Divided We Stand: RSVP and the Limits of Coherence in AI Safety

Flyxion

September 18, 2025

Abstract

This article situates the Relativistic Scalar Vector Plenum (RSVP) framework within current debates on artificial intelligence safety. It contrasts Eliezer Yudkowsky and Nate Soares's book *If Anyone Builds It, Everyone Dies* [Yudkowsky and Soares, 2025] with critiques by Scott Alexander [Alexander, 2025], Clara Collier [Collier, 2025], and Robin Hanson [Hanson, 2024]. The analysis highlights RSVP's distinctive contribution: resilience through entropy-respecting diversity. Where the Machine Intelligence Research Institute (MIRI) emphasizes singular clarity and diamond-hard priors, RSVP insists that survival depends on recursive adaptability, re-bubbling, and dispersion across scalar, vector, and entropic fields (Φ, v, S). Building on foundational thermodynamic and quantum theories [Jacobson, 1995, Verlinde, 2011, Barandes, 2022], RSVP offers a unified architecture for physics, cognition, and governance, providing a counterpoint to MIRI's over-coherent worldview.

Contents

1 Introduction

The publication of Yudkowsky and Soares's *If Anyone Builds It, Everyone Dies* [Yudkowsky and Soares, 2025] has intensified debate about the fate of artificial intelligence and humanity. The Machine Intelligence Research Institute (MIRI) advances a position of near-total certainty: superintelligence entails extinction, and only a global ban on AI development can avert it. Critics have responded from multiple directions. Scott Alexander [Alexander, 2025] questions the practicality of this radical clarity, describing MIRI's stance as rhetorically powerful but politically inert. Clara Collier [Collier, 2025] presses further, arguing that the book represents a regression: the intelligence explosion hypothesis is barely defended, and the MIRI worldview has failed to update since the early blogosphere era. Robin Hanson [Hanson, 2024] adds a cultural critique, warning that over-coherence produces fragility by eroding diversity. Taken together, these responses illuminate the political, epistemic, and cultural costs of excessive coherence.

Against this backdrop, the Relativistic Scalar Vector Plenum (RSVP) framework proposes a different strategy. RSVP models reality as a tripartite field (Φ, v, S) of scalar capacity, vector flow, and entropy. It extends beyond physics into cognition, semantics, and governance, formalizing the principle that coherence without entropy is collapse. RSVP builds on prior theoretical advances: Jacobson's derivation of Einstein's equations from thermodynamics [Jacobson, 1995], Verlinde's entropic account of gravity [Verlinde, 2011], and Barandes's unistochastic reformulation of quantum theory [Barandes, 2022]. Each of these contributions reinterprets fundamental laws as emergent from deeper statistical or probabilistic substrates. RSVP synthesizes these insights into a universal architecture, one that encodes adaptability through entropy-respecting mechanisms.

This article places MIRI's arguments and their critiques in dialogue with RSVP. Where Yudkowsky and Soares call for singular clarity and diamond-hard priors, RSVP insists that survival depends on recursive adaptability, re-bubbling, and dispersion. The central thesis is that divided we stand: diversity and entropy are not liabilities but the conditions of resilience across physics, cognition, and society.

2 The Core Framework

RSVP models reality in terms of scalar semantic density Φ, vector flow v, and entropy S. Its action functional is defined as

$$F[\Phi, v, S] = \int_\Omega \left(\tfrac{\kappa_\Phi}{2} |\nabla \Phi|^2 + \tfrac{\kappa_v}{2} |\nabla \times v|^2 + \tfrac{\kappa_S}{2} |\nabla S|^2 - \lambda \Phi S \right) dx.$$

These fields jointly describe cosmological entropic redshift, neural and cognitive dynamics, and semantic infrastructures. In cosmology, Φ represents matter density, v velocity flows,

and S entropic expansion [Jacobson, 1995]. In cognition, Φ encodes semantic coherence, v attentional flows, and S unresolved ambiguity [Friston, 2010]. The universality of RSVP depends on extending this local partial differential equation structure into broader domains: topological, quantum, ethical, and socio-political.

3 Beyond the Core Framework

RSVP generalizes through four principal extensions, each addressing a distinct domain while preserving the tripartite structure.

3.1 Topological Domain

Local patches of cosmological cells, neural states, or semantic contexts require sheaf-theoretic gluing. RSVP encodes this in the MERGE operator:

$$\text{Emerge} = \int_{\Omega} \Big(\lambda_{\Phi} \|\Phi_1 - \Phi_2\|^2 + \lambda_v \|v_1 - v_2\|^2 + \lambda_S \, \mathrm{KL}(S_1 \| S_2) \Big) \, dx.$$

This MERGE operator functions as an entropic homotopy colimit, minimizing inconsistency across overlapping covers [Lurie, 2009].

3.2 Quantum Domain

Transitions between RSVP states are modeled as unistochastic maps, embedding unitary quantum evolution as a zero-entropy subset within the RSVP action functional [Barandes, 2022].

3.3 Ethical Domain

RSVP defines an entropy-production functional along worldlines:

$$\Sigma[\Gamma] = \int_{\Gamma} \max(0, \dot{S}) \, dt,$$

formalizing irreversible harm as positive entropy growth along trajectories, aligning with thermodynamic ethics [Verlinde, 2011].

3.4 Socio-Political Domain

Governance is reframed as an entropic budgeting problem. The Landauer inequality applies directly:

$$\Delta S(\Omega) \;\geq\; (\ln 2) \, N_{\text{erased}},$$

treating institutional erasure (e.g., data deletion, strategy suppression) as an entropic cost [Landauer, 1961].

3.5 Integrative Closure

All four domains—topological, quantum, ethical, and socio-political—are projections of (Φ, v, S), demonstrating RSVP's methodological parsimony: a single substrate suffices to generate multiple disciplinary projections. This contrasts with MIRI's static priors, emphasizing adaptive coherence over rigid clarity.

4 Comparative Frameworks

4.1 Incrementalism vs. Radical Clarity

Scott Alexander [Alexander, 2025] situates *If Anyone Builds It, Everyone Dies* within the divide between moderate incrementalists and MIRI's radical clarity. Incrementalists advocate for transparency, institutional accountability, and incremental safety research, acknowledging a non-trivial probability of catastrophic failure. Yudkowsky and Soares, by contrast, assert that superintelligence entails extinction with near-certainty, necessitating an immediate global ban on AI development [Yudkowsky and Soares, 2025]. Alexander likens MIRI's strategy to "wearing a hard hat against an asteroid impact," rhetorically compelling but practically inert. From the RSVP perspective, Alexander's critique highlights the limitations of singular interventions. A hard ban, like a hard hat, is a single-attractor strategy that fails to integrate entropic feedbacks. RSVP prioritizes recursive adaptability: multiple partial interventions that inject entropy, diversify responses, and avoid brittle over-commitments.

4.2 Parables and Archetypes

The narrative force of *If Anyone Builds It* lies in its parables: gods playing evolutionary games or misaligned chatbots pursuing perverse goals [Yudkowsky and Soares, 2025]. These echo the tragic machine Mima in Harry Martinson's *Aniara* [Martinson, 1956], which collapses under semantic overload, and the Tower of Babel [The Hebrew Bible], where over-coherence of language is disrupted by dispersion. RSVP interprets these not as mere stories but as field dynamics: runaway attractors (over-concentrated Φ, over-aligned v, suppressed S) collapse without entropy-injection mechanisms.

4.3 The Problem of Non-Update (Collier's Critique)

Clara Collier [Collier, 2025] identifies epistemic rigidity as the book's deepest flaw. The intelligence explosion hypothesis, the load-bearing premise of MIRI's argument, receives only two sentences, despite radical shifts in AI since 2008—from hand-coded architectures to empirically tractable deep learning systems governed by scaling laws. Yet, Yudkowsky and Soares repeat early blogosphere-era arguments almost verbatim, failing to engage with new evidence [Collier, 2025]. Collier calls this a "regression," noting that MIRI shadowboxes outdated opponents rather than contemporary debates. From the RSVP vantage, this reflects epistemic over-coherence. RSVP encodes update mechanisms: the MERGE operator integrates local evidence into global coherence, and the entropy-production functional $\Sigma[\Gamma]$

formalizes responsiveness to novel conditions [Verlinde, 2011]. Where MIRI's priors have the "strength of diamond," RSVP treats adaptability as a principle of resilience.

4.4 Implications for Governance

Both Alexander and Collier converge on governance. Alexander doubts the feasibility of global GPU bans [Alexander, 2025]; Collier questions the empirical assumptions motivating them [Collier, 2025]. RSVP reframes governance as an entropic budget allocation problem: stability arises from distributing flows across multiple channels, preserving resilience through diversity [Landauer, 1961]. Just as Babel scattered language to prevent collapse [The Hebrew Bible], RSVP prescribes dispersion of semantic, technological, and institutional pathways to avert brittle failure.

4.5 Cultural Drift and the Perils of Monoculture (Hanson's Critique)

Robin Hanson [Hanson, 2024] warns that cultural drift erodes resilience, leaving societies brittle through monoculture. In evolutionary dynamics, variation across lineages preserves adaptability; in markets, heterogeneous strategies sustain liquidity; in cognition, divergent perspectives prevent epistemic lock-in [Hanson, 2024]. Excessive alignment collapses adaptive capacity, producing fragile systems. RSVP formalizes this as entropic resilience: perturbations in S preserve adaptive capacity against runaway attractors. The Tower of Babel [The Hebrew Bible] serves as a mythological precedent: a unified semantic field Φ enabled a collective vector v toward heaven, but divine intervention scattered Φ into multiple languages, injecting entropy S and restoring resilience. Modern entropy hacks—student-designed ciphers, Arabic-script reform, retro hardware revival—extend this principle, fragmenting corpora and introducing vector drag to prevent over-coherence [Hanson, 2024].

5 Normative Implications

RSVP extends beyond critique, grounding thermodynamic ethics, governance by entropic budgeting, and a strategy of re-bubbling: creating new attractors of meaning and cooperation rather than resisting collapse. Against MIRI's one-shot framing [Yudkowsky and Soares, 2025], RSVP offers multi-shot adaptation, emphasizing diversity and recursive updates across Φ, v, and S fields.

6 Conclusion

Divided we stand. The debate around AI safety, as crystallized in Yudkowsky and Soares's *If Anyone Builds It, Everyone Dies* [Yudkowsky and Soares, 2025], reveals the costs of excessive coherence. Scott Alexander's critique shows the political impracticality of MIRI's radical clarity: hard bans and absolute prohibitions cannot attract broad coalitions [Alexander, 2025]. Clara Collier's review identifies the epistemic cost: the refusal to update assumptions

in light of new evidence renders the argument brittle and regressive [Collier, 2025]. Robin Hanson's warning about cultural drift highlights the social cost: over-coherence erodes resilience, leaving civilizations vulnerable to unforeseen shocks [Hanson, 2024].

Taken together, these critiques form a triangulation: political impracticality, epistemic rigidity, and cultural fragility. Each identifies a dimension of collapse that arises when entropy is excluded. RSVP responds by formalizing entropy as the substrate of survival. Its MERGE operator encodes continual updating [Lurie, 2009]; its entropy-production functional $\Sigma[\Gamma]$ enforces responsiveness [Verlinde, 2011]; and its socio-political framing treats governance as entropic budgeting rather than prohibition [Landauer, 1961]. Where MIRI insists that humanity's only hope is a single shot at perfect alignment [Yudkowsky and Soares, 2025], RSVP offers a different vision: survival depends on diversity, dispersion, and recursive adaptability. Against monoculture, RSVP prescribes re-bubbling; against rigidity, responsiveness; against brittle singular clarity, entropic resilience. Physics, cognition, and society converge on the same principle: coherence without entropy is collapse.

References

Scott Alexander. Book review: If Anyone Builds It, Everyone Dies. *Astral Codex Ten*, September 2025. URL `https://astralcodexten.substack.com/p/book-review-if-anyone-builds-it-everyone`. Review of Yudkowsky & Soares, If Anyone Builds It, Everyone Dies.

Jacob A. Barandes. Unistochastic quantum theory. *Foundations of Physics*, 52(6):94, 2022. doi: 10.1007/s10701-022-00623-4.

Clara Collier. More was possible: A review of If Anyone Builds It, Everyone Dies. *Asterisk Magazine*, September 2025. URL `https://asteriskmag.com/article/more-was-possible-review-iabied`. Review of Yudkowsky & Soares, If Anyone Builds It, Everyone Dies.

Karl Friston. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11:127–138, 2010. doi: 10.1038/nrn2787.

Robin Hanson. Beware cultural drift. YouTube, SciFuture channel, May 2024. URL `https://www.youtube.com/watch?v=36`. Talk on the risks of cultural monoculture and societal drift.

Ted Jacobson. Thermodynamics of spacetime: The einstein equation of state. *Physical Review Letters*, 75:1260–1263, 1995. doi: 10.1103/PhysRevLett.75.1260.

Rolf Landauer. Irreversibility and heat generation in the computing process. *IBM Journal of Research and Development*, 5(3):183–191, 1961. doi: 10.1147/rd.53.0183.

Jacob Lurie. *Higher Topos Theory*. Princeton University Press, Princeton, NJ, 2009. doi: 10.1515/9781400830558.

Harry Martinson. *Aniara: An Epic Science Fiction Poem*. Bonniers, 1956. English translation by Hugh MacDiarmid and Elspeth Harley Schubert, 1963.

The Hebrew Bible. Genesis 11:1–9, the tower of babel. In *The Old Testament*. Various. Standard reference to the Babel story.

Erik Verlinde. On the origin of gravity and the laws of newton. *Journal of High Energy Physics*, 2011(4):29, 2011. doi: 10.1007/JHEP04(2011)029.

Eliezer Yudkowsky and Nate Soares. *If Anyone Builds It, Everyone Dies*. Machine Intelligence Research Institute, Berkeley, CA, 2025. Published September 2025.