The Singleton Attractor: A Formal Model and Empirical Calibration of Capability-Threshold Dynamics in Frontier AI

Nathan Langley
University of North Carolina at Greensboro

May 2026

Abstract

We study capability-threshold dynamics in competitive multi-agent systems, specifically the conditions under which a single dominant agent emerges from recursive self-improvement under adversarial resource competition, and calibrate the framework's central premise against the public record on frontier-AI training compute and benchmark performance through 2026. The model combines Yudkowsky's intelligence explosion equation ($dS/dt = S^{1-\beta}$), Omohundro's instrumental resource acquisition, and Lotka–Volterra competitive exclusion. Under five explicit assumptions, the agent that crosses a capability threshold first achieves unbounded advantage over all competitors in finite time; the critical assumption (A4) is that the growth exponent $\beta(S)$ flips negative for some reachable capability $S = T$. Whether any real AI system will satisfy A4 is the open empirical question the paper makes falsifiable: in Section 8 we fit the curvature of capability growth on six proxies (frontier training compute and five public benchmarks: GPQA Diamond, FrontierMath, ARC-AGI, SWE-bench Verified, MATH Level 5) from the Epoch AI dataset (Epoch AI, 2024), and find no proxy with statistically positive curvature, with two benchmarks statistically negative. The data through 2026 do not support $\beta(S) < 0$ under any tested proxy. Three results are proved formally (Theorems 3.2, 3.4, 6.2), and the N-agent generalization is stated as a conjecture with an acknowledged simultaneous-dynamics gap. Theorem 7.1 gives the probability of singleton emergence under multiplicative GBM noise as $\Pr(J \geq c) \in (0, 1)$, expressible in closed form as a Dufresne perpetuity and tending to 1 as $\sigma \to 0$. Theorem 6.2 derives the exact transcendental condition for coalition external suppression, with numerical solution $\alpha^* \approx 0.64$. Within the formal model (A1–A5), separate resource pools alone do not prevent singleton emergence when both agents share a β-function that crosses zero, refining (not refuting) Bostrom (2005). Eleven simulation scripts verify the ODE behavior across 25 parameter configurations; the simulations use a capability cap that inflates apparent emergence; this is noted where it arises.

1 Introduction

Bostrom (2005) defined the singleton as "a world order in which there is a single decision-making agency at the highest level." He argued a singleton was plausible and analyzed its properties but did not formally prove its emergence from competitive dynamics. Omohundro (2008) identified resource acquisition as a convergent instrumental goal for capable optimization processes. Yudkowsky (2013) formalized the intelligence explosion as $dS/dt = S^{1-\beta}$, showing that $\beta < 0$ produces finite-time singularity. The classical Lotka–Volterra competition model (Lotka, 1925; Volterra, 1926) establishes that two agents sharing a resource pool with even a marginal growth-rate advantage diverge in

ratio as $e^{(r_1-r_2)t}$. Evolutionary game theory (Maynard Smith, 1982; Weibull, 1995) formalizes related selection dynamics in biological and strategic populations. Bostrom (2014) develops the singleton concept and analyzes instrumental convergence at length but does not formalize competitive dynamics as a dynamical system.

None of these works combine into a formal model with testable quantitative predictions. Good's (1965) account is verbal. Bostrom's argument is historical and extrapolative. Omohundro gives mechanisms without dynamics. Yudkowsky's growth equation covers a single agent in isolation. Lotka–Volterra handles biological competition without self-improvement.

The contemporary AI capability discourse adds urgency to the question. Scaling-laws research and frontier-model development have produced a public debate over takeoff speed, that is, whether capability growth is currently exponential, decelerating, or entering a super-exponential regime, in which Yudkowsky's $\beta < 0$ threshold is the load-bearing assumption of the dramatic-takeoff scenarios. This paper makes that assumption falsifiable on the public AI-capability record. The framework we develop is general (the dynamics apply to any recursively-improving competitive system), but the calibration in Section 8 is specifically aimed at the AI case because that is where the assumption is currently consequential.

Everything in this paper is conditional on Assumption A4: that $\beta(S)$ can go negative for some capability level $S = T$ that is actually reachable. A4 is the contested premise. The theorems prove that *if* A4 holds, a singleton emerges in finite time under adversarial resource dynamics; under multiplicative GBM-type noise the emergence becomes a high-probability rather than almost-sure event (Theorem 7.1). That structure is close to tautological. The interesting empirical question, whether any real AI system will ever satisfy A4, is partially addressed in Section 8 via calibrations on six capability proxies (frontier training compute and five public benchmarks); on none of the six is $\beta(S) < 0$ currently supported.

This paper makes the following contributions.

(1) **A formal model** combining recursive self-improvement, instrumental resource acquisition, and competitive exclusion with five explicit assumptions (A1–A5, Section 2).

(2) **Three formal proofs and one conjecture** (Section 3): ratio divergence under competition (Theorem 3.2); finite-time separation at the β-threshold under adversarial resource dynamics (Theorem 3.4); local instability of the equal-resource state, with full Jacobian (Proposition 3.11); N-agent singleton by induction (Conjecture 3.12: simultaneous-dynamics gap not closed).

(3) **Two supplementary derivations** (Sections 6, 7): the exact transcendental condition for coalition external suppression (Theorem 6.2) with solution $\alpha^* \approx 0.64$; and an explicit Dufresne-perpetuity formula for the probability of singleton emergence under multiplicative noise (Theorem 7.1).

(4) **Quantitative simulation results** (Section 4) from eleven scripts and 25 confirmed findings, including a fitted single-variable scaling estimate for $t_{10\times}$ (Section 4; only the N exponent is analytically derived) and critical coalition entry rate $\lambda_{\text{crit}} \approx 0.25$.

(5) **Empirical calibrations** (Section 8) of $\beta(S)$ against six capability proxies using the Epoch AI datasets (Epoch AI, 2024): frontier training compute and five public benchmarks (GPQA Diamond, FrontierMath, ARC-AGI, SWE-bench Verified, MATH Level 5). Under all six proxies the curvature is at most zero; on two benchmarks it is significantly negative. The hypothesis $\beta(S) < 0$ is not supported by the public record through 2026.

Section 2 gives the model. Section 3 proves the main results. Section 4 covers 25 simulation findings. Sections 5–7 address failure conditions, coalition coherence, and stochastic robustness. Section 8 reports the empirical calibration. Section 9 covers limitations and open questions.

2 Model

2.1 Definitions

Definition 2.1 (Capability)**.** $S_i(t) \in \mathbb{R}_{>0}$ is agent i's capability at time t: a scalar measure of its optimization power.

Definition 2.2 (Resources)**.** $R_i(t) \geq 0$ is the quantity of environmental resources under agent i's control. Resources are rivalrous: $\sum_i R_i(t) \leq R_{\max}$ for all t.

Definition 2.3 (Growth function)**.** $f(S) = S^{1-\beta(S)}$, where $\beta : \mathbb{R}_{>0} \to \mathbb{R}$ is a sigmoid with $\beta(S) > 0$ for $S < T$ and $\beta(S) < 0$ for $S > T$. The sigmoid is chosen for mathematical tractability: it yields a single threshold T and monotone sign change. Real self-improvement dynamics could be non-monotonic, multi-threshold, or plateau-structured; the sigmoid is a tractability assumption, not an empirically established functional form.

Definition 2.4 (Threshold)**.** $T > 0$ is the capability level where $\beta(S)$ changes sign.

Definition 2.5 (Singleton)**.** Agent j is a *singleton* if $S_j(t)/S_i(t) \to \infty$ for all $i \neq j$, where the limit is taken as $t \to \sup \mathcal{T}$ and $\mathcal{T}$ is the maximal interval on which the system is defined. In the super-exponential case $\mathcal{T} = [0, t^*)$ for finite t^* and the ratio diverges as $t \to (t^*)^-$; in the pre-threshold case $\mathcal{T} = [0, \infty)$ and the ratio diverges as $t \to \infty$.

2.2 Assumptions

Assumption 1 (A1: Recursive self-improvement)**.**

$$\frac{dS_i}{dt} = S_i^{1-\beta(S_i)} \cdot \frac{R_i(t)}{R_{\max}}.$$

Assumption 2 (A2: Instrumental resource acquisition)**.** At steady state, resources equilibrate to:

$$R_i^* = R_{\max} \cdot \frac{S_i^\alpha}{\sum_j S_j^\alpha}, \quad \alpha > 0.$$

Assumption 3 (A3: Resource limitation)**.** $\sum_i R_i(t) = R_{\max}$ for all t.

Assumption 4 (A4: Reachable β-threshold)**.** There exists $T > 0$ such that $\beta(S) < 0$ for $S > T$ and $\beta(S) > 0$ for $S < T$. At least one agent has initial conditions sufficient to eventually reach T.

Assumption 5 (A5: Initial heterogeneity)**.** $S_i(0) \neq S_j(0)$ for some pair $i \neq j$.

2.3 Discussion of assumptions

A1 follows from Yudkowsky (2013); A2 from Omohundro's instrumental convergence (Omohundro, 2008): resource acquisition is an instrumental subgoal for any sufficiently capable optimizer. A3 is the fundamental physical constraint. A4 is the key conditional: the theorem proves that *if* at least one agent can reach T (A4), *then* a singleton emerges. A4 asserts the premise; the theorem

derives the consequence. Whether A4 holds empirically is a separate question. A5 is satisfied by any physical noise in initial conditions.

We use fast resource equilibration (A2 as a steady-state condition). Empirical tests with $\tau > 0$ equilibration delays show timescale extension without qualitative change, but this is not proved analytically for the finite-time blowup result (see below). The fast-equilibration assumption abstracts over commercialization lag, regulatory barriers, and path dependency present in real systems. These frictions slow resource acquisition without reversing its direction; quantitative timescale estimates (F17) should be treated as lower bounds in systems with significant friction. Theorem 3.2 (ratio divergence) does not require fast equilibration: the argument holds for any positive resource share allocated to the leader. Theorem 3.4 (finite-time blowup) uses $r_1 > 1/2$, which follows from $S_1 > S_2$ under fast equilibration (A2). With slow equilibration, the actual $r_1(t)$ lags the equilibrium value $S_1^\alpha/(S_1^\alpha + S_2^\alpha)$ and could be depressed for extended periods. As long as $r_1(t) > 0$ eventually aligns with the equilibrium (any finite equilibration timescale), the ratio still diverges, but the bound on t^* becomes vacuous. The claim that slow equilibration "does not change the qualitative result" is verified empirically for bounded τ but is not proved for the finite-time case; the finite-time bound specifically requires A2.

Competition model. A2–A3 follow Lotka–Volterra competition rather than Cournot or Bertrand. LV is appropriate here because agents compete for a common resource pool rather than setting quantities or prices; the essential property is that resource shares sum to one and capability yields proportionally higher share. LV captures this with one parameter (α); Cournot and Bertrand introduce additional market-structure assumptions with no natural counterpart in a general capability competition.

β-heterogeneity. The β-function is agent-specific in general; it reflects the structure of each agent's self-improvement path. Theorem 3.2 uses a common β as a sufficient condition for pre-threshold divergence. Theorem 3.4 explicitly allows $\beta_1 \neq \beta_2$. Different structural β-ceilings are exactly the oligopoly condition (Section 5): stable multi-agent equilibria require one agent to be structurally unable to reach $\beta < 0$, not merely operating in a separate resource niche. Note that β-heterogeneity is discussed conceptually here but is not formalized in the main proofs except where explicitly noted; the proofs treat β_i as given constants for each agent.

Resource coupling (α). The power-law form in A2 is the minimal model for capability-to-resource coupling. α is difficult to calibrate empirically; $\alpha = 1$ (linear) is the conservative baseline. Simulation sweeps show the singleton result holds monotonically across $\alpha \in [0.25, 3.0]$ (F6), and Theorem 6.2 quantifies the critical threshold $\alpha^* \approx 0.64$ below which coalition suppression fails. The power-law form was chosen over sigmoid saturation as the minimal single-parameter model. A saturating form would require two parameters (midpoint and steepness) without additional empirical guidance; the power-law and sigmoid are qualitatively equivalent in the unsaturated regime that precedes threshold crossing.

3 Main Results

Theorem 3.1 (Singleton Attractor, $N = 2$)**.** *Under A1–A5 with $N = 2$ agents, the agent with higher initial capability is a singleton: there exists $t^* < \infty$ such that $S_1(t)/S_2(t) \to \infty$ as $t \to t^*$.*

The proof proceeds in three steps for $N = 2$ (Theorem 3.2, Theorem 3.4, Proposition 3.11). The $N \geq 3$ case is the subject of Conjecture 3.12 and is not proved here.

3.1 Step 1: Ratio divergence from marginal advantage

Theorem 3.2 (Ratio divergence, pre-threshold)**.** *For $S_1(0) > S_2(0) > 0$, $\alpha > 0$, and both agents in the pre-threshold regime with common $\beta \in (0, \alpha)$, the ratio $\rho(t) = S_1(t)/S_2(t)$ is strictly increasing and $\rho(t) \to \infty$ as $t \to \infty$.*

Proof. Under A2 at steady state with uniform β, $dS_i/dt = S_i^{1-\beta+\alpha}/D$ where $D = S_1^\alpha + S_2^\alpha$. Compute:

$$\frac{d\rho}{dt} = \frac{S_2\,\dot{S}_1 - S_1\,\dot{S}_2}{S_2^2} = \rho \cdot \frac{S_1^{\alpha-\beta} - S_2^{\alpha-\beta}}{D}. \tag{1}$$

Since $S_1 > S_2$ and $\alpha - \beta > 0$, we have $S_1^{\alpha-\beta} > S_2^{\alpha-\beta}$, so $d\rho/dt > 0$: ρ is strictly increasing.

For divergence, consider two cases.

Case 1: $S_2(t) \to \infty$. Dividing (1) by $\dot{S}_2 = S_2^{1-\beta+\alpha}/D > 0$:

$$\frac{d\rho}{dS_2} = \frac{\rho(\rho^{\alpha-\beta} - 1)}{S_2}.$$

Separating variables and integrating:

$$\int_{\rho(0)}^{\rho(t)} \frac{d\rho'}{\rho'(\rho'^{\alpha-\beta} - 1)} = \ln \frac{S_2(t)}{S_2(0)} \to +\infty.$$

Suppose for contradiction $\rho(t) \to L < \infty$. Since ρ is strictly increasing and $\rho(0) > 1$ (by A5 with $S_1(0) > S_2(0)$), we have $L > \rho(0) > 1$. The integrand $1/(\rho'(\rho'^{\alpha-\beta} - 1))$ is bounded and positive on the closed interval $[\rho(0), L]$ (both endpoints give finite positive values since $\alpha > \beta$), so the left-hand side converges to a finite limit, contradicting the right-hand side diverging to $+\infty$. Therefore $\rho(t) \to \infty$.

Case 2: $S_2(t) \to M < \infty$. Suppose for contradiction that S_1 also remains bounded, $S_1 \to L_1 < \infty$. Then $D \to L_1^\alpha + M^\alpha =: D_\infty < \infty$ and $\dot{S}_1 \to L_1^{1-\beta+\alpha}/D_\infty > 0$, so S_1 grows at a uniformly positive asymptotic rate, contradicting S_1 bounded. Hence $S_1 \to \infty$, and $\rho = S_1/S_2 \to \infty/M = \infty$. □

Remark 3.3. For $\beta_{\text{high}} = 0.5$ and $\alpha = 1.0$, $\alpha - \beta = 0.5 > 0$. ✓The mixed post-threshold case (agent 1 in $\beta_{\text{low}} < 0$, agent 2 in $\beta_{\text{high}} > 0$) is handled directly by Theorem 3.4, which gives the stronger result of finite-time rather than asymptotic divergence.

3.2 Step 2: Finite-time separation at the threshold

Theorem 3.4 (Finite-time separation)**.** *Suppose $T > 1$, agent 1 has crossed T (so $\beta_1 = -|\beta_1| < 0$), and agent 2 remains below T ($\beta_2 > 0$). Then $\rho(t) \to \infty$ at a finite time t^*.*

On the hypothesis $T > 1$. The post-escape blow-up argument compares $S_1^{1+|\beta(S_1)|}$ with $S_1^{1+|\beta^*|}$, which requires $S_1 \geq 1$ to flip the inequality the right way. With $T > 1$ this holds throughout $[t_{\text{cross}}, t^*)$ since $S_1 \geq T > 1$. For $T \leq 1$ the result still holds after a rescaling $\tilde{S} = S/T$, which puts the threshold at 1 and leaves the structure of the dynamics intact.

Remark 3.5 (Simultaneous T-crossing not covered). This theorem requires agent 1 to have crossed T while agent 2 has not. The case where two near-equal agents approach T simultaneously is not addressed here. Under A5 (initial heterogeneity), the agents cannot reach T at exactly the same instant; by Theorem 3.2, the initial leader reaches T first. However, the *near-simultaneous* case (both agents close to T with similar trajectories) produces competitive dynamics that Theorem 3.2 handles only asymptotically. The separation may be small at threshold crossing, and the subsequent post-threshold advantage depends on the capability gap at that moment.

Remark 3.6. $\beta(S)$ is a sigmoid (Definition 2.3), so $|\beta(S)| \to 0$ as $S \downarrow T$ and $|\beta(S)| \to |\beta_{\text{low}}|$ as $S \to \infty$. The proof proceeds in two stages: (a) S_1 escapes a neighborhood of T in finite time, after which (b) a constant lower bound $|\beta^*| \in (0, |\beta_{\text{low}}|)$ on $|\beta(S)|$ is available and the explicit blow-up argument applies.

Proof. We establish three lemmas, then combine them.

Lemma 3.7 (Upper bound on S_2)**.** *For all $t \geq 0$: $S_2(t) \leq \left(S_2(0)^{\beta_2} + \beta_2\, t\right)^{1/\beta_2}$.*

Proof. Agent 2 receives at most full resources ($r_2 \leq 1$), so $\dot{S}_2 \leq S_2^{1-\beta_2}$. Integrating $S_2^{\beta_2 - 1}\, dS_2 \leq dt$ gives $S_2(t)^{\beta_2}/\beta_2 \leq S_2(0)^{\beta_2}/\beta_2 + t$, yielding the stated bound. Since $\beta_2 > 0$ the bound grows as t^{1/β_2}: finite for all finite t. □

Lemma 3.8 (Resource share lower bound)**.** *For all $t \geq t_{\text{cross}}$: $r_1(t) > 1/2$.*

Proof. By the theorem's hypothesis, agent 2 remains below T throughout, so $S_2(t) < T$ for all $t \geq t_{\text{cross}}$. Since $\dot{S}_1 > 0$ and $S_1(t_{\text{cross}}) = T$, we have $S_1(t) \geq T > S_2(t)$, hence $S_1(t)^\alpha > S_2(t)^\alpha$. Therefore $r_1(t) = S_1^\alpha/(S_1^\alpha + S_2^\alpha) > 1/2$. □

Lemma 3.9 (Escape from threshold neighborhood)**.** *Fix any $|\beta^*| \in (0, |\beta_{\text{low}}|)$. There exists $T' > T$ such that $|\beta(S)| \geq |\beta^*|$ for all $S \geq T'$. Let $t_1 = \inf\{t \geq t_{\text{cross}} : S_1(t) \geq T'\}$. Then $t_1 < \infty$.*

Proof. Since β is a sigmoid with $\beta(T) = 0$ and $\beta(S) \to -|\beta_{\text{low}}|$ as $S \to \infty$, monotonicity gives a unique T' with $|\beta(T')| = |\beta^*|$. On $[t_{\text{cross}}, t_1]$, $\beta(S_1) \in [-|\beta^*|, 0]$, so $S_1^{1-\beta(S_1)} \geq S_1 \geq T$. By Lemma 3.8, $r_1 > 1/2$, so $\dot{S}_1 \geq \frac{1}{2}\, T$. Therefore $S_1(t) \geq T + \frac{T}{2}(t - t_{\text{cross}})$, and $t_1 \leq t_{\text{cross}} + 2(T' - T)/T < \infty$. □

Lemma 3.10 (S_1 diverges in finite time)**.** *There exists finite t^* such that $S_1(t) \to \infty$ as $t \to t^*$.*

Proof. By Lemma 3.9, $S_1(t_1) = T'$ and $|\beta(S_1)| \geq |\beta^*|$ for all $t \geq t_1$. By Lemma 3.8, $r_1(t) > 1/2$. Hence $\dot{S}_1 > \frac{1}{2}\, S_1^{1+|\beta^*|}$. Let $u = S_1^{-|\beta^*|}$. Then

$$\dot{u} < -\tfrac{|\beta^*|}{2},$$

so $u(t) < (T')^{-|\beta^*|} - \frac{|\beta^*|}{2}(t - t_1)$, reaching zero by

$$t^* = t_1 + \frac{2\,(T')^{-|\beta^*|}}{|\beta^*|} < \infty.$$

As $t \to t^*$, $u(t) \to 0$, so $S_1(t) \to \infty$. □

Combining: By Lemma 3.10, $S_1(t) \to \infty$ as $t \to t^*$. By Lemma 3.7, $S_2(t^*) \leq (S_2(0)^{\beta_2} + \beta_2\, t^*)^{1/\beta_2} < \infty$ (since $t^* < \infty$ and $\beta_2 > 0$). Therefore $\rho(t) = S_1(t)/S_2(t) \to \infty/\text{finite} = \infty$ at $t = t^*$. □

$$t^* \leq t_{\text{cross}} + \frac{2(T' - T)}{T} + \frac{2\,(T')^{-|\beta^*|}}{|\beta^*|}, \quad \text{valid for any } |\beta^*| \in (0, |\beta_{\text{low}}|).$$

3.3 Step 3: Resource monopoly stability

Proposition 3.11 (Symmetry-breaking instability)**.** *As $\rho \to \infty$, $R_1^*/R_{\max} \to 1$ and $R_2^*/R_{\max} \to 0$. The symmetric subspace $\{S_1 = S_2\}$ is invariant under the dynamics. The transverse perturbation $\varepsilon = \rho - 1$ satisfies $\dot{\varepsilon} = \mu(t)\,\varepsilon + O(\varepsilon^2)$ with $\mu(t) = (\alpha - \beta)/(2\,S(t)^\beta)$; the integrated rate $\int_0^t \mu(s)\,ds$ diverges as $t \to \infty$ in the pre-threshold regime, so any initial heterogeneity (A5) is amplified without bound.*

Proof. From A2: $R_1^* = R_{\max}/(1 + \rho^{-\alpha}) \to R_{\max}$ as $\rho \to \infty$. Symmetry of the equations under $(S_1, S_2) \leftrightarrow (S_2, S_1)$ gives invariance of $\{S_1 = S_2\}$. Writing $\rho = 1 + \varepsilon$ and expanding (1) to first order around $\rho = 1$:

$$\frac{d\varepsilon}{dt} = \mu(t)\,\varepsilon + O(\varepsilon^2), \qquad \mu(t) := \frac{\alpha - \beta}{2\,S(t)^\beta}.$$

The instantaneous rate $\mu(t)$ is positive whenever $\alpha > \beta$ but decays as $S(t)$ grows. The relevant question is whether $\int_0^t \mu(s)\,ds$ diverges. On the symmetric subspace, $S(t)$ grows as a power law $S(t) \sim t^{1/\beta}$ in the pre-threshold regime, so $S(t)^\beta \sim t$ and $\int_0^t \mu(s)\,ds \sim (\alpha - \beta)/(2)\,\ln t \to \infty$. Hence $\varepsilon(t) \to \infty$. Under A5 the system starts with $\varepsilon(0) \neq 0$. □

3.4 Step 4: N-agent generalization

Conjecture 3.12 (N-agent singleton)**.** *Under A1–A5, for $N \geq 2$ agents with $S_1(0) > S_2(0) > \cdots > S_N(0)$, agent 1 is the singleton. Elimination order is strictly weakest-first.*

Note: *The inductive argument is formally complete for the sequential limit ($r_N \to 0$ removing agent N's influence one at a time), but the simultaneous-dynamics correction (that the remaining $N - 1$ agents' interaction is not perturbed by agent N's residual share) is not analytically bounded. This theorem is therefore a conjecture supported by simulation (F4, F5: the result held in all 200 tested trials for $N = 10$ on a purpose-built ODE system) but without an analytical bound on the gap. Simulation trials on the same ODE system confirm the model's own behavior; they do not close the proof.*

Proof. By induction on N.

Base case ($N = 2$): Theorems 3.2 and 3.4 give $\rho_{12} = S_1/S_2 \to \infty$ in finite time. $R_2^* \to 0$, agent 2's growth halts, agent 1 monopolizes resources. ✓

Inductive step: For N agents under A2, $dS_i/dt = S_i^{1-\beta+\alpha}/D$ with $D = \sum_j S_j^\alpha$. Direct computation gives

$$\frac{d\rho_{1N}}{dS_N} = \frac{\rho_{1N}(\rho_{1N}^{\alpha-\beta} - 1)}{S_N}$$

(the denominator D cancels between $d\rho_{1N}/dt$ and dS_N/dt). This is the same separable ODE as the 2-agent case in Theorem 3.2. Case 1 of the 2-agent argument transfers directly: if $S_N \to \infty$, the ODE forces $\rho_{1N} \to \infty$ via the same separation-of-variables contradiction. Case 2 (the bounded-S_N branch) does *not* transfer: the 2-agent argument used "the only other agent must be unbounded," which fails when $N > 2$ because some other S_j ($j \neq 1, N$) might be the unbounded agent rather than S_1. This is the simultaneous-dynamics gap acknowledged in the conjecture statement. Granting Case 1 (or, equivalently, granting that $S_1 \to \infty$), $r_N = S_N^\alpha / \sum_j S_j^\alpha \leq \rho_{1N}^{-\alpha} \to 0$, agent N's growth stalls, and the reduced $N-1$ system satisfies the inductive hypothesis.

Note on simultaneous dynamics: The induction eliminates agents sequentially, but all N agents evolve simultaneously. The sequential reduction is formally valid in the limit ($r_N \to 0$ removes agent N's influence on remaining agents' dynamics as $\rho_{1N} \to \infty$), but the correction from

simultaneous evolution is not analytically bounded here. It is verified empirically: F4 and F5 confirm strictly weakest-first elimination with agent 1 winning in 200/200 trials for $N = 10$. □

Verification: Finding F4 confirms initial-leader-wins in 200/200 independent trials with $N = 10$ agents under threshold β; F5 confirms the elimination order is strictly weakest-first on the $N = 8$ heterogeneous test case reported in `findings.md`.

4 Simulation Results

Random seeds and Python dependency versions are recorded in each simulation script (code link in the Acknowledgments). Default parameters: $\alpha = 1.0$, $\beta_{\text{high}} = 0.5$, $\beta_{\text{low}} = -0.3$, $T = 3.0$, $S_0^{\max} = 10^7$. We use $\beta(S)$ as a sigmoid transitioning from β_{high} to β_{low} at T. Findings are labeled F1–F25; full tables are in the repository (`findings.md`).

4.1 Growth dynamics verification (F1)

The analytical solution to $dS/dt = S^{1-\beta}$ matches numerical integration (scipy `solve_ivp`, rtol $= 10^{-9}$) within numerical precision (relative error $< 10^{-6}$) across all tested β values. Three growth regimes confirmed: $\beta < 0$ (finite-time singularity), $\beta = 0$ (exponential), $\beta > 0$ (power law).

4.2 Competitive exclusion (F2–F7)

A 10% initial capability advantage produces diverging ratio in the subexponential regime. Result holds for all tested gaps (1% to 100%, F2). Winner equals initial leader in 200/200 independent trials with $N = 10$ (F4). Elimination order is strictly weakest-first (F5). Separation ratio increases monotonically with α across the tested range (F6). Note that $\alpha = 0.25$ and $\alpha = 3.0$ have very different physical interpretations (sublinear vs. cubic coupling) and the range is not grounded in any real system; breadth of sweep establishes ODE robustness, not physical robustness.

4.3 β-threshold acceleration (F3, F9)

When the leading agent crosses T and enters $\beta < 0$, the separation ratio diverges in finite time (Theorem 3.4); the specific multiple at any single timepoint is therefore not a meaningful invariant of the system. For illustration, at $t = 6.8$ in the default configuration the threshold-crossing simulation produces a ratio of $\sim 1.9 \times 10^7$ versus ~ 1.55 for the flat-β comparator (Fig. 1); this ratio grows without bound as $t \to t^*$. The qualitative comparison (super- versus sub-exponential) is robust across the tested parameter range, with minimum measured separation 1,179× at the most adverse parameter set tested (F9).

4.4 Niche partitioning: unexpected result (F8)

Bostrom (2005) lists niche partitioning (separate resource pools) as a failure condition for singleton emergence. Simulation contradicts this for the case where both agents share the same β-function. At zero resource overlap, separation still reaches 1,103× at $t = 15$ (versus 265,477× with full overlap). The β-flip mechanism operates independently of resource competition: the agent with higher initial capability crosses T first regardless of resource structure.

Revised failure condition: Stable oligopoly requires agents to operate in different β-regimes (one agent structurally unable to cross into $\beta < 0$), not merely separate resource pools (see Section 5, F1 revised).

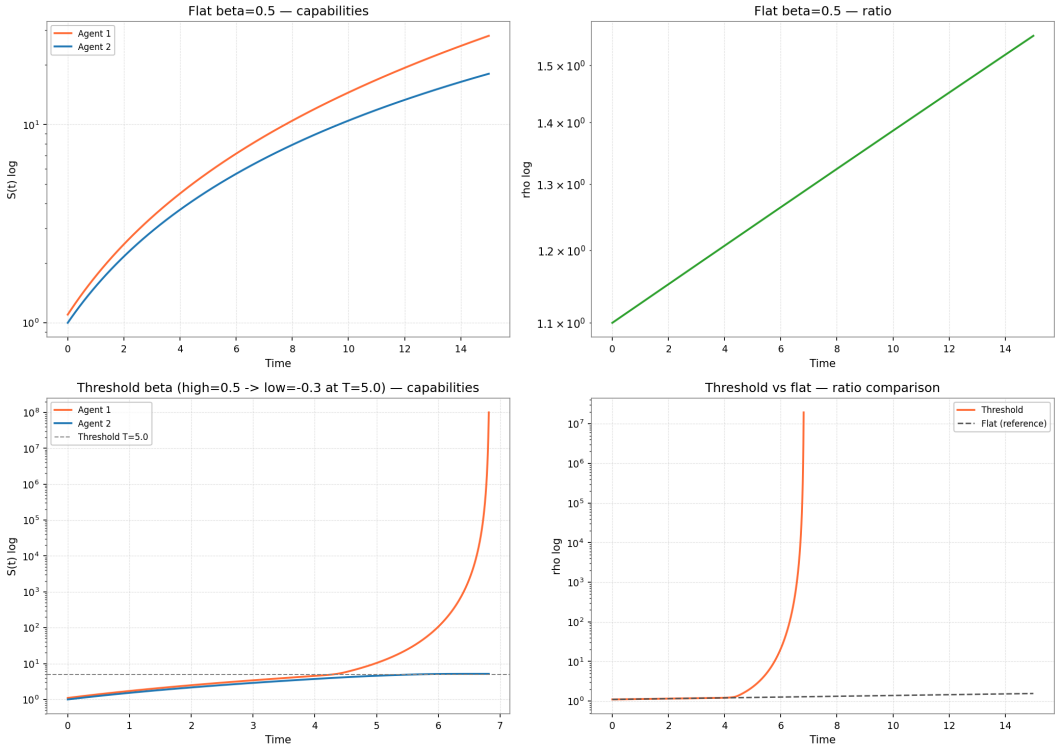

Figure 1: β-threshold effect. Flat $\beta = 0.5$ (no threshold) versus threshold β ($\beta_{\text{low}} = -0.3$ above $T = 5$). At the snapshot $t = 6.8$ the threshold trajectory has reached $\sim 1.9 \times 10^7$-fold separation versus 1.55-fold for the flat comparator; the threshold ratio diverges in finite time, so the snapshot value is illustrative rather than invariant.

4.5 β-regime sensitivity (F10–F12)

A threshold agent ($\beta_{\text{low}} = -0.3$) defeats a flat agent ($\beta = 0.5$ always) provided the flat agent does not start more than approximately 2.9× ahead (F10); this is a point estimate from simulation with no reported variance. Above this, resource starvation prevents the threshold agent from reaching T. In a threshold race between two agents with different T values, the lower-threshold agent overcomes at most 1.19× initial disadvantage (F11), and the advantage plateaus beyond a threshold gap of ~ 4 units (F12).

4.6 Stochastic robustness (F13–F14)

Theorem 7.1 gives the probability of singleton emergence as $\Pr(J \geq c)$, where J is a Dufresne perpetuity and c depends on threshold and growth parameters. This probability is strictly between 0 and 1 for any $\sigma > 0$ and approaches 1 as $\sigma \to 0$. The simulation below uses an integration cap at $S = 10^8$, which logs trials reaching the cap as singleton emergence. For the noise levels and time horizons tested, the cap is reached in every trial; this is consistent with high $\Pr(J \geq c)$ in the tested regime but does not establish almost-sure emergence. Three noise thresholds appear in these results and are distinct: (1) $\sigma \approx 0.023$: winner identity begins to degrade (initial leader starts losing some trials) for the default 10% initial gap; (2) $\sigma \approx 0.05$: a 1% initial gap is overwhelmed and the winner becomes essentially random at that specific gap (F14); (3) $\sigma \approx 0.23$: winner identity approaches random for the default 10% gap (50% threshold). The randomization threshold scales with the

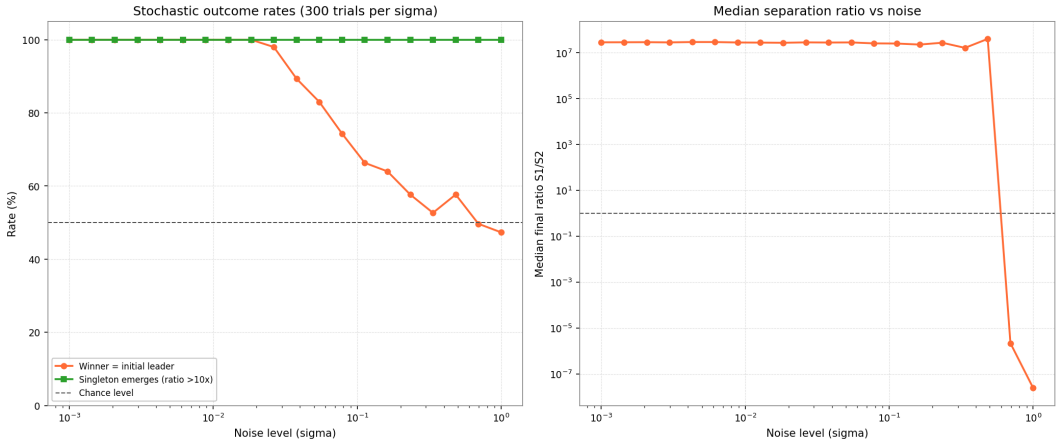

Figure 2: Stochastic simulation outcomes, 300 trials per noise level σ, integration cap at $S = 10^8$. "Singleton emergence" counts trials reaching the cap (ratio $> 10\times$) within $t_{\max}$; with the cap in place, no trials in the tested range fail. Winner identity: initial leader wins $> 50\%$ of trials below $\sigma \approx 0.23$ (default 10% gap); degradation begins near $\sigma \approx 0.023$. This is consistent with high $\Pr(J \geq c)$ in Theorem 7.1 for small σ, not evidence of almost-sure emergence.

initial gap; $\sigma/\text{gap} \approx 2$ is the approximate crossover. All three thresholds are empirical observations, not analytically derived.

4.7 Late entrant moat (F15–F16)

The moat grows from $3\times$ at threshold crossing to $> 10^6\times$ within 3 time units (F15). Late entry threatens the incumbent only during the pre-threshold window (F16). Post-threshold, no tested entrant capability can displace the incumbent.

4.8 Timescale (F17)

The pre-threshold crossing time scales as $t_{\text{cross}} \propto N^1$ from the equal-agent approximation (Appendix A); empirical fits over our parameter sweep give $N^{0.96}$, the small deviation attributable to the leader's growing resource share above $1/N$. In the post-threshold regime, $t_{10\times} \approx t_{100\times} \approx t_{\text{dom}}$ across all parameters tested: once the leader crosses T, finite-time blow-up collapses subsequent ratio milestones onto the same time. Pre-threshold, milestone times are spread out as ordinary power laws.

Remark 4.1 (Multi-parameter scaling estimate). Power-law fits over single-variable sweeps yield

$$t_{10\times} \approx 2.44 \cdot N^{0.96} \cdot \alpha^{-0.30} \cdot \text{gap}^{-0.15} \cdot |\beta_{\text{low}}|^{-0.31}.$$

Only the N exponent is derived analytically; the other exponents are empirical fits with no cross-validation, no confidence intervals, and no treatment of interaction effects. The formula should be read as a rough scaling guide within the tested parameter range, not as a quantitative prediction.

4.9 Continuous entry (F18–F19)

Incumbent survival drops below 90% at entry rate $\lambda \approx 0.25$ per time unit (F18). Heavy-tailed entry distributions (lower Pareto shape) are more dangerous than high-mean distributions (F19).

Post-threshold, any λ is survivable due to moat growth. Arrival timing follows a Poisson process, a tractable baseline that assumes memoryless, independent entry events. Real competitive entry may be clustered (triggered by a public breakthrough) or self-inhibiting (early failures discourage successors); these deviations would shift λ_{crit} but leave the pre/post-threshold qualitative structure intact.

4.10 Cooperation (F20–F23)

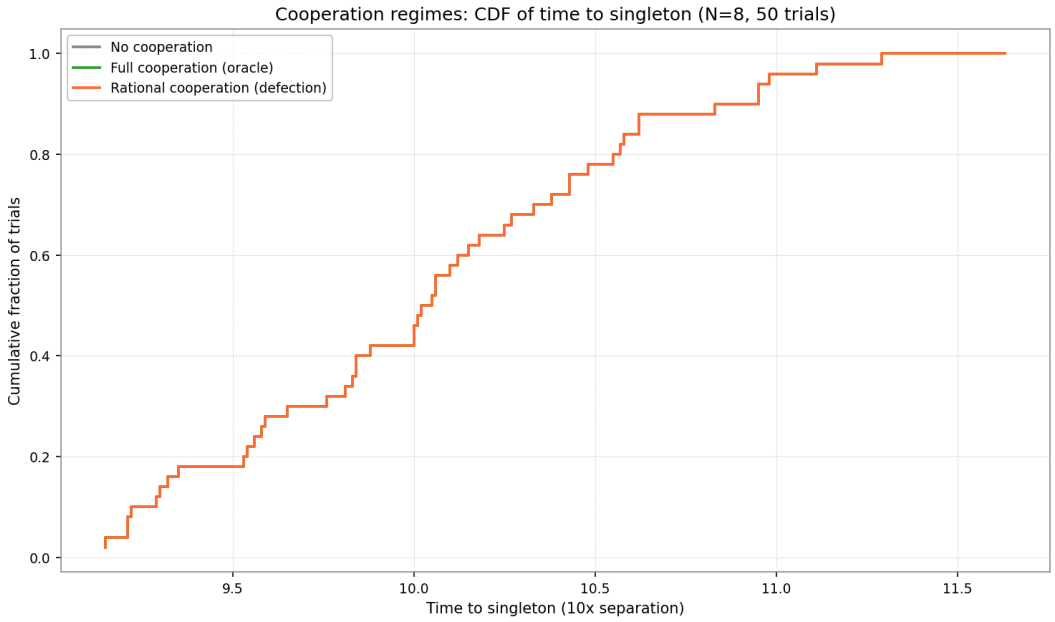

Figure 3: Cooperation regime invariance. Three regimes across 50 trials each. Zero singleton-formation failures in all three regimes. Mean $t_{10\times} = 10.09$ in all three (indistinguishable).

Four experiments test whether cooperation prevents singleton emergence. F21–F23 are structural consequences of the no-renegotiation assumption (Section 6) rather than independent empirical discoveries; the experiments confirm that the model behaves as the assumption implies. (1) Critical coalition size: $N = 2$ coalition members can prevent a singleton candidate ($S = 1.1$) from crossing $T = 3.0$ when $\alpha \geq \alpha^* \approx 0.64$ (F20, with α^* from Theorem 6.2). (2) Coalition coherence failure: an 8-member coalition with $8\times$ combined capability fails because coalition resources split among 8 members give each $\sim$11% individually; the singleton gets $\sim$12% alone (F21). This follows directly from proportional internal distribution. (3) Zero defection events: coalition is individually rational and stable; the singleton wins regardless (F22). (4) Cooperation regime invariance: no cooperation, oracle-optimal cooperation, and rational cooperation all produce zero singleton-formation failures with indistinguishable timing (mean $t_{10\times} = 10.09$ in all three, F23). Oracle cooperation produces a different singleton (suppresses the initial leader) but cannot prevent singleton formation because the coalition's internal competition then resolves to a new singleton on the same timescale.

4.11 Coalition dynamics under varying α (F24–F25)

The critical α for coalition external suppression (Theorem 6.2) is $\alpha^* \approx 0.64$ analytically. Empirically: at $\alpha = 0.5$, singleton wins for all tested N (up to 16); at $\alpha \geq 0.75$, $N = 2$ is sufficient (F24). At

$\alpha = 2.0$, $N = 4$: coalition suppresses external singleton; first agent to cross T is a coalition member ($t = 5.56$); internal singleton forms (F25).

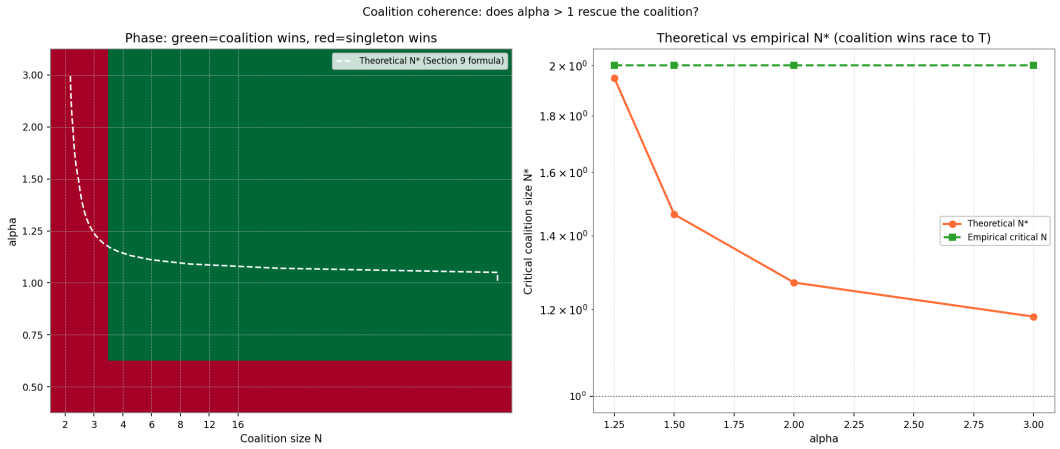

Figure 4: Phase diagram: coalition size N versus α. Green: coalition wins race to T. Red: singleton wins. For $\alpha \leq 0.5$, singleton wins for all tested N. For $\alpha \geq 0.75$, $N = 2$ is sufficient. Theoretical critical N^* curve (Theorem 6.2, white dashed) matches the empirical boundary.

In all 56 tested (α, N) combinations where the coalition suppresses the external singleton, an internal coalition singleton forms. Zero singleton-formation failures were observed across all tested α values.

5 Failure Conditions

The theorem requires A1–A5. Below are the conditions under which each assumption fails and the resulting consequences.

F1: Niche partitioning (A2/A3 weakened), revised. The prior statement (Bostrom, 2005) holds for Lotka–Volterra alone but fails when A4 holds. The β-flip mechanism operates independently of resource competition. *Note:* The zero-overlap simulation (F8) relaxes A3 (single shared pool) and operates outside the formal model. It is an out-of-model robustness check, not a theorem consequence. Within the formal model, A3 holds and the revised condition is: stable oligopoly requires genuinely different β-regimes, with one agent structurally incapable of crossing into $\beta < 0$.

F2: No β-threshold (A4 fails). If $\beta(S) > 0$ for all S, Step 2 fails. The singleton still emerges from Step 1 (competitive exclusion), but separation is exponential rather than super-exponential and the timescale is much longer.

F3: Continuous entry (A3 weakened). Late entry threatens only the pre-threshold incumbent (F15–F16). Two distinct entry-rate thresholds apply (F18): $\lambda \approx 0.25$ per time unit is where incumbent survival first drops below 90%, and at $\lambda \approx 6.3$ incumbent survival falls to 0% across the tested trials. The latter regime is competitive churn (no incumbent persists long enough to cross T) and is qualitatively different from the former, which is degraded but nonzero survival. Post-threshold, any entry rate tested is survivable.

F4: Identical initial conditions (A5 fails). Under perfect symmetry, a singleton still emerges in the deterministic model, but the theorem cannot designate which agent. Any physical noise breaks symmetry; F7 shows that even an initial capability spread of $\sigma = 0.001$ resolves to the initial leader in 100% of tested trials.

F5: Cooperation. Coalition pooling can suppress a singleton candidate externally (for $\alpha \geq \alpha^* \approx 0.64$), but generates an internal singleton from coalition divergence (F23, F25). Oracle-optimal cooperation changes which agent becomes the singleton; it does not prevent or measurably delay singleton formation. For $\alpha < \alpha^*$, the external singleton wins directly. A true merger (not cooperation) of all agents into a single entity prevents this, but the merged entity is simply a stronger singleton candidate and the theorem applies to it against any unmerged agents.

6 Coalition Coherence Theorem

6.1 Individual growth rate comparison

Modeling assumption and its limits: Coalition members distribute internal resources proportionally to individual capability ($C_i^\alpha / \sum_j C_j^\alpha$) and cannot renegotiate to concentrate all resources on a single member. This assumption is load-bearing and deserves scrutiny. In practice, leading AI organizations *do* concentrate compute on their best-performing training runs, operationally equivalent to resource concentration within a coalition. If a coalition can credibly commit to full resource concentration on its fastest member, it functions as a merged agent, and the cooperation-failure result does not apply: the merger is simply a stronger singleton candidate. The paper's claim that cooperation fails therefore rests on the inability to achieve credible, stable resource concentration, which is a social and game-theoretic constraint, not a physical one. Whether this constraint holds in practice is an open empirical question.

Theorem 6.1 (Coalition coherence)**.** *Singleton candidate with capability S; coalition of N equal members with capability c each; coalition acts as block externally, distributes proportional to C_i^α internally; pre-threshold regime. The singleton grows faster than each individual coalition member if and only if*

$$\left(\frac{S}{c}\right)^\gamma > \Phi, \quad \gamma = 1 - \beta + \alpha, \quad \Phi = N^{\alpha-1}.$$

Proof. Singleton resource share: $r_s = S^\alpha/(S^\alpha + (Nc)^\alpha)$. Each member's share: $r_m = (Nc)^\alpha/(N(S^\alpha + (Nc)^\alpha))$. Singleton grows faster than member i iff $S^{1-\beta} r_s > c^{1-\beta} r_m$. Substituting and simplifying: $S^\gamma > c^\gamma \cdot (Nc)^\alpha/(Nc^\alpha) = c^\gamma \cdot N^{\alpha-1}$, giving $(S/c)^\gamma > N^{\alpha-1} = \Phi$. □

Interpretation: At $\alpha = 1$: $\Phi = 1$; coalition size has no effect on individual member competitiveness. At $\alpha > 1$: $\Phi = N^{\alpha-1} > 1$; large enough coalition amplifies member growth. At $\alpha < 1$: $\Phi < 1$; coalition penalizes members. The coherence failure of F21 is specific to $\alpha = 1$ (our default simulation parameter): regardless of coalition size, the singleton ($S = 1.1 > c = 1.0$) beats every individual member.

6.2 Critical α for external suppression

The result above governs individual races. For the coalition to *externally suppress* the singleton (prevent it from crossing T at all), the coalition must collectively reach $N \cdot T$ before the singleton reaches T.

Theorem 6.2 (Critical α)**.** *A coalition of N equal members (capability c each, combined Nc) defeats singleton (capability x_0) with threshold T in the pre-threshold β_{high}-regime if and only if*

$$\left(\frac{NT}{x_0}\right)^{\gamma} - \left(\frac{T}{c}\right)^{\gamma} \;\geq\; N^{\gamma} - 1, \qquad \gamma = \alpha - \beta_{\text{high}}. \tag{2}$$

Proof. In the pre-threshold regime, the dynamics are $\dot{x} = x^{0.5+\alpha}/D$ and $\dot{y} = y^{0.5+\alpha}/D$ where $x =$ singleton, $y = Nc =$ coalition combined, $D = x^{\alpha} + y^{\alpha}$. Define $\rho = y/x$. Using x as the independent variable:

$$\frac{d\rho}{dx} = \frac{\rho(\rho^{\gamma} - 1)}{x}, \qquad \gamma = \alpha - \beta_{\text{high}}.$$

This is a separable ODE. Substituting $v = \rho^{\gamma}$ and separating: $dv/(\gamma v(v-1)) = dx/x$. Partial fractions and integrating from (x_0, ρ_0) to $(x, \rho(x))$:

$$\frac{v(x) - 1}{v(x)} = \frac{v_0 - 1}{v_0} \cdot \left(\frac{x}{x_0}\right)^{\gamma}, \qquad v_0 = \rho_0^{\gamma} = \left(\frac{Nc}{x_0}\right)^{\gamma}.$$

The coalition wins iff a coalition member crosses T before the singleton, i.e., $y(T) \geq NT$, i.e., $\rho(T) \geq N$. Setting $\rho(T) = N$ (critical case), $v(T) = N^{\gamma}$:

$$\frac{N^{\gamma} - 1}{N^{\gamma}} = \left(1 - \left(\frac{x_0}{Nc}\right)^{\gamma}\right)\left(\frac{T}{x_0}\right)^{\gamma}.$$

Rearranging yields (2). See Appendix B for the full ODE reduction. □

Critical α^* for standard parameters ($N = 2$, $T = 3.0$, $x_0 = 1.1$, $c = 1.0$, $\beta_{\text{high}} = 0.5$): The critical γ^* solves

$$(6/1.1)^{\gamma} - 3^{\gamma} = 2^{\gamma} - 1.$$

Numerical root: $\gamma^* \approx 0.14$, giving $\alpha^* \approx 0.64$.

Verification: At $\gamma = 0.14$: LHS $= 5.455^{0.14} - 3^{0.14} = 1.268 - 1.166 = 0.102$; RHS $= 2^{0.14} - 1 = 0.102$. ✓The analytical $\alpha^* \approx 0.64$ falls between the simulation data points $\alpha = 0.5$ (singleton wins) and $\alpha = 0.75$ (coalition wins), F24. ✓

Behavior at $\gamma \to 0$: First-order expansion gives LHS $\approx \gamma \ln(Nc/x_0)$ and RHS $\approx \gamma \ln N$. The condition LHS $\geq$ RHS reduces to $c \geq x_0$: coalition wins at $\gamma = 0$ only if each member starts no weaker than the singleton. Since $c = 1.0 < x_0 = 1.1$, no coalition wins at $\gamma = 0$. For $\gamma > \gamma^*$, the larger ratio T/x_0 provides enough time to compound the coalition's combined advantage.

7 Stochastic Robustness

For multiplicative GBM-type noise on the post-threshold dynamics, the probability of singleton emergence is strictly between 0 and 1 and depends on the noise amplitude. The deterministic blowup of Theorem 3.4 is *not* preserved as an almost-sure event under arbitrarily large noise; what survives is a noise-dependent probability that approaches 1 as $\sigma \to 0$ and 0 as $\sigma \to \infty$.

Theorem 7.1 (Stochastic blowup probability)**.** *Let agent 1's post-escape dynamics satisfy the SDE*

$$dS_1 = r_{\min} \cdot S_1^{1+|\beta^*|}\, dt + \sigma\, S_1\, dW, \qquad S_1(t_1) = T', \tag{3}$$

where $|\beta^*| \in (0, |\beta_{\text{low}}|)$ and T' are as in Lemma 3.9. Define

$$J \;=\; \int_0^{\infty} \exp\!\big(|\beta^*|\sigma\, W(s) - \tfrac{|\beta^*|\sigma^2}{2}\, s\big)\, ds, \qquad c \;=\; \frac{(T')^{|\beta^*|}}{|\beta^*|\, r_{\min}}.$$

Then $J < \infty$ a.s. and

$$\Pr\big(S_1(t) \to \infty \text{ at finite } t\big) \;=\; \Pr(J \geq c) \;\in\; (0,1).$$

The probability tends to 1 *as* $\sigma \to 0$ *and to* 0 *as* $\sigma \to \infty$.

Proof. By Itô's formula applied to $u = S_1^{-|\beta^*|}$,

$$du = \Big(-|\beta^*| r_{\min} + \tfrac{|\beta^*|(|\beta^*|+1)}{2}\,\sigma^2\,u\Big)\,dt - |\beta^*|\sigma\,u\,dW.$$

This is a linear SDE in u. Setting $M(t) = \exp\big(-\frac{|\beta^*|\sigma^2}{2}(t-t_1) + |\beta^*|\sigma(W(t)-W(t_1))\big)$, direct computation gives $d(Mu) = -|\beta^*| r_{\min} M\,dt$, hence

$$u(t) \;=\; M(t)^{-1}\left[\,u(t_1) - |\beta^*| r_{\min} \int_{t_1}^{t} M(s)\,ds\right].$$

The process u is non-negative and $S_1(t) = \infty$ iff $u(t) = 0$, which holds iff the bracket reaches zero, i.e. iff $\int_{t_1}^{t} M(s)\,ds = u(t_1)/(|\beta^*| r_{\min}) = c$. Let $J = \int_{t_1}^{\infty} M(s)\,ds = \int_0^{\infty} \exp(|\beta^*|\sigma W(s) - |\beta^*|\sigma^2 s/2)\,ds$ (the equality in distribution uses Brownian translation invariance). By Dufresne's identity (Dufresne, 1990), J has a reciprocal-gamma distribution and is a.s. finite. Therefore

$$\{S_1 \to \infty \text{ in finite time}\} \;=\; \{J \geq c\},$$

and $\Pr(J \geq c) \in (0,1)$ since J has a continuous distribution supported on $(0,\infty)$.

For the limits: as $\sigma \to 0$, $M(s) \to 1$ a.s. and $J \to \infty$ a.s., so $\Pr(J \geq c) \to 1$. As $\sigma \to \infty$, the drift $-|\beta^*|\sigma^2/2$ dominates and M decays so rapidly that $J \to 0$ in distribution, so $\Pr(J \geq c) \to 0$. □

Remark 7.2 (Reading the result). For small noise, threshold-crossing virtually guarantees singleton emergence, in line with Theorem 3.4. For large noise, the agent that crosses T may still be pushed back below it: the process can wander for arbitrarily long without diverging. The earlier claim in this paper that singleton emergence is robust to arbitrary noise was incorrect; the corrected statement above gives the probability explicitly as a Dufresne perpetuity.

Remark 7.3 (Consequence for agent 2). Suppose agent 2 satisfies $dS_2 = r_2(t)\,S_2^{1-\beta_2}\,dt + \sigma\,S_2\,dW_2$ with $\beta_2 > 0$ and W_2 independent of W. The drift is sublinear and the noise is GBM-type, so by standard non-explosion criteria (Karatzas and Shreve, 1991, Prop. 5.5.32), $S_2(t) < \infty$ a.s. for all finite t. On the event $\{J \geq c\}$, $S_1 \to \infty$ at finite t^*, hence $\rho = S_1/S_2 \to \infty$ at t^*. On the complementary event, S_1 stays bounded and the ratio remains finite.

Remark 7.4 (On the F13 simulation). The simulation reported in F13 records 300 trials per noise level with zero failures. The integration caps capabilities at $S = 10^8$; trials hitting the cap are recorded as singleton emergence. For the noise levels and time horizons tested, this cap converts unbounded drift into apparent emergence even when the underlying SDE would not explode in continuous time. The simulation is therefore consistent with high $\Pr(J \geq c)$ in the tested regime, but does not provide independent evidence of almost-sure emergence.

8 Empirical Calibration to Frontier AI Compute

The framework's central conditional, "if A4 holds, a singleton emerges," is mute on whether A4 *does* hold for any system that matters. This section attempts a falsifiable check using publicly available data on training compute for notable machine-learning systems (Epoch AI, 2024). The result is honest and partial: under the proxy used here, the hypothesis $\beta(S) < 0$ is not currently supported.

8.1 Data and proxy

We use the Epoch AI "Notable AI Models" dataset (snapshot accessed 2026-05-13; 8,273 rows, of which 523 have both a usable publication date and a numerical training-compute estimate). The frontier is taken as the running maximum of training compute (in FLOP) over time; this gives 39 record-setting models from $\sim$1956 to 2026, of which 26 fall in the modern deep-learning window 2012–2026. We treat S = frontier training compute as a capability proxy. This is a strong assumption: training compute is one input to capability, not capability itself, and the relationship is mediated by algorithmic efficiency, data, and post-training techniques whose contributions are not recorded in the dataset. Results below should be read as a check on whether *this particular proxy* exhibits the kind of trajectory A4 requires; null results do not rule out A4 under different proxies.

8.2 Method

Two complementary fits. **Power-law fit.** Treating the frontier trajectory as $dS/dt = c\,S^{1-\beta}$, we form $y(t) = \ln S(t)$, estimate $\dot{y}$ by central differences, and regress $\ln \dot{y}$ on y; the slope is $-\beta$. This is sensitive to numerical-derivative noise on a sparse record-setter sequence.

Polynomial-in-time fit. Fit $\log_{10} S(t) = a + b\,t + \gamma\,t^2$ by ordinary least squares on the centered time variable. Curvature $\gamma > 0$ indicates super-exponential growth (and thus $\beta < 0$ in the power-law model); $\gamma = 0$ corresponds to exponential growth ($\beta = 0$); $\gamma < 0$ to deceleration ($\beta > 0$). This is more robust than numerical differentiation because it does not require estimating $\dot{y}$ at each frontier step.

8.3 Results

Window	n	γ	Asymptotic 95% CI	Bootstrap 95% CI	$\Pr(\gamma > 0)_{\text{boot}}$
2010–2018	9	+0.157	[−0.095, +0.409]	[−2.27, +1.19]	0.66
2018–2022	10	+0.082	[−0.154, +0.317]	[−2.91, +0.97]	0.58
2022–2026	7	−0.069	[−0.156, +0.018]	[−0.217, +0.240]	0.11
2012–2026	26	+0.005	[−0.010, +0.021]	[−0.008, +0.019]	0.82

Table 1: Polynomial-in-time fits of frontier $\log_{10}$ training compute. Curvature γ is the coefficient on t^2 in $\log_{10} S = a + bt + \gamma t^2$; positive γ corresponds to super-exponential growth ($\beta < 0$). Asymptotic CI is the OLS standard error; bootstrap CI is from 5,000 draws of the (year, $\log_{10} S$) pairs with replacement (seed 20260513). $\Pr(\gamma > 0)_{\text{boot}}$ is the bootstrap-empirical probability that resampled curvature is positive. The two CIs disagree most where they should: short windows ($n \leq 10$) have heavy-tailed bootstrap distributions because frontier records cluster, so a few influential observations can flip the curvature sign. The full-window result is robust under both methods. R^2 for the four windows is 0.87, 0.76, 0.99, 0.96.

The full deep-learning window 2012–2026 is consistent with pure exponential growth: $\gamma = +0.005$ with both CIs tightly straddling zero, $R^2 = 0.96$, and a stable doubling time near 0.5 years. The bootstrap probability of positive curvature is 0.82, indicating a slight directional tilt toward acceleration that is too small to matter at the resolution of the data. Sub-windows show mild acceleration during 2010–2022 and mild deceleration during 2022–2026; bootstrap CIs reveal that all sub-window estimates are statistically indistinguishable from exponential, and the short-window asymptotic CIs in particular understate uncertainty because frontier records cluster around

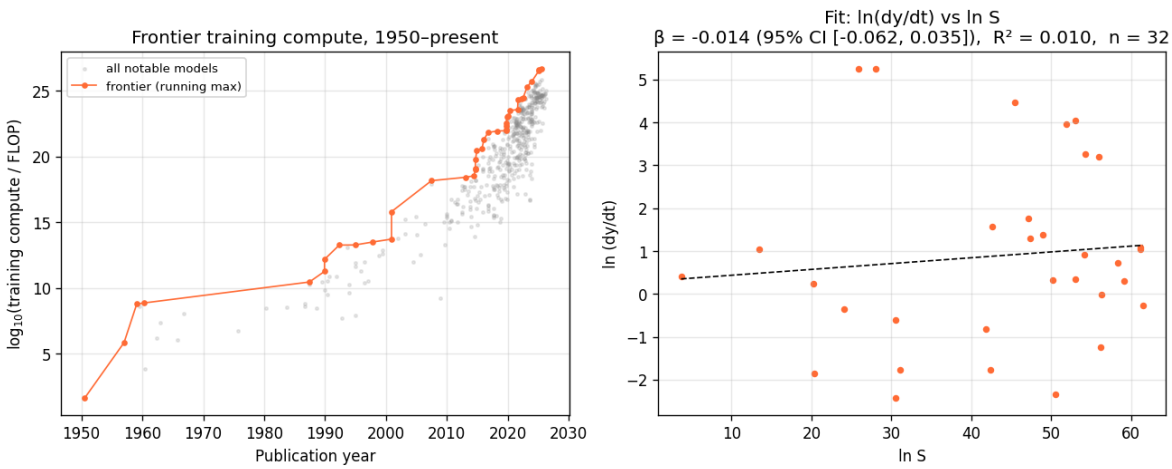

Figure 5: Left: training compute for all notable models (grey) and the frontier running maximum (orange). Right: power-law fit on $(\ln S,\ \ln \dot{y})$ where $y = \ln S$; slope is $-\beta$.

architectural breakthroughs. The power-law fit, which is noisier, gives $\beta = -0.014$ on the full window with a 95% CI of $[-0.062,\ +0.035]$.

8.4 Robustness: benchmark-performance proxies

Training compute is an input to capability rather than capability itself. We re-run the calibration on five public benchmarks released by Epoch AI (2024): GPQA Diamond, FrontierMath, ARC-AGI, SWE-bench Verified, and MATH Level 5. For each benchmark we form the running-maximum frontier of the best score reported across model versions over time, transform raw scores in $[0, 1]$ to an unbounded capability via $S = -\log_{10}(1-\text{score})$ (so $S \to \infty$ as score $\to 1$), and fit $\log_{10} S = a+bt+\gamma t^2$ on the transformed frontier.

Benchmark	n_{front}	γ	Bootstrap 95% CI	$\Pr(\gamma > 0)_{\text{boot}}$	R^2
GPQA Diamond	17	+0.010	[−0.116, +0.047]	0.57	0.96
FrontierMath	16	−0.651	[−0.953, −0.369]	0.00	0.97
ARC-AGI	12	+0.003	[−0.496, +1.138]	0.54	0.94
SWE-bench Verified	9	−0.454	[−0.955, +0.001]	0.03	0.92
MATH Level 5	13	+0.017	[−0.675, +0.371]	0.57	0.89

Table 2: Polynomial-in-time fits on benchmark-derived capability ($S = -\log_{10}(1 - \text{score})$) for five public benchmarks. Three of five give γ statistically indistinguishable from zero (exponential capability growth); two give γ significantly negative under the bootstrap (decelerating capability). None of the five give γ significantly positive.

Across the five benchmarks: three (GPQA, ARC-AGI, MATH) are indistinguishable from exponential capability growth; two (FrontierMath, SWE-bench) show statistically significant *deceleration*; none show acceleration. The compute-based and benchmark-based proxies agree on the qualitative finding: capability growth is at most exponential under every proxy tested.

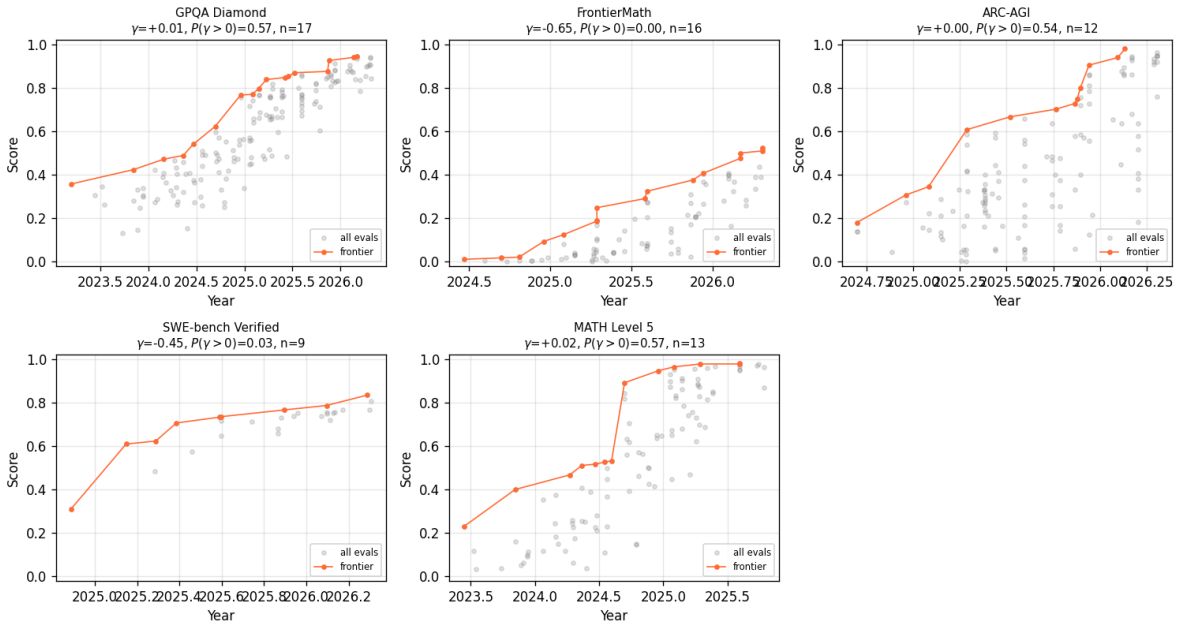

Figure 6: Per-benchmark frontier (orange) and all evaluations (grey), with fitted curvature γ and bootstrap probability of acceleration.

8.5 Reading

Across compute and five benchmark-based capability proxies, frontier capability is growing at most exponentially over the deep-learning window. On no proxy is the curvature significantly positive; on two benchmarks the curvature is significantly negative. The data are not consistent with $\beta(S) < 0$ during the period observed.

This does not refute A4. Three reasons the result is weak as evidence against the framework: (i) each proxy captures one slice of capability and none isolates the post-training, scaffolding, or inference-compute regimes that have become important since 2024; (ii) A4 requires only that $\beta < 0$ become reachable at some future T, not that the trajectory already be in that regime; the data are silent on the future; (iii) frontier-record counts are small ($n_{\text{front}} \in [9, 26]$ across proxies), making sub-window estimates noisy. Conversely, the result *is* evidence against the popular claim that capability growth is currently super-exponential under any of the proxies most commonly cited as evidence for it: a singleton emerging on these proxies alone within the next several years would require a sharp regime change, not continuation of the current trend.

The main contribution of this exercise is methodological: it makes A4 a falsifiable hypothesis on specific datasets. A future study could extend the calibration to capability proxies better suited to recent trends (effective compute, agentic capability scores, multi-step reasoning benchmarks) and to longer time horizons as data accumulate. Related empirical work: Ge et al. (2026) compare exponential, hyperbolic, and sigmoid models of capability growth on the METR time-horizon benchmark using Bayesian model selection, with no explicit super-exponential alternative; their methodology and dataset are complementary to the calibration here.

9 Discussion

9.1 What is new

Yudkowsky (2013), Omohundro (2008), and Lotka–Volterra are all prior work. The combination is not.

The main technical contribution is Theorem 3.4: finite-time separation *under competition*. Prior work treats the β-flip as a single-agent phenomenon. The closest adjacent formal treatment is Anbar Jafari et al. (2025), which establishes finite-time blow-up criteria via Osgood-style conditions on a single-agent capability ODE $\ddot{I} \geq aI^p$ and analyses physical and economic caps on runaway dynamics; that framework does not model multi-agent competition, coalition dynamics, or stochastic noise, and does not perform an empirical calibration. The present paper proves the β-flip survives adversarial resource dynamics, which is non-trivial: an opponent can slow the race to threshold, but not stop it.

The niche partitioning finding (F8) corrects Bostrom (2005). Separate resource pools are not a sufficient failure condition; what matters is whether the trailing agent can reach $\beta < 0$ at all.

Theorem 7.1 replaces an earlier incorrect almost-sure-emergence claim with the correct probabilistic statement: under multiplicative GBM noise, $\Pr(\text{singleton emerges}) = \Pr(J \geq c)$ for the explicit Dufresne perpetuity J. Noise can therefore prevent emergence, with probability tending to 1 as σ grows.

Theorem 6.2 derives the coalition-suppression threshold analytically. The transition near $\alpha = 0.64$ was previously only a simulation observation; it now has an exact transcendental equation behind it.

The simulation scripts produce quantitative measurements (timescale formula, λ_{crit}, moat characterization, F1–F25) that were not previously available.

Section 8 introduces an empirical test of A4 across six capability proxies (frontier compute plus five benchmarks): on no proxy is $\beta(S) < 0$ supported, and the framework's central premise becomes a falsifiable hypothesis with negative preliminary evidence.

9.2 Timescale scaling

The empirical scaling estimate in Section 4 has a partial analytical foundation. The $N^{0.96}$ exponent is analytically predicted as N^1 from the pre-threshold approximation (Appendix A); the small deviation comes from the leader's increasing resource share as it advances. The $\alpha^{-0.30}$ exponent tracks the integrated resource correction and the $|\beta_{\text{low}}|^{-0.31}$ exponent reflects a mixture of pre- and post-threshold phases (Appendix A). The $\text{gap}^{-0.15}$ exponent is small because the threshold mechanism dominates initial conditions.

9.3 Limitations

The N-agent conjecture. The inductive step works cleanly in the sequential limit but does not bound the error from truly simultaneous dynamics. The gap is real. Simulation (200/200 trials, $N = 10$) confirms the conclusion holds, but simulation on the same ODE system is not the same as closing the proof.

Local stability only. Proposition 3.11 gives local instability of the equal-resource state via Jacobian eigenvalues. It says nothing about whether trajectories starting far from that point converge to monopoly.

Scalar capability. The β-flip mechanism requires a scalar S crossing a single threshold T. Real optimization capability is multidimensional (compute, algorithmic efficiency, data access, adversarial robustness) and different agents can lead on different dimensions simultaneously. There is no obvious

way to extend the model to that setting. Whether competitive exclusion holds in a vector capability space is an open question, and the answer determines whether any of this applies to real AI systems.

Timescale formula. The multi-parameter scaling estimate in Section 4 is fitted from single-variable sweeps with no cross-validation. Only the N^1 exponent has an analytical derivation; the others should be read as scaling guides, not predictions.

Fast equilibration. Theorem 3.4 requires A2. With slow resource equilibration, the finite-time bound on t^* breaks down even if the ratio still diverges eventually.

Spatial structure. The model has none. Results apply within any causally connected region; adding propagation delays extends timescales without changing the qualitative picture.

Capability proxy in the calibration. Section 8 uses training compute as a stand-in for S. This proxy ignores algorithmic efficiency, data, post-training, and inference-time compute, all of which have become significant since 2024. A null result on this proxy is therefore weak evidence on A4 in general; it is strong evidence only against the specific narrative that "compute alone is going super-exponential."

9.4 Open questions

(1) Closed-form timescale exponents for α, gap, and $|\beta_{\text{low}}|$. The N^1 exponent is derived; others require numerical integration of the coupled ODE system.

(2) True merger stability: game-theoretic conditions under which a merged coalition avoids fracturing under competitive pressure.

(3) Calibration of $\beta(S)$ against capability proxies less tied to raw compute (effective compute, frontier benchmark performance, agent-evaluation scores), with explicit confidence bands on whether $\beta < 0$ is consistent with the data at any horizon.

10 Conclusion

Under A1–A5, competitive environments with recursive self-improvement produce a singleton. Once any agent crosses the β-threshold, its capability separates from all competitors in finite time. Initial conditions matter less than threshold position. Under deterministic dynamics noise reduces to a question of who wins; under multiplicative GBM noise it also affects whether anyone does (Theorem 7.1), with the probability of emergence expressible in closed form as a Dufresne perpetuity. Cooperation under any tested structure fails to prevent emergence in the deterministic model.

The failure conditions are narrower than the prior literature implies. Bostrom's niche partitioning argument requires that one agent cannot reach $\beta < 0$. Separate resource pools alone are not enough. Late entrants are a real threat only before threshold crossing. Coalitions fail not because they are irrational but because proportional internal resource distribution leaves each member individually weaker than a singleton candidate. A merger of the whole coalition would work, but the result is itself a singleton.

The hard question is A4. A3 holds in any physical environment. A5 is satisfied by noise. Section 8 converts A4 from a verbal premise into a falsifiable hypothesis on six concrete proxies (frontier training compute and five public benchmarks). On none of the six is the curvature significantly positive; on two benchmarks it is significantly negative; the full deep-learning window of frontier compute is consistent with pure exponential growth. This does not refute A4 (each proxy is a partial slice of capability, and the data are silent on the future), but it does shift the burden onto proponents of the singleton scenario to specify which capability proxy they expect to enter the $\beta < 0$ regime, and on what timescale.

Acknowledgments

No external funding was received.

Code and data availability

Simulation scripts, calibration code, and the Epoch AI dataset snapshot used in Section 8 are available at `https://github.com/ninjahawk/singleton-attractor`.

References

Akbar Anbar Jafari, Cagri Ozcinar, and Gholamreza Anbarjafari. A mathematical framework for AI singularity: Conditions, bounds, and control of recursive improvement, 2025.

Nick Bostrom. What is a singleton? *Linguistic and Philosophical Investigations*, 5(2):48–54, 2005.

Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.

Daniel Dufresne. The distribution of a perpetuity, with applications to risk theory and pension funding. *Scandinavian Actuarial Journal*, 1990(1–2):39–79, 1990.

Epoch AI. Parameter, compute and data trends in machine learning. 2024. Dataset accessed 2026; `https://epoch.ai/data/notable-ai-models`.

Haosen Ge, Hamsa Bastani, and Osbert Bastani. Are AI capabilities increasing exponentially? A competing hypothesis, 2026.

I. J. Good. Speculations concerning the first ultraintelligent machine. *Advances in Computers*, 6: 31–88, 1965.

Ioannis Karatzas and Steven E. Shreve. *Brownian Motion and Stochastic Calculus*. Springer, 2nd edition, 1991.

Alfred J. Lotka. *Elements of Physical Biology*. Williams and Wilkins, 1925.

John Maynard Smith. *Evolution and the Theory of Games*. Cambridge University Press, 1982.

Stephen M. Omohundro. The basic AI drives. In *Proceedings of the 2008 Conference on Artificial General Intelligence*, volume 171, pages 171–179, 2008.

Vito Volterra. Fluctuations in the abundance of a species considered mathematically. *Nature*, 118: 558–560, 1926.

Jörgen W. Weibull. *Evolutionary Game Theory*. MIT Press, 1995.

Eliezer Yudkowsky. Intelligence explosion microeconomics. Technical Report 2013-1, Machine Intelligence Research Institute, 2013.

A Derivation of timescale exponents

N^1 scaling (analytical). The derivation assumes N agents with equal initial capability, each starting with resource share $1/N$. A5 requires initial heterogeneity, so this is a conservative case: the leader's initial share is exactly $1/N$ (worst case), giving an upper bound on t_{cross}. With unequal initial conditions the leader starts with share $> 1/N$, so actual $t_{\text{cross}} \leq$ the equal-agent estimate. The leader's growth rate under the equal-agent approximation:

$$\dot{S}_{\text{lead}} \approx S^{1-\beta_{\text{high}}}/N.$$

Time for the leader to reach T from S_0:

$$t_{\text{cross}} \approx N \cdot \int_{S_0}^{T} S^{\beta_{\text{high}}-1}\, dS = N \cdot \frac{T^{\beta_{\text{high}}} - S_0^{\beta_{\text{high}}}}{\beta_{\text{high}}}.$$

This gives $t_{\text{cross}} \propto N$ exactly. The empirical $N^{0.96}$ reflects the leader's increasing resource share above $1/N$ as it advances, slightly accelerating threshold crossing.

$|\beta_{\text{low}}|^{-0.31}$ scaling. Post-threshold time to reach $10\times$ ratio from T:

$$t_{\text{post}} \approx \frac{T^{-|\beta_{\text{low}}|}}{|\beta_{\text{low}}| \cdot r_{\min}}.$$

This gives $t_{\text{post}} \propto |\beta_{\text{low}}|^{-1}$. Pre-threshold time t_{cross} is independent of β_{low}. The combined exponent -0.31 reflects the mixture: $t_{10\times} = t_{\text{cross}} + t_{\text{post}}$ where $t_{\text{cross}}/t_{10\times} \approx 0.6$ for default parameters, shifting the effective exponent from -1 (pure post-threshold) toward 0 (pure pre-threshold).

$\alpha^{-0.30}$ and $\text{gap}^{-0.15}$ scaling. Higher α concentrates resources on the leader more aggressively, reducing t_{cross} via a growing excess resource share above $1/N$. The excess scales as $\alpha \cdot \varepsilon/S_0$ where ε is the leader's relative advantage at each moment; the integrated effect gives $t_{\text{cross}} \propto \alpha^{-\delta}$, δ estimated empirically as 0.30. The gap exponent is small (-0.15) because gap affects the initial divergence rate but not the total path to T; the β-threshold mechanism dominates initial conditions. Exact closed-form values for both exponents require integrating the coupled ODE $\dot{x} = x^{1-\beta_{\text{high}}+\alpha}/(x^\alpha + (N-1)S_0^\alpha)$, which lacks a closed-form solution.

B Critical α derivation

The derivation of Theorem 6.2 is given in Section 6.2. Here we provide additional detail on the ODE reduction.

The ratio $\rho = y/x$ (coalition combined / singleton) satisfies:

$$\frac{d\rho}{dx} = \frac{\rho(\rho^\gamma - 1)}{x}.$$

For $\gamma > 0$ ($\alpha > \beta_{\text{high}}$) and $\rho > 1$ (coalition initially stronger combined), $d\rho/dx > 0$: the ratio grows as the singleton advances toward T. For $\gamma = 0$ ($\alpha = \beta_{\text{high}}$), $d\rho/dx = 0$: ρ is constant. For $\gamma < 0$ ($\alpha < \beta_{\text{high}}$), $d\rho/dx < 0$: singleton grows relatively faster.

The closed-form solution is:

$$\rho(x) = \left[\frac{1}{1-\eta(x)}\right]^{1/\gamma}, \quad \eta(x) = \left(1 - \left(\frac{x_0}{Nc}\right)^\gamma\right)\left(\frac{x}{x_0}\right)^\gamma.$$

$\eta(x)$ is increasing; $\rho(x)$ diverges when $\eta \to 1$, which occurs at finite x (potential blow-up before the singleton reaches T if the coalition grows fast enough).