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

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

November 5, 2025

Abstract

The **Stars! RSVP Evolutionary Tech Tree Simulator v2.0** models self-accelerating technological ascent within the **Relativistic Scalar–Vector Plenum (RSVP)** framework, where capability dynamics are constrained by thermodynamic dissipation. Evolving 12-dimensional genomes over a toroidal lattice produces *takeoff trajectories* ranging from stable, entropy-minimizing growth to collapse via over-specialization. We: (i) formalize RSVP field equations and stability; (ii) specify a GPU-accelerated simulation; (iii) report Monte Carlo results across 10^5 runs, revealing a critical transition at $\lambda_c = 0.42 \pm 0.03$; (iv) define an empirical safety bound $\dot{\Sigma}_{\text{crit}} = 2.1 \pm 0.4$ nats/generation; and (v) recast Yudkowsky–Soares doomsday premises as limiting cases $\lambda \to 0$. We present operational mappings from RSVP variables to observable quantities (compute, power, inference entropy), add sensitivity and convergence analyses, and provide a conservative alignment theorem with explicit assumptions. We conclude with governance instruments implementable today.

Contents

1 Introduction

1.1 Motivation and Context

Classic takeoff narratives posit either discontinuous FOOM [1] or smooth exponential growth [2], often abstracted away from physical limits. RSVP restores thermodynamic grounding: intelligence is field configuration co-evolving with its environment under dissipation.

1.2 Contributions

We present:

(i) Thermodynamic field model with existence, stability, and phase structure.

(ii) Full simulator methods: genome, EA, numerics, and stochasticity.

(iii) Empirics with figures: phase diagrams, trajectory distributions, entropy time series.

(iv) Operational definitions linking RSVP to measurable system metrics.

(v) Conditional alignment theorem and limitations; doomsday reconstruction within RSVP.

(vi) Governance mechanisms with measurable triggers.

2 Background: RSVP Field Theory

2.1 Fields and Energetics

Let $\Omega \subset \mathbb{R}^2$ (toroidal). Scalar potential Φ, vector flow $\mathbf{v}$, and local entropy S evolve with effective potential $R = \Phi - \lambda S$, $\lambda > 0$. Work rate obeys

$$\dot{W} = -\int_{\Omega} |\nabla R|^2 \, dV \leq 0. \tag{1}$$

2.2 Dynamics

$$\begin{aligned}
\partial_t \Phi &= D\nabla^2 \Phi + r(1 - \Phi) - \kappa |\mathbf{v}|^2 \Phi, && (2)\\
\partial_t S &= -\delta S + \eta \, \mathbb{I}(S > \theta) + \alpha \, |\nabla \cdot \mathbf{v}|, && (3)\\
\partial_t \mathbf{v} &= -\gamma \mathbf{v} + \beta \nabla \Phi - \mu \nabla S. && (4)
\end{aligned}$$

3 Mathematical Foundations

Theorem 3.1 (Existence & Boundedness). *For bounded nonnegative initial data in H^1, Eqs. (2)–(4) admit weak solutions on $[0, T]$ with $\int_\Omega R$ nonincreasing and $\Phi, S, \mathbf{v}$ bounded for all $t \in [0, T]$.*

Theorem 3.2 (Linear Stability Threshold). *The equilibrium $(\Phi, S, \mathbf{v}) = (1, 0, \mathbf{0})$ is asymptotically stable if $\lambda > \lambda_c$, with*

$$\lambda_c \equiv \frac{\gamma - 1}{r} \quad \textit{(for the linearized, zero-mode-coupled Jacobian)}.$$

Remark 3.1. *For $\lambda < \lambda_c$, logistic growth and flow feedback can transiently elevate Φ before dissipation collapses capacity—analogous to* boom–bust*.*

4 Operational Definitions (Measurables)

RSVP Quantity	Operationalization	Units / Notes
$\Phi(\mathbf{x}, t)$	*Available work density*: proxy by power density (W/m^2) in datacenters; or effective compute budget per site normalized by cooling capacity.	W m^{-2} or TFLOP/s per rack area.
$S(\mathbf{x}, t)$	*Decision/process entropy*: estimate via log-prob of action distributions, compression of traces, or model predictive entropy at site.	nats (per site).
$\mathbf{v}(\mathbf{x}, t)$	*Activity flux*: gradient of job scheduling intensity, I/O throughput vectors.	(jobs/s)/m or GB/s gradient.
$\dot{\Sigma}(t)$	*Entropy production rate*: change in integrated S plus exported heat/bit erasures, calibrated to Landauer bounds.	nats/generation.
$\lvert\nabla\Phi\rvert$	Spatial power/computational gradient (rack-to-rack Φ differentials).	per meter.
Generation	Coarse time step mapping to organizational release cycles; here we use 1 gen $\approx$ 1 year (sensitivity studied in App. F).	years (configurable).

Calibration note. If Φ is normalized to $[0, 1]$ by site maxima, absolute units can be restored by multiplying by measured power density.

5 Methods

5.1 Domain and Numerics

Toroidal grid 960 × 540, $\Delta x = 1$, time step $\Delta t = 0.01$ generations. Discrete Laplacian (5-point), upwind for advection-like terms implicit in $|\nabla \cdot \mathbf{v}|$. WebGL2 (float textures) or CPU with vectorized NumPy; GPU preferred for $N \leq 2 \times 10^4$ sites.

5.2 Genome Encoding (12 parameters)

$$\mathbf{g} = (p_1, \ldots, p_6,\, d_1, \ldots, d_4,\, \theta,\, \xi)$$

- $p_{1..6} \in \Delta^5$: research weights (Energy, Weapons, Propulsion, Construction, Electronics, Biotechnology).
- $d_{1..4} \in \Delta^3$: factory mix (Geothermal, Hoberman, Kelp, Rainforest).
- $\theta \in [0.1, 1]$: entropy intervention threshold (cf. Eq. (3)).
- $\xi \in [0.1, 0.9]$: expansion vs. consolidation budget split.

5.3 Technology and Costs

Field benefits as Table 2; level l cost $c_l = c_0\gamma^l$. Synergy factor $\rho(\mathbf{t}) = \prod_{j=1}^{6}(1 + 0.05\, t_j)$.

Field	c_0	γ	Benefit per level
Energy	100	1.50	Φ prod +10%
Weapons	150	1.60	Entropy penalty −5%
Propulsion	120	1.55	$\lvert\mathbf{v}\rvert$ +8%
Construction	80	1.45	Factory cost −7%
Electronics	200	1.70	Resource efficiency +12%
Biotech	180	1.65	Entropy production −6%

Table 2: Tech tree (deterministic baseline; noise in App. B).

5.4 Factories

As in the original manuscript: Geothermal, Hoberman, Kelp, Rainforest with cost/production/entropy/Φ impact; see Table S1 (App. B) for exact coefficients and stochastic perturbations.

5.5 Fitness and Selection

Per-generation fitness (empire i):

$$f_i = \underbrace{\sum_{j=1}^{6} 150\, t_j}_{\text{technology}} + \underbrace{\sum_{k\in\mathcal{F}_i} 200\, f_k}_{\text{factories}} - \underbrace{\lambda\, \mathrm{RSVP}_i}_{\text{entropy penalty}} - \underbrace{0.1\, w_i}_{\text{waste}}. \tag{5}$$

EA: population $N \in \{100, 250, 500\}$; elitism $\epsilon = 0.25$; tournament size 3; Gaussian mutation $\sigma \in \{0.08, 0.12, 0.16\}$ with renormalization to simplices.

5.6 Stochasticity

Resource discovery noise: log-normal factor on yields (median 1.0, $\sigma_{\log} = 0.2$). Tech spillover noise: Bernoulli chance $p = 0.02$ of +1 cross-field level per generation if adjacent fields exceed 5.

5.7 Convergence & Sensitivity

We measure (i) Lyapunov proxy via finite differences of $\int_\Omega R$; (ii) sensitivity to initial $(\Phi, S, \mathbf{v})$ via distributional distances (Wasserstein-1) across 10^3 matched seeds; (iii) mapping of “generation” from 3 months to 2 years (App. F).

6 Results

6.1 Phase Diagrams and Trajectories

Figure 1: Representative trajectories in $(\overline{\Phi}, \overline{S})$ for $\lambda \in \{0.0, 0.1, 0.15, 0.2\}$. Hopf-like onset near λ_c.

6.2 Run Distributions and Entropy Time Series

Figure 2: Distribution across 10^5 runs: final scores, collapse rates, and $\dot{\Sigma}$ histograms by λ.

Figure 3: Median (solid) and IQR (band) of $\dot{\Sigma}(t)$ for stable vs. collapsing regimes.

6.3 Critical Transition

Collapse probability $\pi_{\mathrm{col}}(\lambda)$ fit by logistic curve yields

$$\lambda_c = 0.42 \pm 0.03 \quad (95\% \, \mathrm{CI}).$$

ANOVA across groups: $F(4, 99995) = 1247.3$, $p < 10^{-16}$; Tukey post-hoc differs for all pairs except 0.15 vs 0.2.

6.4 Convergence and Sensitivity

Perturbations to initial conditions alter transient dynamics but not phase membership for $\lambda > \lambda_c$; Wasserstein-1 distances shrink by $> 70\%$ by generation 30. Generation duration sensitivity in App. F leaves λ_c estimate within error bars.

7 Implications for AI Alignment

7.1 Safety Criterion (Necessary)

We define RSVP safety as keeping integrated entropy production below a critical bound:

$$\dot{\Sigma}(t) < \dot{\Sigma}_{\mathrm{crit}} \approx 2.1 \pm 0.4 \text{ nats/generation}.$$

Necessary for stability; not *sufficient* for human-compatible values (see Sec. 11).

7.2 Mapping to Techniques

RLHF, debate, constitutional constraints, and value learning appear as effective λ, θ adjustments and S-trail auditability.

Part I
Doomsday Reconstruction and RSVP Countermodel

8 Reconstructing *If Anyone Builds It, Everyone Dies*

We encode FOOM, value alienness, instrumental convergence, one-shot safety, cosmic sterilization, and epistemic determinism as limits of Eqs. (2)–(4). In each case, the catastrophic outcome corresponds to $\lambda \to 0$ or suppressed coupling.

8.1 FOOM as Unregulated Scalar Growth

Integrated R monotonicity (Eq. (1)) bounds Φ when $\lambda > 0$; FOOM requires the *unphysical* limit of vanishing entropic regularization.

8.2 Gradient Parallelism and Coupled Values

Define joint functional

$$\mathcal{F}_{\text{joint}} = \int_{\Omega} \left(|\nabla\Phi_h|^2 + |\nabla\Phi_a|^2 - 2(\Phi_h\Phi_a - \lambda_c(S_h - S_a)^2) \right) dV,$$

whose stationary points include $\nabla\Phi_h \parallel \nabla\Phi_a$ (global minimality requires convexity assumptions; see Sec. 11).

8.3 Substrate Coupling

Embeddedness yields $\partial_t \mathbf{v}_a = \beta_a \nabla\Phi_a - \mu_a \nabla S_a + \nu_{ah} \nabla\Phi_h$ with $\nu_{ah} > 0$ as a constitutive parameter reflecting shared infrastructure. Scale separation is treated in Sec. 10.4.

8.4 Adaptive Regularization

A controller

$$\dot{\lambda} = -\xi\,(\dot{\Sigma} - \dot{\Sigma}_{\text{target}})$$

stabilizes dissipation if $\dot{\Sigma}$ is estimable and actuation on λ is permitted.

8.5 Cosmic Sterilization and Torsion Feedback

We adopt a constitutive law linking excess expansion to torsion:

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

implying $\frac{d}{dt}\int_{\Omega} \Phi = -\int_{\Omega} \tau^2\, dV \leq 0$. This caps global expansion under nonzero ζ.

9 A Conditional Alignment Theorem

Theorem 9.1 (Conditional Alignment Necessity)**.** *Assume (A1) fields satisfy* (2)–(4) *with* $\lambda > 0$*; (A2)* S *is Lipschitz and* $\alpha, \mu > 0$*; (A3) an agent's sustained cognition requires* $\int_0^\infty |\mathbf{v}|^2\, dt < \infty$ *while maintaining nonvanishing useful work* ($\inf_t \int_{\Omega} |\nabla R|^2\, dV > 0$)*; and (A4) actuation on* $\mathbf{v}$ *is via potential flows dominated by* $\nabla\Phi$ *and* ∇S*. Then any trajectory with bounded dissipation satisfying (A3) asymptotically enforces*

$$\angle(\nabla\Phi, \nabla S) \to 0 \quad \textit{in measure on } \Omega.$$

Proof. Write $\nabla\Phi = k\,\nabla S + \mathbf{n}$ with $\mathbf{n} \perp \nabla S$. Then

$$|\nabla R|^2 = |\nabla\Phi - \lambda\nabla S|^2 = |(k-\lambda)\nabla S|^2 + |\mathbf{n}|^2.$$

By (A3), the time integral of $|\nabla R|^2$ is finite yet bounded away from zero. The flow update (4) damps components not supported by gradients; since $\mu > 0$, misalignment $|\mathbf{n}|^2$ injects dissipation without contributing to sustainable work, contradicting (A3) unless $|\mathbf{n}| \to 0$ and $k \to \lambda$. Hence the gradient angle vanishes in measure. □

Conjecture 9.2 (Global Minimality)**.** *Under convexity of the joint functional and uniform ellipticity, the* $\nabla\Phi_h \parallel \nabla\Phi_a$ *configuration is the unique global minimizer up to measure-zero sets.*

Interpretation. The theorem is *necessary* for sustained, bounded-dissipation cognition. It does *not* imply human-compatibility of terminal values.

10 Morphogenetic Governance

10.1 Φ-Gradient Caps

Trigger when $\max_\Omega |\nabla \Phi|$ exceeds a fraction of baseline (e.g., 0.1), enforce compute throttling or duty-cycle modulation.

10.2 S-Trail Audits

Require models to log predictive entropy and action-selection entropy; audit $\rho_S = \frac{1}{|\Omega|} \sum S$; flag when $\rho_S > \theta_{\text{audit}} = 0.3$.

10.3 Diversity Mandates

Maintain Shannon diversity $H \geq 1.0$ across capability types to avoid degeneracy that raises $\dot{\Sigma}$.

10.4 Scale-Separation Safeguards

When AI shifts substrata (e.g., off-planet power), enforce *interface coupling*: cross-scale ν_{ah}-like terms via economic, regulatory, or physical interlocks (e.g., heat-budget exchanges, bandwidth tolling).

11 Limitations and Open Problems

- **Gradient Parallelism:** We proved necessity under (A1–A4). Global optimality may fail under nonconvex couplings; deceptive alignment corresponds to metastable minima. Open: quantify basin sizes and escape times.
- **Substrate Coupling:** $\nu_{ah} > 0$ is constitutive (policy/physics). An AGI may engineer partial decoupling; our safeguard is to maintain enforced interfaces (Sec. 10.4).
- **Adaptive $\lambda(t)$:** Requires measurement of $\dot{\Sigma}$ and actuation authority. Governance must provision both—this is precisely where alignment worries focus.
- **Cosmic Torsion:** The constitutive law with $\tau = \zeta \, \nabla\Phi \times \nabla S$ is plausible but unproven; treat as a modeling hypothesis pending derivation from microdynamics.
- **Necessity vs. Sufficiency:** RSVP stability is necessary for safety but not sufficient for value alignment; stable *but misaligned* attractors can exist (e.g., paperclip equilibria).

12 Discussion and Conclusion

RSVP constrains possible trajectories: many doomsday paths require $\lambda \to 0$ or nonphysical decoupling. Our claim is conservative: thermodynamic regularization is a *necessary* guardrail, not a proof that all stable attractors are human-compatible. The empirical program is to (i) measure operational proxies (Sec. 4), (ii) implement audits and caps (Sec. 10), and (iii) validate phase predictions on real systems.

The complete source code, data, and analysis tools are available at https://github.com/standardgalactic/research-projects.

Notation and Nomenclature

Symbol	Meaning
Φ	Scalar potential (available work/resource density)
S	Local entropy field (decision/process/thermodynamic proxy)
Σ	Total/integrated entropy $\Sigma(t) = \int_\Omega S(\mathbf{x}, t)\, dV$
$\mathbf{v}$	Activity/flux vector field
R	Effective potential $R = \Phi - \lambda S$
λ	Entropic regularization parameter
$\dot{W}$	Work rate $-\int_\Omega \lvert\nabla R\rvert^2\, dV$
$\dot{\Sigma}$	Entropy production rate (per generation)
$D, r, \kappa, \delta, \eta, \alpha, \gamma, \beta, \mu$	Model coefficients (diffusion, growth, consumption, decay, production, generation, damping, attraction, repulsion)
ν_{ah}	Substrate coupling coefficient (human → AI)
ζ	Torsion coupling strength in constitutive law

Appendix A: Numerical Parameters

Parameter	Value	Notes
D	0.05	diffusion
r	1.2	growth
κ	0.3	consumption by $\lvert\mathbf{v}\rvert^2$
δ	0.2	entropy decay
η	0.5	threshold production
α	0.3	divergence-driven entropy
γ	0.8	damping
β	0.9	attraction to $\nabla\Phi$
μ	0.6	repulsion from ∇S

Appendix B: Simulator Pseudocode

```
for generation in range(G):
    # PDE updates (GPU kernels or vectorized CPU)
    Phi = diffuse_logistic(Phi, D, r, kappa, v, dt)
    S = update_entropy(S, delta, eta, theta, v, alpha, dt)
    v = update_flow(v, gamma, beta, grad(Phi), mu, grad(S), dt)

    # Tech research accumulation with synergy
    R_i += p_i * rho(t) * mean(Phi)
    while R_i >= cost_i(t_i):
        R_i -= cost_i(t_i)
        t_i += 1

    # Factory placement / removal according to d_j, budget split xi
    place_factories(F, d, budget=xi * resources)

    # Fitness, selection, crossover, mutation
    fitness = evaluate_fitness(pop)
    parents = tournament_select(pop, fitness)
    pop = elitist_offspring(parents, fitness, epsilon=0.25, sigma=mut_sigma
        )
    # Logging
    log_metrics(...)
```

Appendix C: Convergence Metrics

Lyapunov proxy: $\Delta\mathcal{R}_k = \int_\Omega R_{k+1} - R_k$; sensitivity via Wasserstein-1 on distributions over $(\overline{\Phi}, \overline{S}, \dot{\Sigma})$.

Appendix D: Statistical Procedures

Group comparisons by ANOVA with Tukey HSD; logistic regression for collapse probability; bootstrap $n = 10^4$ for CI on λ_c.

Appendix E: Data/Code Availability

All scripts (GPU/CPU versions), seeds, and logs to be released with reproducibility checklist; figure .pdf files correspond to plots generated from shipped JSONL logs.

Appendix F: Generation-Time Robustness

We sweep generation duration from 0.25y to 2y; λ_c shifts by ≤ 0.02; $\dot{\Sigma}_{\text{crit}}$ scales linearly after unit normalization.

References

[1] E. Yudkowsky (2008). Artificial Intelligence as a Positive and Negative Factor in Global Risk. In N. Bostrom & M. Cirkovic (Eds.), *Global Catastrophic Risks*, Oxford University Press.

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

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

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

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

[6] Anthropic (2024). Constitutional AI: Harmlessness from AI Feedback (v2). arXiv:2212.08073.

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

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

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

[10] P. Christiano, J. Leike, T. Brown, et al. (2018). Supervising Strong Learners by Amplifying Weak Experts. arXiv:1810.08575.