gamma \vdash u : A}{\Gamma \vdash \mathrm{Pop}(t, u) : B[u/x]}$$
$$\frac{\Gamma \vdash t : A \quad \Gamma \vdash u : A}{\Gamma \vdash \mathrm{Merge}(t, u) : A}$$
$$\frac{\Gamma \vdash t : A \quad \Gamma \vdash u : A}{\Gamma \vdash \mathrm{Choice}(p, t, u) : A}$$

0.19 DSL Lowering Example

```
sphere f(type: Pi x:A.B, body: pop k with x)
pop f with a
```

lowers to:

$$f = \mathrm{Sphere}(x : A.\ \mathrm{Pop}(k, x)), \quad \mathrm{Pop}(f, a)$$

0.20 Haskell Backend Sketch

```
data Tm = Var Name | Sphere Name Ty Tm | Pop Tm Tm | Merge Tm Tm | Choice Double Tm Tm
```

Epilogue: The Trust Singularity

The RSVP and SPC $frameworksrevealthattrustbetweenminds---humanorartificial---mirrorscosmicevolution.The**TrustSingularity**isaphasetransitiontouniversalcoherence, wherealignmen$

References

- Christian, B. (2020). *The Alignment Problem.* W. W. Norton & Company.
- Russell, S. (2019). *Human Compatible.* Viking.
- Yudkowsky, E., & Soares, N. (2025). *If Anyone Builds It, Everyone Dies.*