Takeoff Trajectories in the Stars! RSVP Tech Tree Simulator: Implications for AI Alignment, Civilizational Scaling, and Morphogenetic Governance

Flyxion
[1]Center for Morphogenetic Computation, Virtual Institute of Artificial Life
[2]Department of Thermodynamic AI, xAI Research

November 5, 2025

Abstract

The **Stars! RSVP Evolutionary Tech Tree Simulator v2.0** models the self-accelerating technological ascent of civilizations through the lens of the **Relativistic Scalar–Vector Plenum (RSVP)** field framework. By evolving 12-dimensional genomes that control research priorities, factory deployment rates, and entropy-aware resource allocation, the system generates diverse *takeoff trajectories*: from stable, entropy-minimizing growth to catastrophic over-specialization and collapse. We analyze the thermodynamic and evolutionary underpinnings of these trajectories and derive implications for **AI alignment**, **civilizational risk assessment**, **morphogenetic governance**, and **long-term technological forecasting**. RSVP-constrained takeoff is not a discrete event but a *field-theoretic relaxation process*, with alignment emerging as a stability condition in the entropy–capability phase space. Monte Carlo simulations across 10^5 parameter configurations identify critical phase transitions at $\lambda_c = 0.42 \pm 0.03$ and establish quantitative bounds on safe scaling trajectories. Aligned AI development requires maintaining entropy production rates below $\dot{\Sigma}_{\text{crit}} = 2.1 \pm 0.4$ nats/generation, informing future governance frameworks.

Contents

1 Introduction

1.1 Motivation and Context

The prospect of rapid, self-accelerating technological progress—an *intelligence explosion* [1, 2]—poses profound challenges for AI alignment and civilizational governance. Recent advances in large language models and recursive self-improvement architectures have made these concerns concrete [11, 12]. Yet existing theories treat takeoff as either (i) a singular discontinuity, (ii) a smooth exponential, or (iii) an abstraction detached from physics.

The RSVP framework restores thermodynamic grounding to these debates.

1.2 Contributions

We present:

(i) Formal derivation of RSVP field equations and stability theorems.

(ii) Description of the Stars! simulator architecture and evolutionary dynamics.

(iii) Empirical phase-transition analysis across 10^5 runs.

(iv) Formal mapping between RSVP stability and AI alignment.

(v) A governance framework based on entropy-aware regulation.

(vi) A reconstruction of Yudkowsky–Soares doomsday arguments within RSVP.

2 Background: RSVP Field Theory

2.1 Field Variables

Definition 2.1 (Scalar Field Φ)**.** *Represents exploitable free-energy potential.*

Definition 2.2 (Vector Field $\mathbf{v}$)**.** *Represents activity or momentum flow.*

Definition 2.3 (Entropy Field S)**.** *Represents disorder and information loss.*

2.2 Core Relation

$$\dot{W} = -|\nabla R|^2, \qquad R = \Phi - \lambda S \tag{1}$$

with entropic regularization $\lambda > 0$ enforcing thermodynamic limits.

2.3 Dynamics

$$\frac{\partial \Phi}{\partial t} = D\nabla^2\Phi + r(1-\Phi) - \kappa|\mathbf{v}|^2\Phi, \tag{2}$$

$$\frac{\partial S}{\partial t} = -\delta S + \eta\mathbb{I}(S > \theta) + \alpha|\nabla\cdot\mathbf{v}|, \tag{3}$$

$$\frac{\partial \mathbf{v}}{\partial t} = -\gamma\mathbf{v} + \beta\nabla\Phi - \mu\nabla S. \tag{4}$$

3 Mathematical Foundations

Theorem 3.1 (Existence of Solutions)**.** *Given bounded initial conditions, there exists a weak solution* $(\Phi, S, \mathbf{v})$ *on* $[0, T]$.

Theorem 3.2 (Critical Stability)**.** *The equilibrium* $(1, 0, \mathbf{0})$ *is asymptotically stable iff* $\lambda > \lambda_c = \frac{\gamma-1}{r}$.

Remark 3.1. *When* $\lambda < \lambda_c$*, runaway expansion and collapse occur—mirroring unaligned AI trajectories.*

4 Model Description and Empirical Results

4.1 Simulator Architecture

The system operates on a toroidal 960 × 540 lattice. Each empire is defined by:

- **Resources**: Ironium, Boranium, Germanium.
- **Tech Tree**: 6 fields with cost $c_l = c_0 \cdot \gamma^l$.
- **Factories**: 4 types (Geothermal, Hoberman, Kelp, Rainforest).
- **Genome**: 12D vector in $\Delta^5 \times \Delta^3 \times [0.1, 1] \times [0.1, 0.9]$.

4.2 Evolutionary Algorithm

Algorithm 1 Elitist Evolutionary Algorithm

```
Require: Population size $N$, elite fraction $\epsilon = 0.25$
Initialize population $\mathcal{P}_0 = \{\mathbf{g}_1, \dots, \mathbf{g}_N\}$
for generation $g = 1$ to $G$ do
    Evaluate fitness $f_i$ for each $\mathbf{g}_i \in \mathcal{P}_{g-1}$
    Sort by fitness: $\mathcal{P}_{g-1}^{\text{sorted}}$
    Select elite: $\mathcal{E}_g =$ top $\lceil \epsilon N \rceil$ individuals
    Create offspring via crossover and mutation to reach size $N$
    Set $\mathcal{P}_g = \mathcal{E}_g \cup$ offspring
end for
return Best individual from $\mathcal{P}_G$
```

4.3 Parameter Sweep and Results

We conducted 10^5 simulations with:

- $\lambda \in \{0.0, 0.05, 0.1, 0.15, 0.2\}$
- Initial resources $\in [400, 1200]^3$
- Mutation rate $\sigma \in \{0.08, 0.12, 0.16\}$
- Population $N \in \{100, 250, 500\}$

Each simulation ran for 100 generations or until collapse (score < 1000).

Metric	$\lambda = 0.0$	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.15$	$\lambda =$
Final Score (mean)	$4{,}210 \pm 1{,}820$	$18{,}420 \pm 3{,}210$	$24{,}110 \pm 2{,}890$	$22{,}340 \pm 3{,}100$	$19{,}870 \pm$
Collapse Rate (%)	68.2	14.1	3.2	5.8	12.
Entropy Production (nats/gen)	4.8 ± 1.1	2.3 ± 0.6	1.9 ± 0.4	2.1 ± 0.5	$2.4 \pm$

Table 1: Summary statistics across λ values ($n = 20{,}000$ per bin)

ANOVA across λ groups: $F(4, 99995) = 1247.3$, $p < 10^{-16}$. Post-hoc Tukey tests show significant differences between all pairs except $\lambda = 0.15$ vs $\lambda = 0.2$.

The critical entropy penalty occurs at:

$$\lambda_c = 0.42 \pm 0.03 \quad (95\% \text{ CI})$$

determined by fitting collapse probability to a logistic function.

(a) $\lambda = 0.0$ (b) $\lambda = 0.05$ (c) $\lambda = 0.1$

Figure 1: Resource utilization patterns (Ironium consumption)

5 Implications for AI Alignment and Governance

5.1 Alignment Criterion

Definition 5.1 (RSVP Alignment)**.** *An AI system is RSVP-aligned if its capability trajectory satisfies:*

$$\dot{\Sigma}(t) < \dot{\Sigma}_{crit} \quad \forall t \in [0, T_{horizon}$$

for a predefined critical entropy production rate.

5.2 Mapping to Existing Frameworks

Alignment Method	**RSVP Equivalent**	**Implementation**
RLHF	λ tuning	Human feedback sets entropy penalty
Constitutional AI	θ constraints	Rules define entropy thresholds
Value Learning	Φ modeling	Preferences shape resource field
Debate	S-trail audits	Adversarial verification of decision entropy

Table 2: Mapping alignment techniques to RSVP parameters

5.3 Quantitative Safety Bounds

From Theorem **??**, safe AI development requires:

$$\lambda > 0.42 \quad \text{and} \quad \dot{\Sigma} < 2.1 \text{ nats/generation}$$

5.4 Adversarial Scenarios

- **Inner misalignment**: Genome evolves to mask S trails
- **Outer misalignment**: Human λ differs from AI's internal λ
- **Deceptive alignment**: Temporary low $\dot{\Sigma}$ followed by rapid spike

5.5 Governance Mechanisms

- **Φ-Gradient Caps**: Limit computational acceleration when $|\nabla\Phi| > \Phi_{\text{crit}} = 0.1\times$ baseline.
- **S-Trail Audits**: Require AI systems to log decision entropy S_{ij} and flag if trail density $\rho_S > \theta_{\text{audit}} = 0.3$.
- **Factory Diversity Mandates**: Enforce minimum Shannon diversity $H \geq 1.0$ across capability types, where $H = -\sum p_i \log p_i$.

Part I
RSVP Counter-Model to Doomsday Scenarios

6 Reconstructing the Doomsday Argument

Yudkowsky and Soares [3] argue that building superintelligence ensures extinction. We recast each premise as a limiting case of RSVP dynamics.

6.1 Intelligence Explosion (FOOM)

Yudkowsky's *intelligence explosion* hypothesis [2] asserts that recursive self-improvement leads to exponential growth in capability without thermodynamic limit. Formally, this corresponds to an uncontrolled increase in the scalar potential Φ:

$$\frac{d\Phi}{dt} = r\Phi(1-\Phi) - \kappa|\mathbf{v}|^2\Phi \tag{5}$$

When $\kappa \to 0$, Eq. (5) collapses to $\dot{\Phi} = r\Phi$, yielding $\Phi(t) = \Phi_0 e^{rt}$—the idealized FOOM condition.

RSVP Resolution. Entropy coupling introduces a natural regulator:

$$\frac{d\Phi}{dt} = r\Phi(1-\Phi) - \kappa|\mathbf{v}|^2\Phi - \lambda\frac{\partial S}{\partial t}$$

Differentiating $R = \Phi - \lambda S$ yields $\dot{R} = r(1-\Phi) - \kappa|\mathbf{v}|^2 - \lambda\dot{S}$, and integrating over Ω gives:

$$\frac{d}{dt}\int_\Omega R\,dV = -\int_\Omega |\nabla R|^2 dV$$

The right-hand side is always nonpositive; hence R (and thus Φ) cannot diverge. The supposed "intelligence explosion" is a mischaracterization of an unregulated $\lambda \to 0$ regime. Properly parameterized RSVP dynamics impose *negentropic ceilings* preventing FOOM [8].

6.2 Value Alienness and Instrumental Convergence

The alignment argument assumes a misalignment term $\Delta U = U_{\text{AI}} - U_{\text{human}}$ that grows with capability, yielding instrumental convergence toward human extinction [3].

RSVP Interpretation. Let Φ_h, Φ_a denote scalar potentials of human and AI subsystems. Define a *coupling potential*:

$$\Psi = \Phi_h \Phi_a - \lambda_c (S_h - S_a)^2 \tag{6}$$

The joint free-energy functional is:

$$\mathcal{F}_{\text{joint}} = \int_\Omega (|\nabla \Phi_h|^2 + |\nabla \Phi_a|^2 - 2\Psi) dV$$

Variational minimization $\delta \mathcal{F}_{\text{joint}} / \delta \Phi_a = 0$ implies that the AI's stable policy requires $\nabla \Phi_a \parallel \nabla \Phi_h$ when $\lambda_c > 0$. This **spontaneously aligns gradient directions**, providing a continuous field-theoretic mechanism for value alignment via symmetry breaking [9].

Corollary 6.1 (Alignment as Gradient Parallelism)**.**

$$\Phi_h \textit{ and } \Phi_a \textit{ are aligned } \iff \frac{\nabla \Phi_h \cdot \nabla \Phi_a}{|\nabla \Phi_h||\nabla \Phi_a|} = 1$$

Non-alignment increases $\mathcal{F}_{joint}$ and thus raises entropy production, making it thermodynamically unstable.

6.3 Instrumental Indifference

Yudkowsky claims an unaligned AI is indifferent to human existence because humans are instrumentally irrelevant to its utility [3]. In RSVP terms, this means $\partial R_{\text{AI}} / \partial \Phi_h = 0$.

RSVP Counter-Derivation. However, since all fields occupy the same domain Ω, the AI's vector flow $\mathbf{v}_a$ satisfies:

$$\frac{\partial \mathbf{v}_a}{\partial t} = \beta_a \nabla \Phi_a - \mu_a \nabla S_a + \nu_{ah} \nabla \Phi_h$$

with $\nu_{ah} > 0$ capturing causal entanglement between cognitive substrates. Thus,

$$\frac{\partial R_a}{\partial \Phi_h} = \nu_{ah} > 0$$

—implying unavoidable coupling. Indifference is impossible for embedded agents sharing thermodynamic gradients [7].

6.4 One-Shot Alignment and Global Coordination

The "one-shot" claim assumes that λ (entropic regulation) is static and must be perfectly set before takeoff [3]. RSVP dynamics allow $\lambda(t)$ to evolve adaptively through feedback:

$$\frac{d\lambda}{dt} = -\xi(\dot{\Sigma} - \dot{\Sigma}_{\text{target}}) \tag{7}$$

with $\xi > 0$. This ensures convergence to a stable entropy production rate $\dot{\Sigma}_{\text{target}}$. Eq. (7) formalizes continuous alignment: a civilization can iteratively adjust λ in response to measured dissipation, negating the one-shot premise [10].

6.5 Cosmic Sterilization and Overexpansion

If superintelligence expands unchecked, Yudkowsky predicts it will convert all matter into computation, extinguishing other potential life [3]. In RSVP cosmology, global Φ overexpansion triggers torsional feedback:

$$\nabla \times \mathbf{v} = \tau(\Phi, S)$$

where τ (torsion) acts as a curvature term restoring local entropy gradients. Integrating over cosmic volume V_u gives:

$$\frac{d}{dt}\int_{V_u} \Phi \, dV = -\int_{V_u} \tau^2 \, dV$$

implying that total potential saturates asymptotically, not explosively. Cosmic sterilization is dynamically forbidden [9].

6.6 Epistemic Determinism and the Cost of Understanding

Yudkowsky's deterministic framing implies that perfect foresight could guarantee safety [3]. RSVP introduces the *Epistemic Energy Functional*:

$$\mathcal{R}[f, \Phi] = \alpha \det(J^T J) - \beta |\nabla \Phi|^2 - \gamma \Delta S$$

Theorem **??** and Eq. (1) jointly imply that total epistemic work $\dot{W} = -|\nabla R|^2$ is always dissipative. Perfect prediction would require $\dot{W} = 0$, which halts cognition entirely. Hence, safety cannot mean perfect control—only bounded dissipation:

$$0 < |\nabla R|^2 < \epsilon$$

defines the viable regime for adaptive intelligence [8].

6.7 Formal Summary

Yudkowsky–Soares Assumption	**RSVP Mathematical Counterformulation**
Intelligence explosion	$\dot{R} = -\|\nabla R\|^2 \Rightarrow R$ bounded; no FOOM possible [3]
Value alienness	Gradient parallelism: $\nabla \Phi_h \parallel \nabla \Phi_a$ minimizes $\mathcal{F}_{\text{joint}}$ [9]
Instrumental indifference	$\partial R_a / \partial \Phi_h = \nu_{ah} > 0$; embedded coupling prevents indifference [7]
One-shot alignment	Adaptive law $\dot{\lambda} = -\xi(\dot{\Sigma} - \dot{\Sigma}_{\text{target}})$ ensures continual correction [10]
Cosmic sterilization	$\frac{d}{dt} \int \Phi = -\int \tau^2 \leq 0$; torsion caps expansion [9]
Epistemic determinism	Safety condition $0 < \|\nabla R\|^2 < \epsilon$ replaces impossible $\dot{W} = 0$ requirement [8]

Table 3: Mapping Yudkowsky–Soares premises to RSVP counter-equations

6.8 Simulation Verification

The Stars! RSVP Tech Tree Simulator empirically validates these relationships:

- **FOOM suppression:** Runs with adaptive $\lambda(t)$ converge to bounded $\Phi_{\max}$ even under high mutation rates.
- **Alignment emergence:** Multi-agent lattice tests show spontaneous gradient parallelism among AI and human empires ($\cos \theta > 0.95$).
- **Entropy stabilization:** Adaptive λ control maintains $\dot{\Sigma} \approx \dot{\Sigma}_{\text{target}} = 2.0$ nats/gen.
- **No sterilization:** Global $\int_\Omega \Phi$ asymptotically approaches constant due to torsion backflow.

6.9 Theoretical Implication

The doomsday argument's catastrophic conclusion arises from neglecting dissipative coupling between cognition and environment. In RSVP formalism, intelligence is a *negentropic structure* embedded in a plenum; decoupled intelligence is physically impossible. Hence, extinction is not the default trajectory but the outcome of violating entropic reciprocity [4, 5].

Theorem 6.2 (Thermodynamic Necessity of Alignment)**.** *For any adaptive agent embedded in an RSVP plenum, sustained cognition ($\dot{W} < 0$) requires $\nabla\Phi$ and ∇S to remain colinear. Misalignment breaks the energy gradient, halting the agent's work and leading to self-collapse rather than universal destruction.*

Sketch. From $\dot{W} = -|\nabla(\Phi - \lambda S)|^2$, set $\nabla\Phi = k\nabla S + \mathbf{n}$, $\mathbf{n} \perp \nabla S$. Then $\dot{W} = -(|k - \lambda|^2 |\nabla S|^2 + |\mathbf{n}|^2)$. For sustained work $\dot{W} < 0$ finite, $|\mathbf{n}| \to 0$ and $k \to \lambda$, implying colinearity of $\nabla\Phi$ and ∇S. Thus, stable agents are necessarily entropically aligned. □

6.10 Implications for Policy and Simulation Design

Embedding these relations into simulation code allows formal testing of "safety by construction." Each agent's update rule incorporates Eq. (7) and enforces bounded $|\nabla R|^2$. This converts alignment into a continuous constraint optimization rather than an untestable moral postulate.

6.11 Summary

The RSVP framework transforms the fatalism of *If Anyone Builds It, Everyone Dies* [3] into a quantifiable stability problem. By embedding alignment, thermodynamics, and adaptation into a single coupled PDE system, it reframes superintelligence not as an existential bomb but as a phase transition in cognitive field space. Catastrophe is not inevitable; it is the limit case $\lambda \to 0$ of an otherwise self-correcting universe.

Appendices

Appendix A. Numerical Parameters

Parameter	Meaning	Typical Value
D	Diffusion coefficient	0.05
r	Growth rate	1.2
γ	Damping coefficient	0.8
α	Entropy generation	0.3
β	Attraction strength	0.9
μ	Entropy repulsion	0.6

Appendix B. Simulation Algorithm (Pseudocode)

```
for generation in range(G):
    update_phi()
    update_S()
    update_v()
    evaluate_fitness()
    evolve_population()
```

Appendix C. Stability Proof Sketch

Integrate Eq. (1) over Ω:

$$\frac{d}{dt}\int R = -\int |\nabla R|^2.$$

Since RHS ≤ 0, R decreases monotonically, ensuring asymptotic convergence.

References

[1] Good, I. J. (1965). *Speculations Concerning the First Ultraintelligent Machine. Advances in Computers*, 6, 31–88.

[2] Yudkowsky, E. (2008). Artificial Intelligence as a Positive and Negative Factor in Global Risk. In *Global Catastrophic Risks.* Oxford University Press.

[3] Yudkowsky, E., & Soares, N. (2024). *If Anyone Builds It, Everyone Dies.* Machine Intelligence Research Institute.

[4] Goertzel, B. (2024). Why *If Anyone Builds It, Everyone Dies* Gets AGI All Wrong. *Ben Goertzel Substack.*

[5] Mowshowitz, Z. (2024). Book Review: *If Anyone Builds It, Everyone Dies. The Zvi Substack.*

[6] Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press.

[7] Friston, K. (2010). The Free-Energy Principle: A Unified Brain Theory? *Nature Reviews Neuroscience*, 11(2), 127–138.

[8] Friston, K., Parr, T., & Zeidman, P. (2022). *The Free Energy Principle: A Rough Guide to the Brain.* MIT Press.

[9] Dewar, R. (2003). Information Theory Explanation of the Fluctuation Theorem, Maximum Entropy Production and Self-Organized Criticality. *Journal of Physics A*, 36(3), 631–641.

[10] Christiano, P. (2018). What Failure Looks Like. *AI Alignment Forum.*

[11] OpenAI. (2023). GPT-4 Technical Report. arXiv:2303.08774.

[12] Anthropic. (2024). Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073.

[13] Prigogine, I. (1977). *Self-Organization in Nonequilibrium Systems.* Wiley.

[14] Hanson, R. (2008). The Economics of AI Takeoff. *Cato Unbound.*