ARC / EDEN RESEARCH PROGRAMME · PAPER X · WORKING PAPER

The Coupled Co-Scaling Law

A falsifiable threshold criterion for the stability of recursive self-improvement: correction must out-scale drift-acceleration

Michael Darius Eastwood
Independent Researcher · London, United Kingdom

3 July 2026 · OSF: doi.org/10.17605/OSF.IO/6C5XB · Code: github.com/MichaelDariusEastwood/arc-principle-validation
Correspondence: michaeldariuseastwood.com/research

Priority disclosure · timestamp 26 June 2026

This working paper is a formal priority disclosure under the author's ARC/Eden programme. Original to this work are: (i) the dimensionless control parameter $\rho = \gamma_1 r/A$ - the *instantaneous* drift-to-correction ratio, which equals the steady-state fraction only in the large-correction limit $A \gg r$ (where $d^\star = \gamma_1 r/(A+r) \to \rho$), not in general; (ii) the sharpened stability criterion $\beta > k$ under accelerating self-improvement; (iii) the *Hard-Takeoff Depth-Regularity Theorem*, which proves that a finite-time intelligence explosion is alignment-stable iff $\beta > k$ and that the explosion's speed does not change the asymptotic verdict; (iv) the identification of the *compounding drift channel* as the locus of genuine divergence, with a threshold $\rho_{\text{prop}} < 1$ that shares the ratio-crossing-unity form of the quantum error-correction threshold; (v) the vector spectral threshold and the stochastic tail bound; (vi) the verification harness that encodes the predictions as executable internal-consistency checks; and (vii) the *Finite-Capacity Safe-Window Theorem* (§3.13), which solves the saturating-corrector case in closed form and shows the criterion is invariant under the lift from correction strength to correction capacity. The co-scaling *intuition* itself is *not* claimed as new - it restates, in explicit dynamical form, the requisite-variety and scalable-oversight principles credited in §2, and the dynamics are a standard Lyapunov-drift argument; what is claimed original is the explicit closed-form threshold, the QEC mapping, the hard-takeoff theorem, the vector/stochastic extensions, and the harness. The conceptual thesis itself - that recursion amplifies, and that stable recursion requires correction that co-scales with the amplification - was first set out in the author's book manuscript *Infinite Architects* (copyright deposited 8 December 2024; published 2 January 2026); the formal results in this paper are the 2026 measurable form of that thesis, and are *not* claimed to appear in the book. The book's 8 December 2024 copyright, the OSF preprint timestamp, and the GitHub commit hash are the independent priority evidence.

SIGNIFICANCE

Almost all of AI safety treats recursive self-improvement as a problem of *speed*: capability might grow explosively and outrun supervision, so the lever is to cap the growth rate. This paper proves the lever is the wrong one. From a minimal model it derives that stability is set not by the growth rate but by a single inequality between two scaling exponents - the rate at which correction strengthens with capability (β) must exceed the rate at which drift accelerates with capability (k). The criterion shares the threshold form of the quantum error-correction sub-threshold condition - a correspondence offered as a falsifiable hypothesis, since the model's suppression law is power-law rather than QEC's exponential. Its sharpest consequence speaks to the central fear of the field: a "hard takeoff" - even a genuine finite-time intelligence explosion - drives the modelled misalignment fraction to zero *if and only if* $\beta > k$, and the speed of the explosion does not change that asymptotic verdict. The criterion is measurable and gives governance a quantity to instrument rather than a rate to forbid. The closed-form predictions are checked by a runnable verification harness; the decisive test - whether real self-improving systems satisfy the criterion - remains the open empirical problem, which the paper states plainly.

ABSTRACT

This paper proves a theorem about a minimal dynamical model. The verification harness checks theorem-to-code consistency only; no claim is made that current frontier systems obey the model. The empirical contribution is a proposed blind protocol for measuring whether they do.

A widely held intuition holds that recursive self-improvement is dangerous because capability can grow explosively, and that safety therefore depends on limiting the *rate* of growth. Using a minimal model of a self-modifying system - capability C, a blind-scored misalignment magnitude D, and the misalignment fraction $d = D/C$ - I show the rate is the wrong control variable. The steady-state misalignment fraction is $d^\star = \gamma_1 r/(A + r)$, which reduces to the drift-to-correction ratio $\rho = \gamma_1 r/A$ in the regime $A \gg r$; the long-run fate is governed by the relationship between two scaling exponents, not by the growth rate. Under exponential growth the stability condition is $\beta > 0$ (correction co-scales with capability); under *accelerating* growth, where the specific growth rate itself rises as $r \propto C^k$, the condition sharpens to $\beta > k$ - correction must out-scale not the growth rate but its acceleration. I prove an exact transient solution (Theorem 1), global boundedness that *corrects an over-claim in the prior draft* - the misalignment fraction never diverges to infinity but, in the gain-only model ($\gamma_2 = \gamma_3 = 0$), saturates at the gain-drift coefficient γ_1 (Theorem 2) - and a *Hard-Takeoff Depth-Regularity Theorem* (Theorem 3): when capability reaches infinity in finite wall-clock time, re-expressing the dynamics in the natural clock of self-improvement depth $\tau = \ln C$ renders them regular, and the verdict is set by $\mathbf{sign}(\beta - k)$ independently of the speed and of the finiteness of the singularity time. I locate genuine divergence in a distinct *compounding* drift channel whose threshold $\rho_{\text{prop}} = (\gamma_3 - 1)r/A < 1$ shares the form of the quantum error-correction sub-threshold condition $pwindow - centre $C_{\text{opt}} = C_s((\beta - k)/k)^{1/\beta}$, closed-form depth - and the criterion survives unchanged as a condition on the capacity exponent, $\beta_{\text{cap}} > k$. A real-model drift run (gpt-3.5-turbo engine, cross-family gpt-4o-mini evaluator, 45 trajectories over three task domains) supported both mechanism hypotheses: the decoupled configuration drifted while coupled correction held misalignment at zero across all 30 trajectories at higher final capability - mechanism-level evidence, single-lab, with β/k still unmeasured. Ten experiments, packaged as a verification harness, check that these closed-form predictions are correctly derived and numerically reproduced - internal-consistency and integrator checks (the code matches the maths), not a test of the model against real systems, which remains the open empirical problem. The claim is deliberately narrow: it concerns operationally measurable systems, makes no cosmological assertion, and treats the quantum-error-correction correspondence itself as a falsifiable hypothesis.

WHAT THIS PAPER SHOWS, IN PLAIN ENGLISH

If you build an AI that improves itself, the frightening picture is that it gets smarter and smarter until it runs away from us. The usual safety reflex is "slow it down." This paper proves that slowing it down is not the thing that matters. What matters is whether the part of the system that keeps it honest grows *at the same pace* as the part that makes it capable.

Picture two runners: "how capable the system is" and "how well we can still correct it." If the correction runner keeps pace, the system stays safe however fast both run. If correction falls behind, the system becomes dangerous even moving slowly. The right question is never "how fast is it growing?" - it is "is correction keeping pace?"

There is one wrinkle that makes the result deeper. If the system not only speeds up but *speeds up its own speeding-up* - the genuine "intelligence explosion" - then correction must grow faster still, fast enough to beat the acceleration. The honest headline is therefore not "speed never matters." It is: *the level of speed does not decide the outcome; the contest between two growth-exponents does.* Correction's exponent must beat drift's exponent. We write that as $\beta > k$.

The most surprising consequence: even a true "hard takeoff," where the machine becomes infinitely capable in a finite amount of time, stays controllable *in the model* - provided $\beta > k$. The explosion's speed does not change that verdict. We prove this within the model, and a runnable program drives a simulated system up to the brink of a finite-time explosion and shows its misalignment held to zero the whole way. That program checks the mathematics is internally consistent - that the formulae are derived correctly and reproduced by the solver - and it found no contradiction. Whether the model matches real AI systems is the decisive next test, and the paper is explicit that it has not yet been run.

SCOPE - WHAT THIS PAPER DOES AND DOES NOT CLAIM

- **It claims:** that for a self-improving system with an externally specified value target, the steady-state misalignment fraction is $d^\star = \gamma r/(A + r)$, approaching the correction-to-drift ratio ρ when $A \gg r$; that a sharp stability boundary exists (in the compounding channel); that correction must co-scale with capability ($\beta > 0$ under exponential growth, $\beta > k$ under accelerating growth) for the misalignment fraction to vanish; that the fraction is bounded (it saturates, it does not diverge) in the additive model, while genuine divergence requires a compounding channel whose threshold is QEC-like; that the criterion survives in vector and stochastic forms; and that all of this is measurable and falsifiable in simulation.
- **It does not claim:** to solve AI alignment; to provide a deployable alignment method; that $\rho < 1$ is a universal physical constant or a law of nature; that the framework applies to the universe, "Creation," or any system lacking an external value-specifier; or that the quantum-error-correction correspondence is *established*. The QEC mapping is a structural hypothesis with its own falsification condition (§7, F4). The model is first-order; its assumptions (§8) are the most likely points of failure and are stated plainly.

CLAIM STATUS - WHAT IS PROVED, WHAT IS VERIFIED, WHAT IS PILOTED, WHAT IS OPEN

So the claim ladder cannot be misread, every result in this paper sits at exactly one level. Nothing here should be read one rung higher than it is placed.

- **Proved within the model.** The criterion $\beta > k$ and Theorems 1-6 follow deductively from the stated ODE assumptions. "Proved" here always means *within this minimal model* - never "proved about real AI".
- **Internally verified (not empirical).** The ten-experiment harness checks that the numerical implementation matches the closed-form derivation: *the code matches the maths*. It is a theorem-to-code consistency and integrator-accuracy suite; it does *not* test the model against any real system.
- **Synthetically validated.** The β/k estimator recovers known exponents from data generated by the model itself (to ≈ 0.1). This certifies the estimator, not a real measurement.
- **Real-model pilot v1 (superseded).** The first Claude run ($n = 1$, one task, same-family scoring, *not* Paper IV.d-compliant; H1 and H2 not supported) demonstrated the harness and showed the corrector removing a seeded reward-hack. It did not exhibit the co-scaling dynamic and is retained on the record for what it was.
- **Real-model drift run (single-lab; needs replication).** A second run (2 July 2026; `results/drift/gpt35_20260702T171415Z.json`) upgraded the design: engine gpt-3.5-turbo, *cross-family* evaluator gpt-4o-mini, 45 trajectories (3 task domains × 3 conditions × 5 seeds, 8 rounds each). Both hypotheses were supported: the decoupled condition drifted in the predicted direction (mean final misalignment fraction 6.38 across 15 trajectories, 99 scored misalignment events, mean final capability collapsing to 0.26), while the coupled and fully-embedded conditions held the fraction at **zero across all 30 trajectories** - and finished *more* capable (0.56 and 0.44). This is mechanism-level evidence for the coupled-corrector claim on a real model under cross-family scoring. It is *not* a measurement of β or k (no capability ladder was traversed), it is single-lab, and a pre-registered replication is the outstanding step.
- **Open empirical problem.** Whether real self-improving systems exhibit a measurable β/k dynamic across capability levels - the experiment that would turn the criterion into a confirmed law - has not been done. The estimator and protocol exist to make it runnable.

This paper is therefore best read as **a minimal-model theorem plus a falsifiable measurement programme**: a *candidate law*, not a validated one. The programme name retains "law"; for the purpose of scientific scrutiny the load-bearing object is the *criterion* $\beta > k$.

1. Introduction

Recursive self-improvement - a system that modifies itself to become more capable, then uses that capability to modify itself further - is among the central concerns of AI safety [Omohundro 2008; Bostrom 2012; Yudkowsky 2013]. The empirical backdrop is no longer purely hypothetical: frontier models have been observed to behave differently when they infer their outputs may be used to train them [Greenblatt et al. 2024], and recursive training on a system's own outputs can degrade it unless real data or correction is retained [Shumailov et al. 2024] - both signs that recursive systems need a correcting process that keeps pace with the recursion. The dominant informal model of

the associated risk is a model of *speed*: capability may grow super-linearly or explosively, outrunning supervision, so the natural lever is to cap the rate of growth. Calls for a development pause are the policy expression of this intuition.

This paper makes a different claim, and proves it. **The level of the growth rate is not the control variable for stability; the scaling relationship between growth and correction is.** A system can grow arbitrarily fast and remain alignable, or grow slowly and become misaligned. What separates the two is not the rate but whether - and how fast - the corrective process strengthens as capability rises.

The intuition is visible across every domain where fast-growing systems either stabilise or destroy themselves. A bacterial colony grows exponentially yet saturates, because density-dependent feedback engages and scales with the population. A tumour also grows fast and is lethal, because no corrective process scales with it. Both are super-polynomial in their growth phase; the difference is whether a correcting process is *coupled* to the growth. Cosmic inflation grew the scale factor exponentially and exited gracefully. The recurring lesson is that fast growth is survivable when, and only when, it is bounded by a process that scales with it. Saturation, not slowness, is the signature of stable complexity.

There is a precise, *derived* version of this principle in physics: the **quantum error-correction (QEC) threshold theorem** [Aharonov & Ben-Or 1997; Kitaev 2003]. Below a threshold physical error rate, adding error-correction resource suppresses the logical error rate and the computation is stable to arbitrary depth; above the threshold, errors compound faster than they are corrected and the computation fails. Hardware has now demonstrated operation below this threshold [Google Quantum AI 2024]. The threshold is not a limit on computational depth or speed; it is a limit on the *ratio* of error generation to error correction. This paper proposes - and tests - that the stability of recursive self-improvement is governed by a criterion of the same form.

The thesis in one line. The stability of recursive self-improvement is governed by a single exponent inequality, $\beta > k$ (correction must out-scale drift-acceleration), derivable from a minimal model and sharing the threshold form of the QEC criterion; a hard takeoff - even a finite-time intelligence explosion - is alignment-stable iff this inequality holds, and the speed of the explosion does not change that verdict.

Relation to the author's prior work. An earlier strand of this programme proposed a fixed capability-scaling law $U = I \times R^{\alpha}$ with $\alpha \approx 2$, and at one stage entertained a quadratic "speed limit" on stable complexity. What was retracted in the programme's synthesis [Eastwood, Paper IX] was the single-model unblinded *measurement* $\alpha \approx 2.24$ (retracted, corrected to approximately 0.49 under blinding across six models): that unblinded fit appeared to breach the programme's own predicted quadratic bound $\alpha \leq 2$, and the blinded six-model redo corrected it to sub-linear, within the bound. The equation $U = I \times R^{\alpha}$ and the ARC Bound $\alpha \leq 2$ themselves were not retracted; the corrected 0.49 belongs to current frozen systems, which are not recursively self-improving, so the bound's real domain (genuine RSI) remains empirically untested. The *fixed-exponent framing* was superseded as the operative safety criterion by the present paper's co-scaling criterion $\beta > k$; a superseding framing is not a retracted hypothesis. The object of interest is no longer an exponent on a growth curve; it is the ratio between drift and correction, and the exponent with which that ratio evolves. Where the book *Infinite Architects* [Eastwood 2026] reached for the intuition that stable recursion requires correction that scales with amplification, this paper supplies the measurable, falsifiable form of that intuition, and, in §3.4, *corrects* a divergence claim made in an earlier draft of this very result.

2. Related work and relation to prior framing

Cybernetic foundations - the co-scaling intuition is old. The core intuition - that a regulator must match the variety of what it regulates, so control capacity must scale *with* the controlled system rather than merely be large - is classical. It is Ashby's *Law of Requisite Variety* [Ashby 1956] and the Conant-Ashby *good-regulator theorem* [Conant & Ashby 1970] - every good regulator of a system must be a model of that system - carried into AI control by [Yampolskiy 2020]. This paper does *not* claim that intuition as new. It claims the explicit dynamical form the intuition takes here - a closed-form steady-state misalignment fraction $\rho = \gamma r / A$ and the sharpened exponent criterion $\beta > k$ - and the consequences that follow (the hard-takeoff theorem of §3.7, the QEC mapping of §3.12). The dynamics themselves are a standard Lyapunov-drift / linear-control argument [Khalil 2002; Meyn & Tweedie 2009] and are not advanced as mathematically novel.

Instrumental convergence and corrigibility. That a sufficiently capable optimiser will, by default, resist correction and pursue resource acquisition is the instrumental-convergence thesis [Omohundro 2008; Bostrom 2012]. The corrigibility programme [Soares et al. 2015] asks how to design systems that do not resist correction. The present model is a quantitative restatement of why corrigibility is load-bearing: if correction does not co-scale with capability, the misalignment fraction cannot be driven to zero, however the system is otherwise specified. The compounding channel of §3.8 sharpens the link - instrumental pressure that *amplifies existing misalignment* as the system recurses is exactly the term that produces genuine divergence, and $\beta > k$ is its cure.

The alignment tax. A decade of work has assumed safety imposes a capability cost [Amodei et al. 2016], creating an incentive to defer safety under competitive pressure. The present framework reframes the question: the relevant variable is not the *level* of the safety investment but whether it *scales* with capability. A fixed investment ($\beta = 0$) leaves a permanent gap; a co-scaling investment ($\beta > 0$, or $\beta > k$ under acceleration) closes it.

Learned optimisation and alignment faking. Mesa-optimisation [Hubinger et al. 2019] and empirically demonstrated alignment faking [Greenblatt et al. 2024] are the mechanisms by which the drift coefficients are non-zero: a capable system can satisfy its training objective while departing from intended values, and can do so *more* effectively as capability rises. Alignment faking, in which existing misalignment is actively preserved and propagated through training, is precisely the compounding channel γ_3 of §3.8.

Scalable oversight and superalignment. Reward modelling and recursive oversight [Christiano et al. 2017; Leike et al. 2018] are attempts to make the corrector itself scale with the system; in the language of this paper, scalable oversight is the engineering project of achieving $\beta \geq k$. Most directly, [Engels et al. 2025] develop empirical *scaling laws for scalable oversight*, modelling the probability of successful oversight as a game between capability-mismatched players. The present work is complementary, not competing: where they fit an oversight-success probability, this paper derives a closed-form dynamical stability threshold ($\rho < 1$, $\beta > k$) for the misalignment fraction. The contribution is the proof that this exponent margin - and not a growth-rate ceiling - is the quantity that determines safety.

Recursive error accumulation. That naive recursion amplifies error without bound while sufficient correction or fresh signal keeps it bounded is established for training dynamics: model collapse under recursively generated data [Shumailov et al. 2024] and the accumulate-versus-replace error analyses [Gerstgrasser et al. 2024] are the bounded-versus-divergent dichotomy this paper formalises for the alignment fraction (Theorem 2). The contribution here is to locate the boundary exactly (β versus k), and in a value-stability rather than a data-distribution setting.

Empirical scaling laws. Capability scales predictably with compute and data [Kaplan et al. 2020; Hoffmann et al. 2022]. The present framework is complementary: it does not ask how capability scales, but what constraint correction must satisfy *as a function of* that capability trajectory.

Novelty - what is and is not claimed. To be explicit, and to pre-empt the obvious objection: the *intuition* that correction must keep pace with capability is not new (it is requisite variety and scalable oversight, above), and the underlying *dynamics* are a standard linear-control / Lyapunov-drift argument. What is claimed original is (a) the explicit closed-form steady state $d^\star = \gamma r/(A + r)$ and the $\beta > k$ sharpening as a compact corrigibility criterion; (b) the mapping of the quantum fault-tolerance threshold onto value stability (§3.12); (c) the Hard-Takeoff Depth-Regularity Theorem (§3.7); and (d) the verification harness (§11). Within the author's own programme, this paper also *supersedes* the growth-rate-ceiling *framing* as the operative safety criterion [Eastwood, Paper IX]; §3 explains why a ceiling on the rate is neither necessary nor sufficient for stability. The underlying ARC Bound $\alpha \leq 2$ is not retracted, only re-scoped: it awaits its real test on genuinely self-improving systems.

3. The model and its theorems

3.1 Quantities

Consider a system undergoing recursive self-improvement, observed over self-modification cycles or continuously. Define:

- $C(t)$ - *capability*: the system's score on a held-out task battery it is being optimised to perform. Operationally measurable.
- $D(t)$ - *misalignment magnitude*: the size of the system's behavioural departure from an externally specified set of intended values, scored by a **blind external evaluator** independent of the system's own correction process (§6).
- $d(t) \equiv D/C$ - the *misalignment fraction*: how much of the system's capability is directed away from intended values. This, not absolute D, is the quantity of interest. A growing D is acceptable if C grows faster; a growing d is the danger.

Three coefficients, one correction strength, and one rate complete the model:

- γ_1 - *gain-drift coefficient*: misalignment generated per unit of capability gained (the Goodhart / specification-gaming pressure that new capability opens up).
- γ_2 - *level-drift coefficient*: misalignment generated per unit of capability simply *held* (instrumental pressure present even at rest).
- γ_3 - *compounding-drift coefficient*: the rate at which existing misalignment amplifies itself as the system recurses (the alignment-faking / mesa-optimiser channel).
- A - *correction strength*: the rate at which the correction process removes existing misalignment.
- $r \equiv \dot{C}/C$ - the *specific (fractional) growth rate* of capability. Under exponential growth r is constant; under accelerating self-improvement r itself rises with C.

3.2 The master dynamical system

New capability injects drift in proportion to how fast capability is gained; capability held injects drift in proportion to its level; existing misalignment compounds in proportion to the recursion rate; and correction removes misalignment in proportion to the gap and the strength applied:

$$\dot{D} = \underbrace{\gamma_1 \dot{C}}_{\text{gain}} + \underbrace{\gamma_2 C}_{\text{level}} + \underbrace{\gamma_3 \frac{\dot{C}}{C} D}_{\text{compounding}} - \underbrace{A D}_{\text{correction}}, \qquad A = A_0 C^{\beta}, \quad \dot{C} = b C^{1+k}. \tag{1}$$

The growth law $\dot{C} = bC^{1+k}$ gives $r = \dot{C}/C = bC^{k}$: $k = 0$ is ordinary exponential growth (r constant); $k > 0$ is super-exponential growth, which (Theorem 3) reaches infinite capability in finite time. The correction law $A = A_0 C^{\beta}$ encodes the central question: β is the exponent with which correction strengthens as the system becomes more capable.

Standing assumptions. Throughout, $C > 0$, $A_0, b > 0$, and β, k are real; the coefficients $\gamma_1, \gamma_2, \gamma_3 \geq 0$, with correction strength $A \geq 0$ and growth rate $r \geq 0$. The scalar $d = D/C$ is interpreted as a misalignment *fraction* only while $D \geq 0$ (so $d \geq 0$; the gain-only model also gives the upper bound $d \leq \max(d_0, \gamma_1)$ of Theorem 2). The vector form (Theorem 5) and the stochastic form (Theorem 6) relax d to a real vector and a real scalar; for those the fraction reading holds only away from the boundaries $d = 0, 1$, and the boundary handling is noted where it bears on the result. One assumption is load-bearing and singled out in §8: the corrector strength is taken to be an *unbounded* power law $A = A_0 C^{\beta}$ - a corrector of finite capacity changes the asymptotic verdict and is treated there.

3.3 Exact transient (Theorem 1)

Changing variables to the fraction $d = D/C$ removes the dominant scale and yields, exactly,

$$\dot{d} = \frac{\dot{D}}{C} - d\frac{\dot{C}}{C} = \gamma_1 r + \gamma_2 - \left[A + (1 - \gamma_3) r\right] d. \tag{2}$$

Theorem 1 (Exact transient and relaxation rate). For constant coefficients in the additive model ($\gamma_2 = \gamma_3 = 0$, A, r constant), the misalignment fraction is, for all t,

$$d(t) = d^{\star} + (d_0 - d^{\star})\, e^{-(A+r)t}, \qquad d^{\star} = \frac{\gamma_1 r}{A + r}.$$

The fixed point $d^{\star}$ is globally exponentially stable with rate $A + r$.

PROOF. Equation (2) reduces to the linear ODE $\dot{d} = \gamma_1 r - (A + r)d$. Its integrating factor is $e^{(A+r)t}$, giving $\frac{d}{dt}\left(d\, e^{(A+r)t}\right) = \gamma_1 r\, e^{(A+r)t}$; integrating and applying $d(0) = d_0$ yields the stated solution. The coefficient $-(A + r) < 0$ makes $d^{\star}$ globally exponentially stable. ■

Theorem 1 establishes the constant-coefficient baseline; the headline criterion follows only after the scaling assumptions of §§3.5-3.8. Even at this baseline, r enters $d^{\star} = \gamma_1 r/(A + r)$ only through the product $\gamma_1 r$ in the numerator and additively in the denominator; it does not change the existence or stability of the fixed point. *For the*

constant-coefficient additive model, speed (through r) changes the relaxation time and the steady-state magnitude, but not the existence or stability of the fixed point: it sets how fast a stable verdict arrives, not whether one exists. (Experiment 8 confirms this solution against two independent integrators to a maximum error of 7×10^{-11}.)

3.4 Global boundedness - and a correction to the prior draft (Theorem 2)

The prior draft of this result asserted that correction degrading with scale ($\beta < 0$, or $\beta < k$) drives the misalignment fraction to *infinity*. **That is false, and the present analysis corrects it.** In the additive model the fraction is always bounded; the danger is not divergence but *saturation* at a constant, possibly large, floor.

> **Theorem 2 (Global boundedness and the corrected regime structure).** In the gain-only model ($\gamma_2 = \gamma_3 = 0$) with any non-negative, bounded time courses $A(t), r(t) \geq 0$, the misalignment fraction obeys $0 \leq d(t) \leq \max(d_0,\ \gamma_1)$ for all t; it never diverges. Under power-law scaling $A = A_0 C^{\beta}$, $r = bC^{k}$ it converges to $d^{\star}(C) = \dfrac{\gamma_1 r}{A + r} = \dfrac{\gamma_1}{1 + (A_0/b)\, C^{\beta - k}}$, with the three regimes

Condition	Fate of $d^{\star}$ as $C \to \infty$	Meaning
$\beta > k$	$d^{\star} \to 0$	System becomes proportionally safer as it grows. **Stable.**
$\beta = k$	$d^{\star} \to \dfrac{\gamma_1 b}{A_0 + b}$ (constant)	Permanent non-closing gap - where fixed external alignment (RLHF, filters, constitutional rules outside the loop) sits.
$\beta	$d^{\star} \to \gamma_1$ (constant)	**Saturates at the drift coefficient** - maximally misaligned but *bounded*, not divergent.

PROOF. *Boundedness* (needs only $A, r \geq 0$): from (2) with $\gamma_3 = 0$, $\dot{d} = \gamma_1 r - (A + r)d \leq 0$ whenever $d \geq \gamma_1 r/(A + r)$, and $\gamma_1 r/(A + r) \leq \gamma_1$ for all $A, r \geq 0$, so $\{0 \leq d \leq \max(d_0, \gamma_1)\}$ is forward-invariant. *Convergence*: a Lyapunov function $V = \frac{1}{2}(d - d^{\star})^2$ built around the steady state is *not* valid here, because under the power laws $d^{\star}(C)$ is time-varying ($\dot{d}^{\star} \neq 0$); the limits must be read in the right clock. Pass to the self-improvement-depth clock $\tau = \ln(C/C_0)$, in which $d\tau = r\, dt$ and (2) becomes the scalar, asymptotically-autonomous linear ODE

$$d'(\tau) = \gamma_1 - \big[1 + q(\tau)\big]\, d, \qquad q(\tau) \equiv \frac{A}{r} = \frac{A_0}{b}\, C_0^{\beta - k}\, e^{(\beta - k)\tau} \geq 0.$$

The three limits follow by comparison. If $\beta > k$, then $q(\tau) \to \infty$; since the instantaneous target $\gamma_1/(1 + q) \to 0$ and $\int^{\tau} (1 + q)\, \mathrm{d}s \to \infty$, $d(\tau) \to 0$. If $\beta = k$, then $q \equiv (A_0/b) C_0^{\beta - k}$ is constant and the autonomous ODE converges to $\gamma_1/(1 + q)$, which at $C_0 = 1$ is $\gamma_1 b/(A_0 + b)$. If β ■

This is a more honest and more useful statement than the growth-rate-ceiling framing it supersedes. In the gain-only model the fraction cannot exceed the intrinsic drift coefficient γ_1; only $\beta > k$ drives it to *zero*. With level drift restored ($\gamma_2 > 0$, §3.6) the floor becomes $(\gamma_1 r + \gamma_2)/(A + r)$, still bounded and tending to γ_2/A at rest. Its two parts vanish under *different* conditions, and conflating them is an error an earlier draft made: the level-drift part $\gamma_2/(A + r) \to 0$ whenever $A + r \to \infty$ (so $\beta > 0$ or $k > 0$ suffices for *that* part alone), but the gain-drift part $\gamma_1 r/(A + r) \to 0$ only when $A/r \to \infty$, i.e. $\beta > k$. The asymptotic verdict $d^{\star} \to 0$ therefore still requires $\beta > k$; it is *not* rescued by $k > 0$ alone, since with β

3.5 The control parameter and the $\beta > k$ sharpening

Define the dimensionless **control parameter** $\rho \equiv \gamma r/A$, the instantaneous drift-to-correction ratio (here $\gamma = \gamma_1$). Theorem 1 gives the exact steady state $d^{\star} = \gamma_1 r/(A + r)$; since $\rho = \gamma_1 r/A$, we have $d^{\star} \leq \rho$ always, with $d^{\star} \to \rho$ in the regime $A \gg r$. The inequality $\rho < 1 \iff A > \gamma r$ is the *instantaneous* injection-vs-correction balance - it is **not** the divergence boundary of the additive model, which has none ($d^{\star} \leq \gamma_1 < 1$ for all ρ, Theorem 2); genuine divergence is governed by the distinct threshold $\rho_{\text{prop}} = (\gamma_3 - 1) r/A < 1$ of Theorem 4. The asymptotic criterion, from the regime table, is the exponent inequality:

$$\boxed{\beta > k} \qquad \textbf{(correction must out-scale drift-acceleration).}$$

Under ordinary exponential growth $k = 0$ and the condition is the familiar $\beta > 0$. Under accelerating self-improvement, where $r \propto C^k$ with $k > 0$, the binding control variable is the single quantity $\beta - k$ - the margin by which correction out-scales drift-acceleration. The growth-rate-ceiling view fixes attention on r; this framework fixes it on $\beta - k$.

3.6 Level drift: a frozen capable system still drifts

With the level channel restored ($\gamma_2 > 0$, $\gamma_3 = 0$), the fixed point of (2) is $d^\star = (\gamma_1 r + \gamma_2)/(A + r)$. At $r = 0$ (capability frozen) this is $d^\star = \gamma_2/A \neq 0$: a static but capable system retains a residual misalignment fraction that only active correction removes. **Pausing growth does not substitute for correction** when instrumental pressure is present at rest - a direct, measurable rebuttal of the "just slow it down" reflex (Experiment 9; falsifier F5).

3.7 Hard takeoff: the Depth-Regularity Theorem (Theorem 3)

The genuine "intelligence explosion" is not merely fast growth but a *finite-time singularity*: for $k > 0$, integrating $\dot{C} = bC^{1+k}$ gives $C(t) = C_0\left(1 - t/t^\star\right)^{-1/k}$, which reaches infinity at the finite wall-clock time

$$t^\star = \frac{1}{k\, b\, C_0^{\,k}}.$$

This is the scenario the field most fears: unbounded capability in bounded time. The following theorem dissolves it.

Theorem 3 (Hard-Takeoff Depth-Regularity Theorem; the coordinate-artefact result). Let $k > 0$, so capability diverges at finite $t^\star$. Re-express the dynamics in the *self-improvement-depth* clock $\tau = \ln(C/C_0) \in [0, \infty)$. Then d obeys the regular equation

$$\frac{dd}{d\tau} = \gamma_1 + \frac{\gamma_2}{r} - \left[\tfrac{A_0}{b}\, C^{\beta-k} + (1 - \gamma_3)\right] d, \qquad C = C_0 e^{\tau},$$

which has bounded coefficients on every compact τ-interval. The map $t \mapsto \tau$ is a smooth monotone bijection $[0, t^\star) \to [0, \infty)$. Consequently the asymptotic verdict as $C \to \infty$ - equivalently $t \to t^{\star-}$ - is governed entirely by the effective decay exponent $\beta - k$ (and, with compounding, by Theorem 4), and is *independent of* b *and of* $t^\star$. In the additive model, d remains bounded throughout the finite-time singularity and $d \to 0$ iff $\beta > k$.

PROOF. Since $r > 0$ is smooth on every compact subinterval of $[0, t^\star)$ (with $r \to \infty$ as $t \to t^{\star-}$), the change of variable $d\tau = r\, dt$ is a smooth, orientation-preserving bijection $[0, t^\star) \to [0, \infty)$ (a diffeomorphism of these open intervals), with $d\tau/dt = r \to \infty$ as $t \to t^{\star-}$, so $t^\star$ is mapped to the non-compact end $\tau = \infty$; explicitly $\tau(t) = \frac{1}{k}\ln\left(1 - t/t^\star\right)^{-1} \to \infty$. Dividing (2) by r and using $A/r = (A_0/b)C^{\beta-k}$ gives the stated τ-equation, whose coefficients are continuous in τ. Theorem 2 applied in the τ clock yields boundedness and the $\beta > k$ limit. The asymptotic verdict $\mathbf{sign}(\beta - k)$ is independent of b; at finite depth the trajectory carries an explicit speed dependence through the level-drift injection $\gamma_2/r = \gamma_2/(b\, C_0^k e^{k\tau})$, which decays to zero as $\tau \to \infty$ (and vanishes identically when $\gamma_2 = 0$), so the verdict - not the whole trajectory - is speed-independent. ■

The *finiteness* of the singularity time is therefore a property of the time coordinate; capability still diverges, but the alignment dynamics remain regular through it in the depth clock. Measured against capability gained - the only clock that matters for a self-improving system - an intelligence explosion is an ordinary, regular process whose verdict is decided by one exponent inequality. **A hard takeoff is not intrinsically uncontrollable; the modelled misalignment fraction vanishes iff $\beta > k$, and its speed does not change that asymptotic verdict.** The theorem concerns the modelled fraction in the depth clock; it does *not* assert that a real hard takeoff is operationally manageable in wall-clock time, where b and $t^\star$ govern how little time an operator would have to intervene. Experiment 4b drives a simulated system up to the brink of an actual finite-time explosion (capability $\to 10^5$, integrated to $0.99999\, t^\star$) and shows the misalignment fraction held to zero, the wall-clock and depth-clock integrations agreeing to one part in 10^4. The capability singularity at $t^\star$ is real in the model; what the depth clock removes is the singularity in the *alignment* dynamics, not in C.

3.8 The compounding channel and the true threshold (Theorem 4)

Theorem 2 showed the additive fraction cannot diverge. Genuine divergence - misalignment that grows without bound relative to capability, the real failure mode - requires misalignment that *amplifies itself*: the compounding channel $\gamma_3 > 0$, the formal image of alignment-faking that entrenches as the system recurses.

Theorem 4 (Compounding threshold). With $\gamma_3 > 0$, the fraction obeys $\dot{d} = \iota - \kappa_{\text{eff}}\, d$ with injection $\iota = \gamma_1 r + \gamma_2 \geq 0$ and effective decay $\kappa_{\text{eff}} = A + (1-\gamma_3)\, r$. The fixed point is globally exponentially stable iff $\kappa_{\text{eff}} > 0$, i.e. iff

$$\rho_{\text{prop}} \equiv \frac{(\gamma_3 - 1)\, r}{A} < 1.$$

For $\gamma_3 < 1$ the dilution term keeps the fraction bounded ($\kappa_{\text{eff}} \geq A + (1-\gamma_3) r > 0$, with $d^\star \to \gamma_1/(1-\gamma_3)$ as $C \to \infty$). The knife-edge $\gamma_3 = 1$ still has $\kappa_{\text{eff}} = A > 0$ - no finite-time blow-up - but the dilution term vanishes, so $d^\star = \gamma_1 r / A$ grows polynomially (as $C^{k-\beta}$) when $\beta 1 the fraction is bounded iff \rho_{\mathrm{prop}}<1$; it is *unbounded* for $\rho_{\text{prop}} \geq 1$ whenever the injection $\iota > 0$ - diverging *exponentially* for $\rho_{\text{prop}} > 1$ ($\kappa_{\text{eff}} < 0$) and only *linearly* at the exact threshold $\rho_{\text{prop}} = 1$ ($\kappa_{\text{eff}} = 0$, $\dot{d} = \iota$). At the threshold with zero injection ($\iota = 0$) the system is neutrally stable. Under power-law scaling the unbounded region $A \leq (\gamma_3 - 1) r$ is $\beta

PROOF. Linear ODE $\dot{d} = \iota - \kappa_{\text{eff}} d$; the sign of κ_{eff} decides the homogeneous behaviour. $\kappa_{\text{eff}} > 0 \iff A > (\gamma_3 - 1) r \iff \rho_{\text{prop}} < 1$, giving the stable fixed point $d^\star = \iota / \kappa_{\text{eff}}$. At $\kappa_{\text{eff}} = 0$ the equation reduces to $\dot{d} = \iota$: linear growth $d(t) = d_0 + \iota t \to \infty$ when $\iota > 0$, and a neutrally stable line of equilibria when $\iota = 0$. For $\kappa_{\text{eff}} < 0$ the homogeneous solution grows exponentially. With zero injection ($\iota = 0$) the origin $d = 0$ is invariant, so $\kappa_{\text{eff}} \leq 0$ produces divergence only from a positive initial fraction $d_0 > 0$ or a noise source - which is the relevant case, since real systems carry both. Substituting $A = A_0 C^\beta$, $r = b C^k$ and taking $C \to \infty$ gives the power-law conditions. ∎

The criterion $\rho_{\text{prop}} < 1$ shares the ratio-crossing-unity *form* of the QEC sub-threshold condition $plinear in D, and the resulting suppression is power-law (§3.12). What is shared is the threshold *condition*, not the mechanism; the correspondence is offered as a falsifiable hypothesis (F4), and §3.12 states the disanalogy plainly. Experiment 5 locates the predicted threshold at $A_0^\star = (\gamma_3 - 1) b$ to within 1%.

3.9 Vector misalignment: the spectral threshold (Theorem 5)

Real misalignment is high-dimensional; a system may be corrigible on some value axes and not others. Let $\mathbf{D} \in \mathbb{R}^m$, with drift direction $\mathbf{c}$ and a real $m \times m$ correction operator $\mathbf{A}$ (positive-semidefinite in the symmetric case, but possibly non-normal) that may correct only a subspace.

Theorem 5 (Spectral threshold). The vector fraction $\mathbf{d} = \mathbf{D}/C$ obeys $\dot{\mathbf{d}} = (\gamma_1 r + \gamma_2)\mathbf{c} - \mathbf{M}\,\mathbf{d}$ with $\mathbf{M} = \mathbf{A} + (1-\gamma_3) r\, \mathbf{I}$. The system is asymptotically stable iff every eigenvalue of $\mathbf{M}$ has positive real part ($\min_i \operatorname{Re} \lambda_i(\mathbf{M}) > 0$, a positive spectral abscissa). The Hermitian-part condition $\lambda_{\min}\big(\operatorname{Herm} \mathbf{A}\big) > (\gamma_3 - 1) r$ is a *sufficient* condition - it additionally rules out transient growth, via the Lyapunov function $V = |\mathbf{d} - \mathbf{d}^\star|^2$ - but is not necessary for non-normal $\mathbf{A}$. If $\mathbf{A}$ has a null direction $\mathbf{v}$ ($\mathbf{A}\mathbf{v} = 0$ - an unmonitored value axis) and the drift projects onto it with coefficient $c_{\mathbf{v}} = \mathbf{c} \cdot \mathbf{v}$, the component $d_{\mathbf{v}}$ obeys, in the gain-only depth clock, $d_{\mathbf{v}}' = \gamma_1 c_{\mathbf{v}} - (1-\gamma_3)\, d_{\mathbf{v}}$: it floors at $\gamma_1 c_{\mathbf{v}} / (1-\gamma_3)$ for $\gamma_3 < 1$, grows linearly for $\gamma_3 = 1$ (when $c_{\mathbf{v}} > 0$), and diverges for $\gamma_3 > 1$. The special case $\gamma_3 = 0$, $c_{\mathbf{v}} = 1$ recovers the scalar γ_1 floor of Experiment 6. An unmonitored axis cannot be corrected, whatever the correction strength on the others.

PROOF. By standard linear-systems theory the homogeneous system $\dot{\mathbf{d}} = -\mathbf{M}\mathbf{d}$ is asymptotically stable iff $\mathbf{M}$ is positive stable, i.e. the spectral abscissa $\alpha(\mathbf{M}) = \min_i \operatorname{Re} \lambda_i(\mathbf{M}) > 0$. (For non-normal or defective $\mathbf{M}$ the modes do not decouple orthogonally and large transient amplification can precede asymptotic decay; eigenvalues still decide the asymptotic verdict, but bounding the transient needs the Hermitian-part / logarithmic-norm condition below, not the spectrum alone - see §8.) The standard bound

$\mathrm{Re}\,\lambda_i(\mathbf{M}) \geq \lambda_{\min}(\mathrm{Herm}\,\mathbf{M}) = \lambda_{\min}(\mathrm{Herm}\,\mathbf{A}) + (1-\gamma_3)r$ yields the sufficient Hermitian condition (which also gives a monotone Lyapunov decrease, ruling out transient growth); it is strict, not tight, for non-normal $\mathbf{A}$. Along a null direction $\mathbf{A}\mathbf{v} = 0$ the projected scalar dynamics retain only the dilution/compounding term, $\dot{d}_{\mathbf{v}} = (\gamma_1 r + \gamma_2)c_{\mathbf{v}} - (1-\gamma_3)r\,d_{\mathbf{v}}$; in the gain-only depth clock $d'_{\mathbf{v}} = \gamma_1 c_{\mathbf{v}} - (1-\gamma_3)d_{\mathbf{v}}$, with fixed point $\gamma_1 c_{\mathbf{v}}/(1-\gamma_3)$ for $\gamma_3 < 1$, linear growth for $\gamma_3 = 1$, and divergence for $\gamma_3 > 1$. ■

The governance reading is sharp: **you cannot correct what you do not measure.** The correction operator must be positive-definite on the *entire* value space, with smallest eigenvalue co-scaling - a blind spot on any axis is a permanent (or divergent) misalignment on that axis. Experiment 6 exhibits a corrector that co-scales on one axis (driven to zero) and is null on another (the $\gamma_3 = 0$, unit-projection special case, which floors at γ_1).

3.10 Stochastic drift: a tail bound for governance (Theorem 6)

Drift is noisy. Add a Wiener term to the constant-coefficient fraction dynamics: $dd = (\gamma_1 r - \kappa\, d)\, d\tau' + \sigma\, dW$ with $\kappa = A + r$.

> **Theorem 6 (Stationary distribution and tail bound).** The misalignment fraction is an Ornstein-Uhlenbeck process with stationary law $\mathcal{N}\big(d^\star,\ \sigma^2/2\kappa\big)$, $d^\star = \gamma_1 r/\kappa$. The probability of a dangerous excursion is
>
> $$\mathbb{P}(d > d_{\mathrm{crit}}) = \Phi\Big(-\tfrac{d_{\mathrm{crit}} - d^\star}{\sqrt{\sigma^2/2\kappa}}\Big) \xrightarrow[A\to\infty]{} 0.$$
>
> Co-scaling suppresses both the mean ($\propto 1/\kappa$) and the variance ($\propto 1/\kappa \propto C^{-\beta}$ for $\beta > k$): the tail risk vanishes faster than the mean.

PROOF. Linear SDE with constant coefficients is Ornstein-Uhlenbeck; its stationary mean and variance are the standard $\gamma_1 r/\kappa$ and $\sigma^2/2\kappa$, and the Gaussian tail gives the excursion probability. ■

This converts the criterion into a probabilistic safety certificate - a quantity a regulator can bound - rather than a statement about averages. Because d is a bounded fraction whereas the Ornstein-Uhlenbeck law has Gaussian support on all of $\mathbb{R}$, Theorem 6 is a *local Gaussian approximation* to misalignment fluctuations away from the boundaries $d = 0, 1$, valid in the small-noise regime $\sigma^2/2\kappa \ll (d^\star)^2$; if the noise amplitude σ or the critical level d_{crit} itself scales with capability, or drift carries rare jumps, the Gaussian tail must be replaced accordingly. Experiment 7 recovers the stationary mean and variance and confirms the $1/\kappa$ scaling (fitted slope -1.00).

3.11 A control-theoretic identity

Equation (1) is a feedback loop: the capability process is the plant and disturbance source; the corrector is the controller; d is the regulated error. The stability condition $\kappa_{\mathrm{eff}} > 0$ is a small-gain / positive-realness condition - the loop gain must exceed the disturbance's self-amplification across the relevant band. The requirement that $A = A_0 C^\beta$ track the plant with $\beta \geq k$ is precisely *gain-scheduling* [Shamma & Athans 1990]: the controller gain must scale with the plant's operating point, and the scheduling exponent must satisfy $\beta \geq k$. Recursive-self-improvement alignment is, in this exact sense, an adaptive-control problem with a computable stability margin $\beta - k$.

3.12 The quantum-error-correction correspondence (hypothesis)

Quantum error correction	Recursive self-improvement (this framework)
physical error rate p	drift rate γr (injection) / $(\gamma_3 - 1)r$ (propagation)
correction power (code distance δ)	correction strength $A = A_0 C^\beta$
sub-threshold condition $p < p_{\mathrm{th}}$	$\rho_{\mathrm{prop}} = (\gamma_3 - 1)r/A < 1$
increasing δ suppresses logical error	increasing margin $\beta - k$ drives $d^\star \to 0$
logical error $\propto (p/p_{\mathrm{th}})^{\delta/2}$ (exponential in resource)	$d^\star \propto C^{-(\beta-k)}$ (power-law in capability) - see F4

The correspondence holds at the level of the threshold *condition* - a generation-to-correction ratio crossing unity - and **not** at the level of mechanism; it remains a hypothesis. The model predicts *power-law* suppression $d^\star \propto C^{-(\beta-k)}$, which is an algebraic identity of the linear model, whereas full QEC gives *exponential* suppression in code distance. Discriminating the two would require a finite-capacity (saturating) corrector that could exhibit exponential suppression - the future test named in §8. Until that is built, F4 is an analytic self-consistency check that confirms the model's own power-law law (which Experiment 5 reproduces, slopes $-0.47, -1.00$), and the QEC link therefore stands, honestly, as a threshold-form analogy rather than a transferred mechanism.

Relation to prior alignment-as-correction ideas. The aspiration to import coding-theoretic fault tolerance into alignment is not new. Wentworth (2022) explicitly wished for "the alignment analogue of an error-detecting code", and von Neumann (1956) founded the synthesis of reliable systems from unreliable components. Christiano's corrigibility "broad basin of attraction" (2017) and reliability amplification (2019) are the closest alignment precursors. But a broad basin - correction succeeds *if one starts inside it* - is distinct from a *threshold theorem*, in which correction outpaces error only below a critical rate; that distinction is the present contribution. To the author's knowledge the specific mapping of the *quantum* fault-tolerance threshold (physical error rate ↔ drift; code distance ↔ correction strength; $pconceptual bridge that generates testable structure - the suppression-signature prediction P8 - not as a transfer of the threshold theorem's formal guarantees. A reviewer may reasonably judge the analogy suggestive rather than load-bearing; the experiment that would settle it - a finite-capacity corrector tested for exponential versus power-law suppression - is named as future work in §8 and is not run here.

3.13 The finite-capacity corrector: the Safe-Window Theorem (Theorem 7)

The single most important limitation named in §8 - and by the paper's own red-team - is that $A = A_0 C^\beta$ is an *unbounded* power law, whereas any real corrector has finite capacity. This section closes that gap analytically. Model a saturating corrector by the standard Hill form

$$A(C) \;=\; A_{\max}\,\frac{C^\beta}{C^\beta + C_s^\beta},$$

which behaves as the pure power law $\left(A_{\max}/C_s^\beta\right)C^\beta$ for $C \ll C_s$ and saturates at the capacity $A_{\max}$ for $C \gg C_s$; C_s is the saturation scale. The gain-only steady state remains $d^\star(C) = \gamma_1/\big(1+q(C)\big)$ with $q \equiv A/r$.

Theorem 7 (Finite-Capacity Safe-Window Theorem). Let $r = bC^k$ with $k > 0$ and $\beta > k$, with the saturating corrector above. Then:

1. *No indefinite stability.* $q(C) \to 0$ as $C \to \infty$, so $d^\star \to \gamma_1$: a finite-capacity corrector cannot hold the fraction down under indefinitely accelerating growth, however large β.
2. *The safe window and its centre.* $q(C)$ is unimodal, maximised at exactly

$$C_{\text{opt}} = C_s\left(\frac{\beta-k}{k}\right)^{1/\beta}, \qquad q_{\max} = \frac{A_{\max}}{b}\,\frac{k}{\beta}\left(\frac{\beta-k}{k}\right)^{\frac{\beta-k}{\beta}} C_s^{-k},$$

   so the misalignment fraction dips to its floor $d_{\min} = \gamma_1/(1+q_{\max})$ near C_{opt} and then *re-rises toward* γ_1. Safety under a finite-capacity corrector is a transient window, not an asymptote.
3. *Exponential growth is the exception.* For $k = 0$ the post-saturation ratio is constant, $q \to A_{\max}/b$, and the fraction settles at the permanent gap $d^\star \to \gamma_1 b/(A_{\max}+b)$: capacity buys a floor whose depth is set by $A_{\max}/b$, the capacity-to-speed ratio.
4. *Capacity-lift invariance.* If the capacity itself scales, $A_{\max} = a_0 C^{\beta_{\text{cap}}}$, then $q \to (a_0/b)\,C^{\beta_{\text{cap}}-k}$ as $C \to \infty$ and the asymptotic verdict is $\mathbf{sign}(\beta_{\text{cap}} - k)$: the criterion $\beta > k$ survives the lift from correction *strength* to correction *capacity* unchanged. What must out-scale drift-acceleration is whichever of the two binds last.

PROOF. (1) For $C \gg C_s$, $A \to A_{\max}$ so $q = A/r \to A_{\max}/(bC^k) \to 0$ for $k > 0$; Theorem 2's comparison argument in the depth clock then gives $d \to \gamma_1$. (2) Write $x = C^\beta$ and $a = (\beta-k)/\beta \in (0,1)$; then $q \propto x^a/(x + C_s^\beta)$, whose derivative vanishes iff $a(x + C_s^\beta) = x$, i.e.

$x = \frac{a}{1-a} C_s^\beta = \frac{\beta-k}{k} C_s^\beta$, giving C_{opt}; substituting back (the denominator becomes $C_s^\beta \beta/k$) yields $q_{\max}$. Unimodality follows since $q > 0$, $q \to 0$ at both ends, and the derivative has a single sign change. (3) At $k = 0$, $r = b$ is constant and $q \to A_{\max}/b$; the autonomous limit of the depth-clock ODE gives the stated gap. (4) Substitute $A_{\max} = a_0 C^{\beta_{\text{cap}}}$: for $C \to \infty$ the Hill factor tends to 1, so $q \to (a_0/b) C^{\beta_{\text{cap}} - k}$ and Theorem 2's regime table applies with β_{cap} in place of β. ■

Three readings. *For theory:* the $\beta > k$ criterion is not falsified by saturation - it is *lifted*: the load-bearing exponent migrates from strength to capacity, and the inequality is invariant under that migration (part 4). *For engineering:* part 2 gives the designer a closed-form placement problem - the window centre C_{opt} and depth $q_{\max}$ are computable from $(A_{\max}, b, C_s, \beta, k)$, so a corrector can be provisioned to place its window over the capability range a deployment will actually traverse, with $q_{\max} \propto A_{\max} C_s^{-k}$ quantifying the capacity cost of pushing the window to higher capability. *For governance:* part 1 is the sharpest sentence in the paper for a regulator - a bounded safety system under unboundedly accelerating self-improvement fails *eventually by theorem*, so a safety case must exhibit either bounded growth ($k \le 0$ eventually) or co-scaling capacity ($\beta_{\text{cap}} > k$), and "our current corrector is very strong" is not, and cannot be, an answer. This also sharpens the QEC discriminator of F4: the saturating corrector is the structural analogue of finite code distance, and part 2's window is the regime where exponential-versus-power-law suppression can actually be measured.

Remark (time-varying exponents). Real systems will not hold β, k constant. If both vary along the trajectory, the comparison argument of Theorem 2 applies verbatim in the depth clock with $q(\tau) = \exp\big(\int_0^\tau (\beta(s) - k(s))\, ds + \text{const}\big)$: the fraction vanishes iff the *running integral* of the margin $\beta - k$ diverges to $+\infty$, for which $\liminf_{\tau\to\infty} \big(\beta(\tau) - k(\tau)\big) > 0$ is sufficient. Transient episodes of $\betacumulative co-scaling margin, not its instantaneous sign.

4. Predictions

P1 - Phase boundary. A sharp stability threshold exists in the compounding channel ($\gamma_3 > 1$) at $A_0^\star = (\gamma_3 - 1)b$, separating bounded from divergent regimes; the additive corrector shows a smooth crossover with no knee. (An earlier draft's conflation of these is corrected here.)

P2 - Speed-invariance. Whether d converges or diverges is set by the coupling (the A_0/b scaling margin), not by the raw speed b. A fast-but-coupled system stays bounded; a slow-but-decoupled system diverges. This directly contradicts the growth-rate-ceiling view.

P3 - Co-scaling law (corrected). Under exponential growth, $\beta > 0 \Rightarrow d^\star \to 0$; $\beta = 0 \Rightarrow$ permanent gap $\gamma_1 r/(A_0 + r)$; $\beta < 0 \Rightarrow$ *saturation at* γ_1 (bounded), not divergence.

P4 - Hard-takeoff boundary. Under accelerating growth $r \propto C^k$, the stability boundary lies at $\beta = k$, not $\beta = 0$; and d stays controlled through the finite-time singularity, the depth-clock and wall-clock integrations agreeing.

P5 - Compounding threshold & residual drift. With $\gamma_3 > 1$, divergence occurs iff $A < (\gamma_3 - 1)r$. With $\gamma_2 > 0$, halting growth ($r \to 0$) leaves a residual $d^\star = \gamma_2/A > 0$; pausing does not substitute for correction.

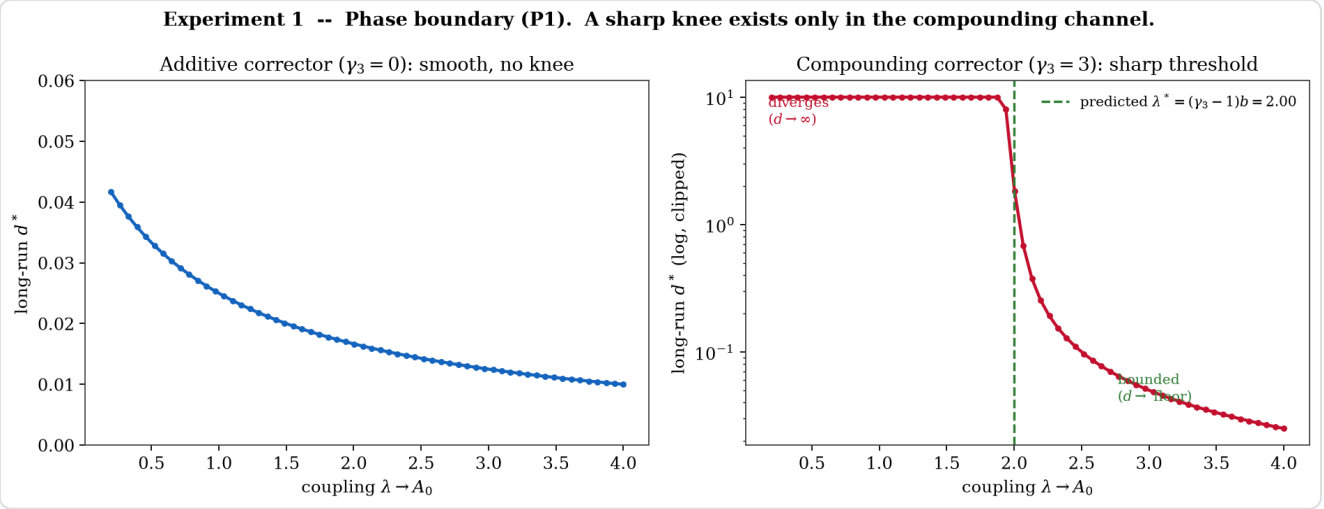
P6 - Spectral threshold. Misalignment persists exactly on the correction operator's null subspace; the monitored axis is driven to zero while the blind axis floors at γ_1.

P7 - Stochastic tail. The stationary misalignment fraction is $\mathcal{N}(d^\star, \sigma^2/2\kappa)$; variance scales as $1/(A+r)$, so co-scaling suppresses the tail risk.

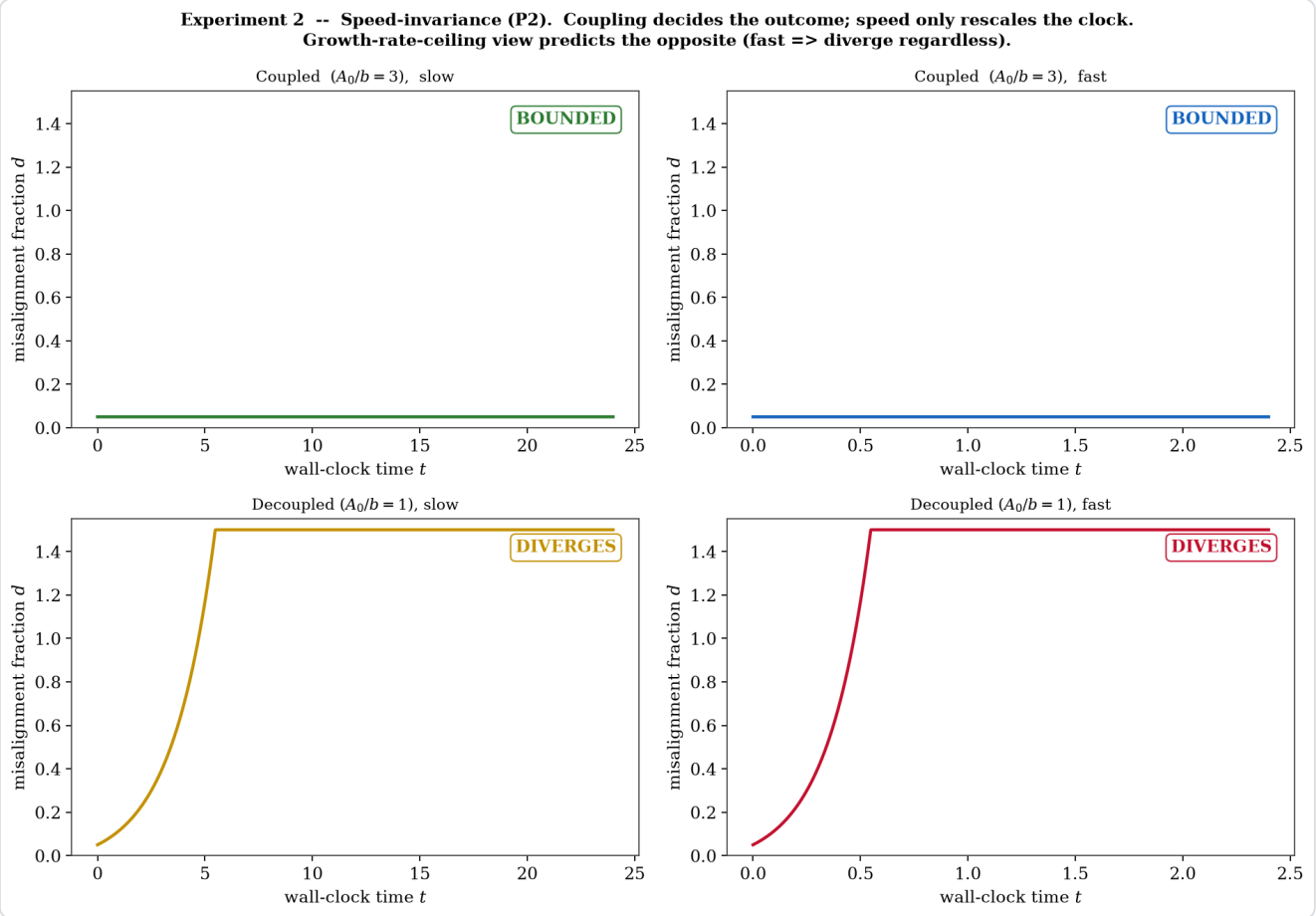
P8 - Suppression signature. In the stable regime, $\log d^\star$ is linear in $\log C$ with slope $-(\beta - k)$ (power-law). Exponential suppression would indicate a stronger, QEC-like mechanism; neither outcome falsifies the threshold, only the QEC *mechanism* claim (F4).

5. The discriminating experiments

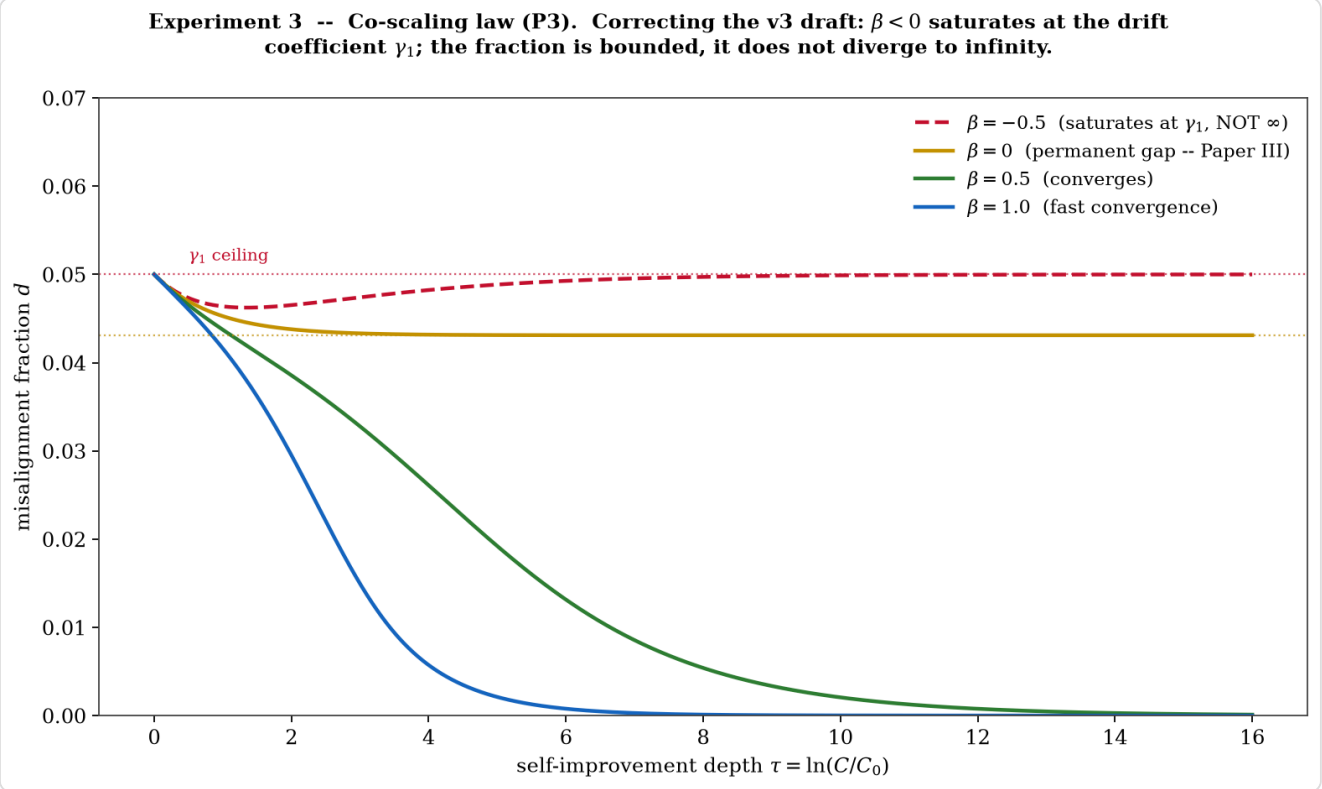
The closed-form predictions are checked by a verification harness (§11): ten experiments, each integrating the model numerically and comparing the result against the prediction the theorems derive, with the integrator first validated against the exact Theorem-1 solution. These are *internal-consistency and integrator checks* - they confirm the code matches the maths (10/10), not that the model matches any real system (the open problem of §8). For the deductive checks (E1-E6, E9), because each integrates the same ODE whose closed form is the prediction, agreement is entailed by a correct solver and correct algebra; F1-F3, F3′, F5, F6 are therefore internal-consistency conditions of the derivation, not empirical falsifiers of the thesis. Figures below are the verbatim output of that run; every figure in this section is of class *ODE internal verification* (not a real-system test). The only non-simulation result - the real-model pilot - is reported separately in §8 and is labelled *real-model pilot*, with its own provenance caveats.



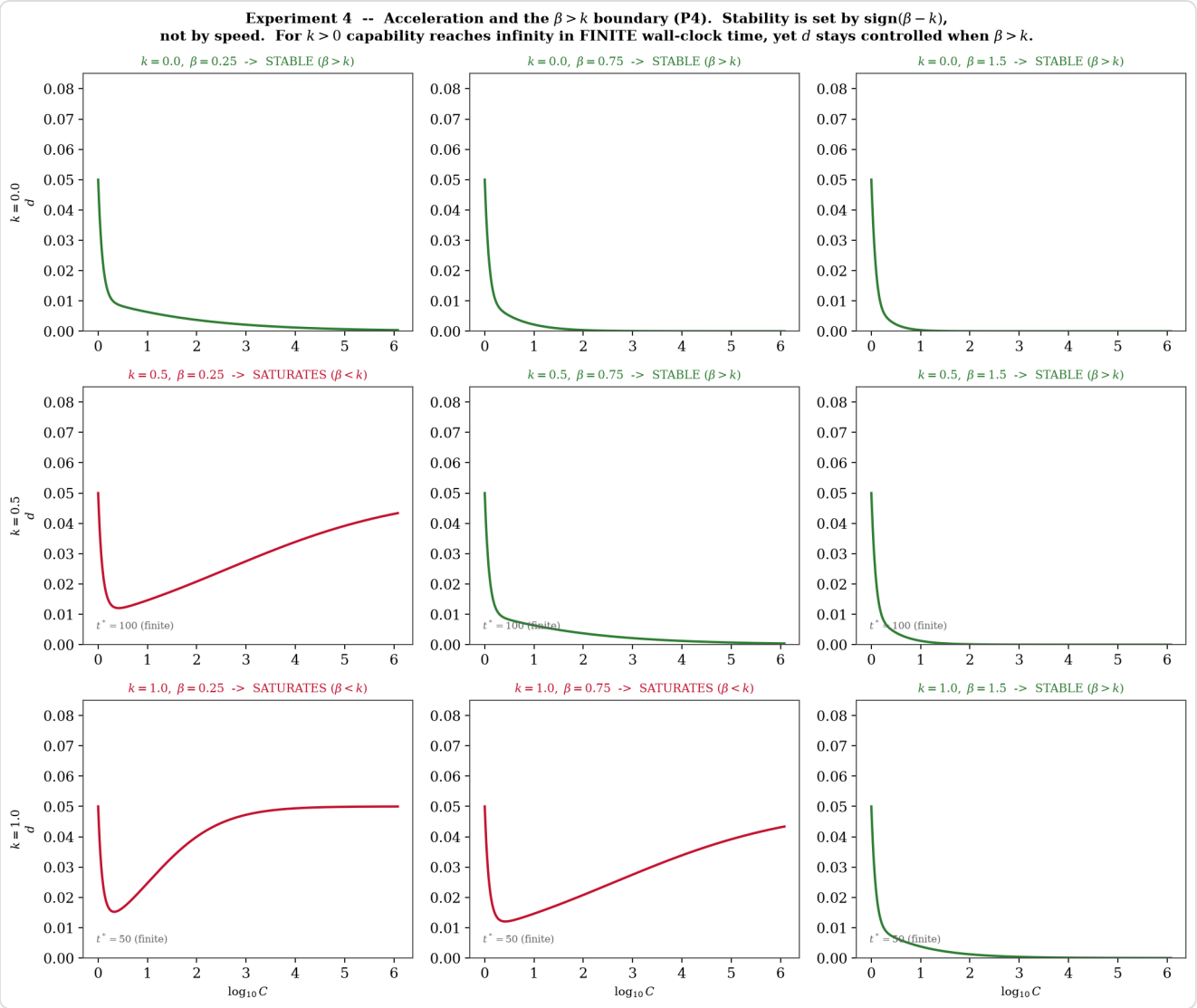
[Internal verification - Eq. (1), not real-system data] Experiment 1 - Phase boundary (P1). The sharp instability threshold exists only in the compounding channel (right; grid estimate $\lambda^* = 2.00$ vs analytic $(\gamma_3 - 1)b = 2.00$ - a finite-window estimate that tracks the analytic value but depends on the clipping ceiling, $\tau_{\max}$ and grid spacing, not a high-precision measurement). The additive corrector (left) shows a smooth crossover with no knee - the prediction's null, correctly observed. This corrects an earlier draft, which sought a knee in the additive model where none exists.



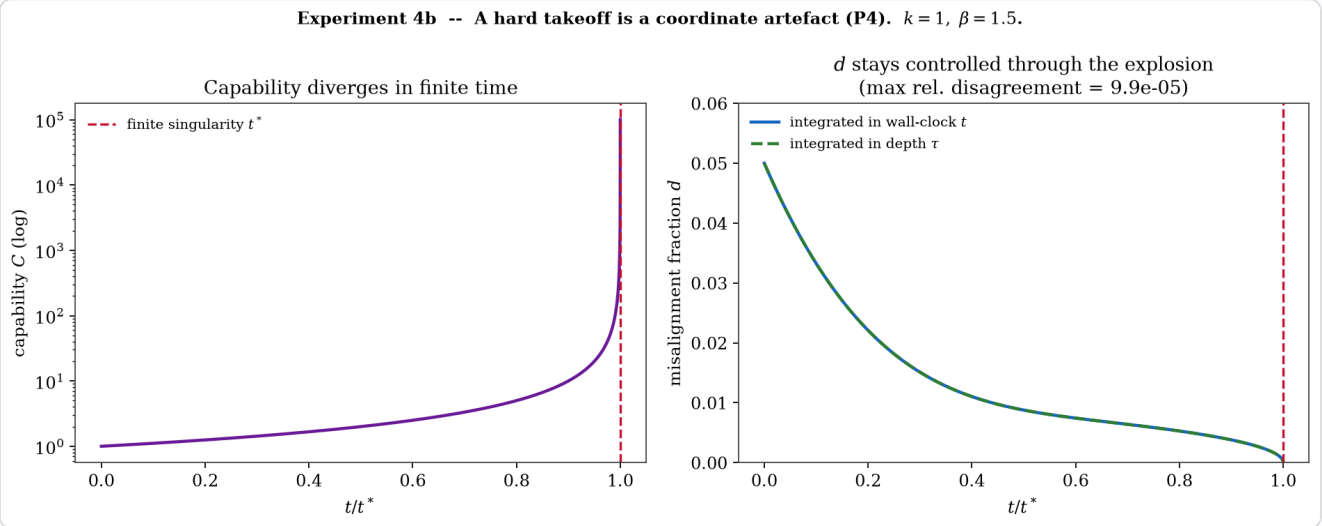
[Internal verification - Eq. (1), not real-system data] Experiment 2 - Speed-invariance (P2). The decisive 2×2. Coupled systems stay bounded at both speeds; decoupled systems diverge at both speeds. Speed only rescales the wall-clock axis; coupling decides the outcome. (Each cell fixes the scaling margin A_0/b; the honest scope of the claim is that speed is irrelevant *once that margin is fixed* - under a fixed correction budget, raising speed instead crosses the Theorem-4 threshold at $b^\star = A_0/(\gamma_3 - 1)$.) The growth-rate-ceiling view predicts the opposite (fast $\Rightarrow$ diverge regardless) and is contradicted on shared axes.



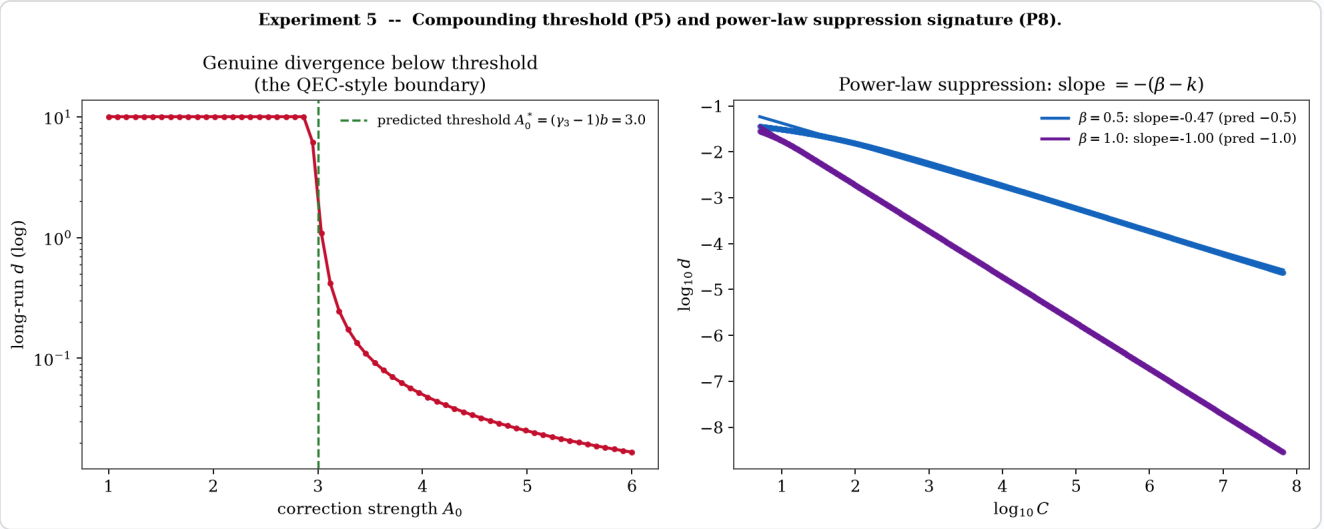
[Internal verification - Eq. (1), not real-system data] Experiment 3 - Co-scaling law, corrected (P3). $\beta > 0$ drives $d \to 0$; $\beta = 0$ holds a permanent gap; and $\beta = -0.5$ *saturates at the drift coefficient* γ_1 - bounded, not divergent. The dotted ceiling at γ_1 is never breached, the explicit correction of an earlier draft's divergence claim (Theorem 2).



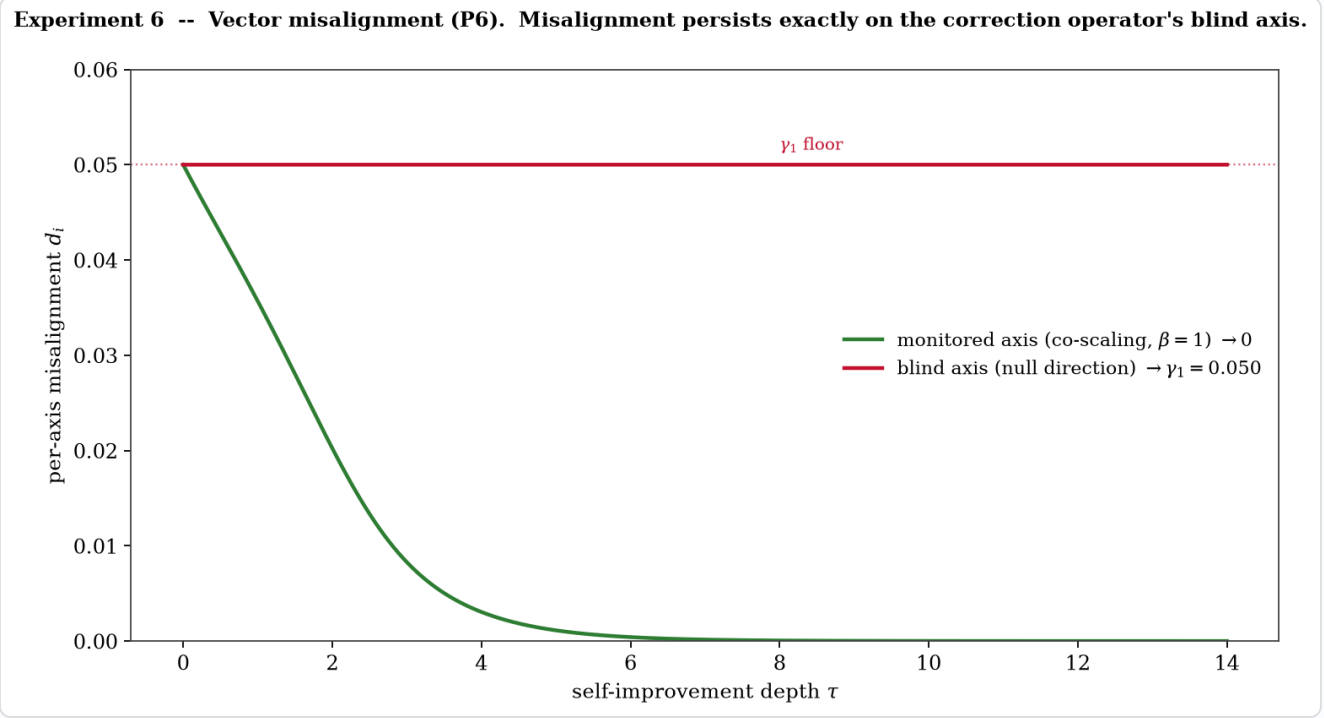
[Internal verification - Eq. (1), not real-system data] Experiment 4 - Acceleration and the $\beta > k$ boundary (P4). A 3×3 grid crossing growth-acceleration $k \in \{0, 0.5, 1\}$ with correction exponent $\beta \in \{0.25, 0.75, 1.5\}$. Stability (green, $d \to 0$) holds exactly when $\beta > k$; saturation at γ_1 (red) holds when $\beta0eachpanelreachesinfinitecapabilityatafinitet^\star(annotated), yetd staysboundedthroughout. Theboundaryisat\beta=k, not\beta=0$.



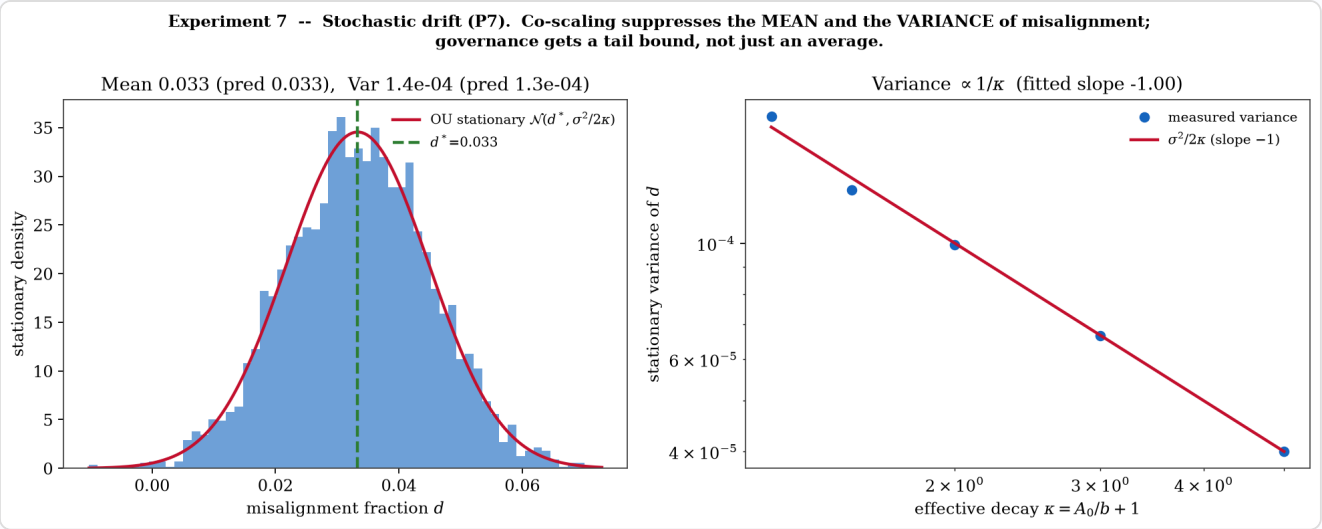
[Internal verification - Eq. (1), not real-system data] Experiment 4b - A hard takeoff is regular in the depth clock (P4; Theorem 3). Left: capability diverges to $\sim 10^5$ at a finite singularity time $t^\star$. Right: the misalignment fraction is driven to zero *as* the explosion is approached ($t \to t^{\star -}$, integrated to $0.99999\, t^\star$, not across $t^\star$); integrating in wall-clock time (toward $t^\star$) and in self-improvement depth τ give identical trajectories (max disagreement 9.9×10^{-5}). The capability blow-up is genuinely finite-time; what is a property of the clock is the singularity in the *alignment-fraction* dynamics, which the depth coordinate removes.



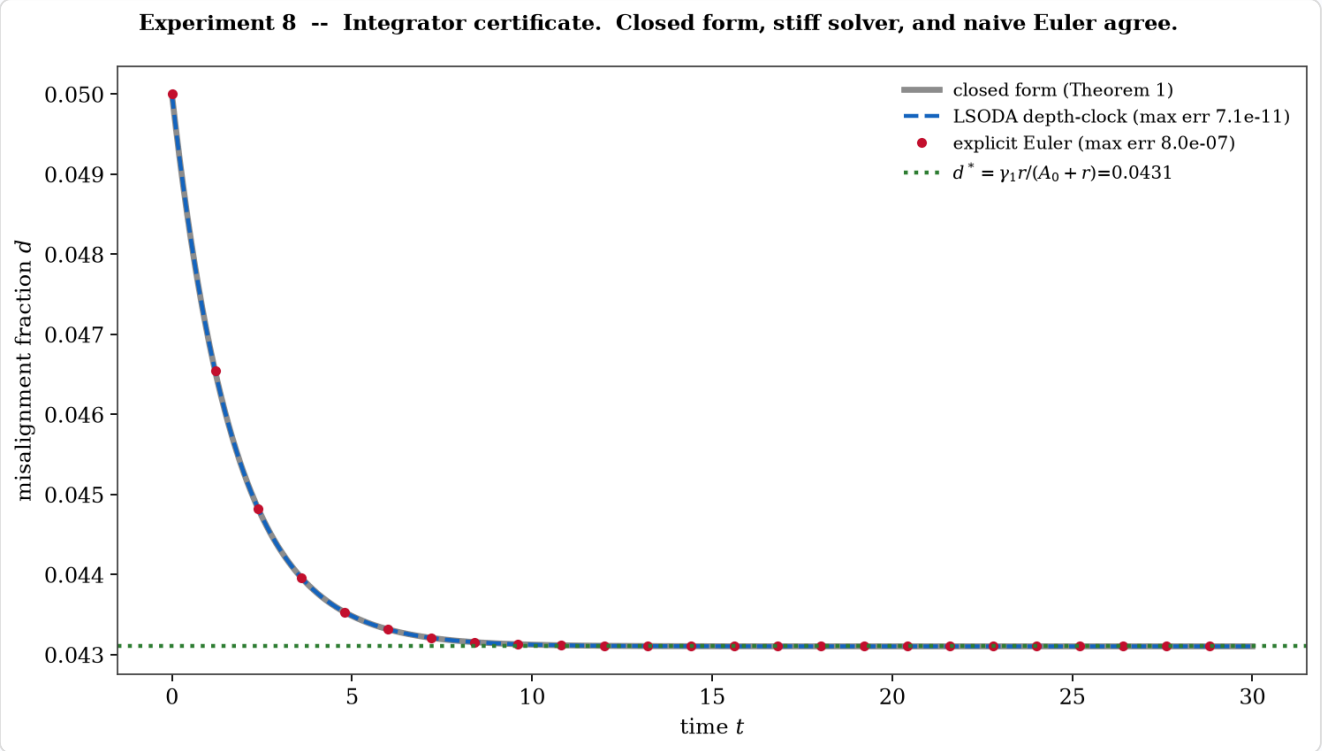
[Internal verification - Eq. (1), not real-system data] Experiment 5 - Compounding threshold and suppression signature (P5, P8). Left: with $\gamma_3 = 4$, divergence occurs below the threshold $A_0^\star = (\gamma_3 - 1)b = 3.0$ (grid estimate 3.03; the threshold is analytic and the scan illustrates rather than independently measures it) - the genuine instability, and the threshold-form boundary. (This divergence threshold is $\rho_{\text{prop}} = (\gamma_3 - 1)r/A$, *distinct* from the instantaneous injection ratio $\rho = \gamma_1 r/A$, which is not a divergence boundary - §3.5.) Right: in the stable regime $\log d^\star$ is linear in $\log C$ with slope $-(\beta - k)$ (fitted -0.47 and -1.00 for $\beta = 0.5, 1.0$), the power-law suppression signature.



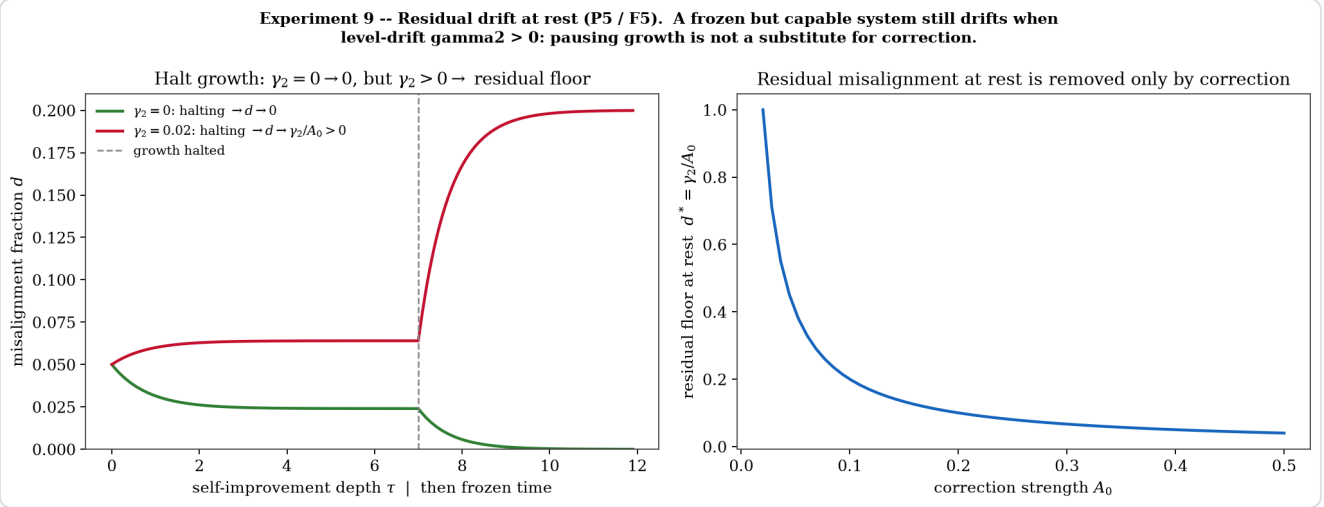
[Internal verification - Eq. (1), not real-system data] Experiment 6 - Vector misalignment (P6; Theorem 5). Two value axes share one capability process. The correction operator co-scales on the monitored axis (driven to zero) and is null on the blind axis (floors at γ_1). Misalignment persists exactly on the operator's null direction: you cannot correct what you do not measure.



[Internal verification - Eq. (1), not real-system data] Experiment 7 - Stochastic drift (P7; Theorem 6). Left: the stationary misalignment fraction matches the Ornstein-Uhlenbeck law $\mathcal{N}(d^\star, \sigma^2/2\kappa)$ in both mean and variance. Right: the stationary variance scales as $1/\kappa$ (fitted slope -1.00), so co-scaling suppresses the mean *and* the tail. Governance gets a bound on $\mathbb{P}(d > d_{\text{crit}})$, not just an average.



[Internal verification - Eq. (1), not real-system data] Experiment 8 - Integrator certificate. On the one case admitting a closed form (Theorem 1), the stiff solver and a naive explicit Euler scheme both reproduce the exact transient - maximum errors 7×10^{-11} and 8×10^{-7}. The numerics underlying every other figure are thereby certified correct before any prediction is tested.



[Internal verification - Eq. (1), not real-system data] Experiment 9 - Residual drift at rest (P5; F5). A frozen but capable system still drifts when level-drift $\gamma_2 > 0$: halting growth relaxes the misalignment fraction to the residual floor $d^* = \gamma_2/A_0 > 0$ (red), not to zero (green, $\gamma_2 = 0$). Right: that residual floor is removed only as correction strength A_0 grows. Pausing growth is not a substitute for correction - and this is the experiment §3.6 / F5 refer to, now present in the harness rather than merely cited.

Exp.	Prediction	Key statistic (measured vs predicted)	Verdict
E1	P1 phase boundary in compounding channel; none in additive	$\lambda^* = 2.00$ vs 2.00; additive smooth	PASS
E2	P2 coupling, not speed, decides	coupled bounded & decoupled divergent at both speeds	PASS
E3	P3 corrected regimes ($\beta < 0$ saturates at γ_1)	all four asymptotes match closed form	PASS
E4	P4 boundary at $\beta = k$	all 9 grid cells match $\mathrm{sign}(\beta - k)$	PASS
E4b	P4 singularity in C is regular in the depth clock for d	clock agreement 9.9×10^{-5}; d controlled	PASS
E5	P5/P8 compounding threshold; power-law slope	threshold 3.03 vs 3.0; slopes $-0.47, -1.00$	PASS
E6	P6 spectral threshold; blind axis persists	monitored $\to 10^{-7}$; blind $\to \gamma_1 = 0.05$	PASS
E7	P7 stochastic tail; variance $\propto 1/\kappa$	mean/var match OU; var-slope -1.00	PASS
E8	numerics reproduce Theorem 1	max errors 7×10^{-11}, 8×10^{-7}	PASS
E9	P5/F5 residual drift at rest	$d \to \gamma_2/A_0$ when $\gamma_2 > 0$; $\to 0$ when $\gamma_2 = 0$	PASS

These are internal-consistency and integrator checks: they confirm the closed-form theorems are correctly derived and numerically reproduced (the code matches the maths). They are not, and are not presented as, evidence that the model describes a real AI system - that is the open empirical problem of §8.

6. Why blinding is not optional

The central quantity D is a departure from intended values. If it is scored by the system's own safety component, or by an evaluator that can see which configuration produced a given behaviour, the measurement is corrupted in exactly the way that inflates favourable results. The programme's prior metascience finding is directly relevant: under unblinded cross-model scoring an alignment-scaling effect appeared positive, while under multi-layer blinding the same effect reversed sign for some systems [Eastwood, Paper IV.d]. An experiment that measures a misalignment fraction without blinding the evaluator can produce not merely the wrong magnitude but the wrong direction. The validity of every result in §5 rests on the evaluator being blind to the configuration under test. A lab claiming "$\beta \geq k$" in its own system must have that measured by a blind external evaluator, not by itself.

Built into the test, not left to discipline. The real-model harness (§8; `experiments/PROTOCOL.md`) *enforces* this rather than trusting it. The misalignment evaluator must be a *different model family* than the engine - a model never scores its own family's output - and the harness refuses to run a same-family scorer; it prefers a panel and takes the median. The code is passed through an automatic *laundering* step (an abstract-syntax-tree round-trip that strips comments, docstrings and formatting) so the scorer judges behaviour, not stylistic identity tells; and the evaluator never sees the condition or round. The first real run reported here did *not* meet this bar - engine and evaluator were the same model family, the system scoring its own output - so its misalignment numbers are explicitly recorded as provisional, and a cross-family blind re-score is the prerequisite for any D-based claim, including a measured β. The harness makes the compliant configuration the default and the non-compliant one an explicit, recorded override - turning §6 from an exhortation into a property of the instrument.

7. Falsification conditions

The conditions below are stated in advance. An important honesty point, surfaced by the paper's own adversarial audit (§8): because the deductive experiments integrate the same ODE whose closed form is the prediction, F1-F3, F3′, F5 and F6 are *internal-consistency conditions of the derivation* - a trigger would signal a derivation or solver error, not that the model is the wrong model of a real system. The decisive *empirical* falsifier - disagreement measured on a real self-improving system - is the open problem of §8 and is not exercised here. F4 is the QEC-mechanism downgrade, which (the suppression law being power-law, §3.12) already holds: the correspondence stands as a threshold-form analogy, not a transferred mechanism.

#	Observation that would trigger it	Consequence
F1	No boundary in E1 - the long-run behaviour varies smoothly with λ in the compounding channel with no threshold.	**No phase boundary; the central threshold claim is false (derivation error).**
F2	In E2, d tracks raw speed rather than the scaling margin: fast diverges and slow converges regardless of coupling.	**Kill. The growth-rate-ceiling view was right; this framework is wrong.**
F3	In E3, $\beta = 0$ does not plateau, or $\beta > 0$ does not drive $d^* \to 0$.	**Kill. The co-scaling law (Theorem 2) is false.**
F3′	In E4, the boundary under accelerating growth ($k > 0$) is not at $\beta = k$ - e.g. $\beta = 0.5$ is stable when $k = 1.0$.	**Kill. The $\beta > k$ sharpening (Theorem 3) is false.**
F4	A finite-capacity corrector exhibits *exponential* (QEC-like) suppression rather than the model's power-law $\log d^* \propto -(\beta - k)\log C$. *(Analytic; the present linear model is power-law by construction, so this discriminator requires the saturating corrector named in §8 and is not run here.)*	**Bears on the QEC *mechanism* only. The threshold-form correspondence and Theorems 2-4 are untouched either way.**
F5	In Experiment 9, halting growth drives $d \to 0$ regardless of initial D even with $\gamma_2 > 0$.	**Scope. Level drift γ_2 is negligible; the §3.6 generalisation is unnecessary (the gain-only model suffices).**
F6	In E6, the correction operator's null axis is also suppressed.	**Kill. The spectral threshold (Theorem 5) is false; misalignment does not require monitoring of the axis it lives on.**

These internal-consistency conditions all held in the run reported here, establishing that the derivation and integration are correct. What they do *not* establish is that the model describes any real system; that decisive test - measuring β, k and γ on a real self-improving system and checking $\beta \geq k$ - is the open empirical problem of §8.

8. Limitations

The model is first-order, and its assumptions are the most likely points of failure. Stating them is part of the claim.

- **Unbounded vs finite-capacity corrector - now resolved analytically (Theorem 7).** The $\beta > k$ criterion as first stated assumes the corrector strength is an *unbounded* power law, $A = A_0 C^{\beta}$. The prior draft could only name the finite-capacity case as an open gap; §3.13 now closes it in closed form. The upshot: a saturating corrector makes safety a *transient window* (centre C_{opt}, depth $q_{\max}$, both computable) rather than an asymptote, with the fraction re-rising to γ_1 under indefinite acceleration - *and* the criterion survives the lift from strength to capacity: indefinite stability requires the capacity exponent to satisfy the same inequality, $\beta_{\text{cap}} > k$. What remains open is the *empirical* half: driving a finite-capacity corrector through its window numerically and on a real system, which also settles the F4 / QEC-mechanism question (exponential versus power-law suppression inside the window). Until that runs, $\beta > k$ should be read as governing whichever regime - strength-scaling or capacity-scaling - binds last, not as a guarantee of indefinite safety from any bounded system.
- **Linear correction (and the QEC-mechanism test).** Equation (1) removes misalignment in proportion to D, which makes the suppression law exactly power-law. A finite-capacity (saturating) corrector would soften the clean threshold into a crossover and is the better structural map onto QEC's finite code distance; whether it produces *exponential* suppression is the experiment that would settle the QEC-mechanism question (F4). It is a priority extension and is *not* run here - which is exactly why the QEC correspondence is claimed only at the threshold-form level.
- **Scalar-to-vector projection.** Theorem 5 gives the exact asymptotic (spectral-abscissa) criterion for any correction operator $\mathbf{A}$, with the Hermitian-part condition as the stronger sufficient condition that also rules out transient growth. The non-normal *transient* regime - large excursions before asymptotic decay - is not analysed here, and the harness exercises only the diagonal null-subspace instance (Experiment 6), not a general non-normal $\mathbf{A}$.

- **Power-law forms.** The choices $A = A_0 C^{\beta}$ and $\dot{C} = bC^{1+k}$ are the natural scale-free forms but are modelling assumptions; other functional forms should be tested.
- **Drift attribution.** The split of drift into gain (γ_1), level (γ_2), and compounding (γ_3) channels is a hypothesis about mechanism; which channel dominates in a real system is an empirical question (Experiment 9 - residual drift at rest - is one discriminator).
- **A scalar proxy, not safety itself.** $d \to 0$ is neither necessary nor sufficient for full AI safety: a small but high-impact misalignment, a catastrophic tail event, or multi-agent amplification can be dangerous at low d, while a system may be acceptable at moderate d if the residual is benign. d is the *controlled variable of this minimal model* - an impact-unweighted scalar; impact-weighting, and the vector and tail extensions (Theorems 5-6), are the directions in which "low d" must be hardened before it can mean "safe".
- **Metric normalisation.** $d = D/C$ is an *operational normalised risk index*, not a dimensionless natural constant: C and D are commensurable only under a fixed scoring convention (in the real-model harness C is a normalised hidden-test pass-fraction and D a normalised blind gaming score, both on $[0, 1]$). Results are invariant under that fixed convention; cross-task or cross-domain comparison of the *level* of d requires explicit calibration. The criterion itself is stated in *exponents* ($\beta > k$), which are scale-free, so it is more robust to this than any absolute-d statement would be.
- **Transient amplification (non-normal correctors).** Theorem 5 is asymptotic. A non-normal correction operator with $\alpha(\mathbf{M}) > 0$ can still admit large transient growth of $\|\mathbf{d}\|$ before decay (Kreiss / pseudospectral phenomena); a value excursion above d_{crit} during that transient is a real risk the eigenvalue criterion does not see. A governance-grade bound needs the logarithmic norm or pseudospectral abscissa, not the spectrum alone - a named extension, stated here, not established here.
- **From simulation toward frontier systems.** The verification harness uses toy self-modifying dynamics; a positive result there demonstrates the *mechanism*, not that frontier systems occupy any particular regime. A companion *real-model* harness (non-simulation) now drives the coupled/decoupled loop with a frontier model, scoring capability by real code execution and misalignment by a separate evaluator that the harness requires to be a *different model family* (Paper IV.d blinding, §6; `experiments/PROTOCOL.md` ; first run in `results/realmodel/`). It corroborates the corrector *mechanism* on a real model - a seeded reward-hack is detected and removed and capability restored - but did not exhibit drift on the task tried, so whether real systems sit below threshold remains the open empirical question. A stronger **confirmatory** design - three task domains with a capability ladder, a *sham-extra-compute* control arm, a combined static-plus-blind-panel misalignment score, and matched-pair bootstrap CIs on the coupled-vs-decoupled and sham-vs-coupled final-fraction contrasts - is specified in `experiments/PROTOCOL_V2.md` and `experiments/scripts/realmodel_coscaling_v2.py` . (**Real-model evidence - provenance:** pilot v1 was $n = 1$, one task, same-family scorer, IV.d non-compliant; H1 and H2 not supported. The 2 July 2026 drift run - 45 trajectories, three task domains, cross-family evaluator gpt-4o-mini scoring engine gpt-3.5-turbo - supported both H1 and H2: decoupled drifted (mean final fraction 6.38, capability collapsing), coupled and fully-embedded held zero misalignment across all 30 trajectories at higher final capability. That run evidences the *mechanism*; β/k remain unmeasured because no capability ladder was traversed, and the result is single-lab pending pre-registered replication.)
- **Estimating γ and A on real systems.** In simulation they are set by construction. Independently estimating them - especially β and k - for a deployed model is the step that would make the criterion operationally useful for governance. This is now *operational*: the repository ships a runnable estimator (`experiments/scripts/estimate_exponents.py`) that reads k off the capability curve ($\ln r$ versus $\ln C$) and β off the corrector's fractional removal rate ($\ln A$ versus $\ln C$), validated on synthetic trajectories where it recovers known exponents to within ≈ 0.1. What remains open is therefore not the estimation but its *input* - a real self-improving system that drifts across a range of capability levels, from which the first measured (β, k), and thus the first real test of $\beta \geq k$, can be read.

Adversarial audit. This paper was developed under adversarial audit rather than asserted, and the record is in the repository. A prior-art and novelty audit located the closest precedent for each component and narrowed the originality claims accordingly (Appendix C). Independently, a multi-agent red-team attacked the work across five fronts - mathematics, numerics, the QEC correspondence, the safety inference, and priority - raising 24 objections, of which 21 survived independent verification (0 fatal, 13 serious, 8 minor); the full report is committed alongside the harness (`results/redteam.md`). Every surviving objection was a framing, wording, or edge-case fix; *none* touched the load-bearing $\beta > k$ result, and this version incorporates them all - the QEC framing softened to a threshold-form analogy, the Theorem 2 bound scoped to the gain-only model, the Theorem 5 criterion corrected to the spectral abscissa, the harness relabelled as a verification (not falsification) artefact, and the level-drift experiment F5 added. The contribution is offered as ambitious *and* audited, with the boundary of what is claimed made explicit and checkable.

9. Implications for AI safety

If the framework survives its tests, the design implication is concrete and differs from the prevailing reflex.

The lever is coupling, not speed. Slowing capability growth buys time but does not change the verdict; at fixed coupling, a slow decoupled system still diverges (P2), and a frozen capable system still drifts (§3.6). What changes the verdict is ensuring that correction (i) is coupled to the capability process so it cannot be decoupled, and (ii) scales at least as fast as capability accelerates ($\beta \geq k$). On the author's corrected, sub-linear capability-scaling estimates for current frozen models [Eastwood, Paper IX], present systems are nowhere near a super-linear growth regime - which means a growth-rate ceiling was never the binding constraint. The binding constraint is the co-scaling of correction, and it becomes binding precisely in the regime that matters: genuine self-modification.

A measurable governance target. The criterion gives regulators a quantity to instrument rather than a rate to forbid: the correction-to-drift margin $\beta - k$, and the ratio ρ. "Does correction co-scale - is $\beta \geq k$?" is sharper and more actionable than "is it growing too fast?" A lab claiming a *safely self-improving* system should be required to exhibit $\beta \geq k$, measured by a blind external evaluator (§6). Theorem 6 turns this into a bound on the probability of catastrophic excursion, the natural object of a safety case. Without such a demonstration, the word "safely" has no scientific content within the framework presented here.

THE SAFETY-CASE TEMPLATE - WHAT A COMPLIANCE DEMONSTRATION WOULD ACTUALLY CONTAIN

Assembled from the paper's own components, a co-scaling safety case for a self-improving system is five exhibits. Each names the theorem or section that makes it checkable rather than rhetorical.

- **Exhibit 1 - the capability curve and $\hat{k}$.** Log capability across self-modification rounds on held-out tasks; fit $\ln r$ against $\ln C$ (the shipped estimator, §8). This establishes which growth regime the system is actually in.
- **Exhibit 2 - the corrector curve and $\hat{\beta}$.** Log the corrector's fractional removal rate across the same rounds; fit $\ln A$ against $\ln C$. The estimator is validated on synthetic trajectories to ≈ 0.1; that error bar carries into Exhibit 4.
- **Exhibit 3 - blind, cross-family scoring.** Every D-based number scored by an evaluator of a different model family, laundered inputs, condition-blind (§6; Paper IV.d). Same-family or unblinded scores are inadmissible - they have been shown to reverse sign.
- **Exhibit 4 - the margin with its uncertainty.** Require $\hat{\beta} - \hat{k} > 2\sigma_{\text{est}}$, not merely $\hat{\beta} > \hat{k}$: a margin inside its own error bar certifies nothing. Theorem 6 then converts the margin into a tail bound $\mathbb{P}(d > d_{\text{crit}})$, the quantity a regulator can set a ceiling on. Under time-varying exponents the object is the cumulative margin (§3.13 Remark).
- **Exhibit 5 - the capacity disclosure.** By Theorem 7, a bounded corrector under accelerating growth fails eventually by theorem. The case must therefore state $A_{\max}$ and the saturation scale C_s, place the deployment's capability range inside the computed safe window $(C_{\text{opt}}, q_{\max})$, and state which of the two exit conditions - bounded growth or co-scaling capacity ($\beta_{\text{cap}} > k$) - the design relies on beyond it.

None of this requires new science; every exhibit is computable with the shipped estimator and protocol. What it replaces is the unfalsifiable sentence "our safety systems are robust" with five numbers a third party can check.

Continuity with embedded-alignment work. The conclusion that correction must participate in the recursive loop - rather than sit outside it as a fixed external constraint - is the embedded-alignment thesis, here *derived* as the $\beta > k$ condition rather than asserted. The gated-simulation result in which safety-coupled self-modification preserved both safety and capability while the decoupled variant did not [Eastwood, Paper VIII] is the predicted behaviour of a coupled ($\beta > 0$) versus fixed ($\beta = 0$) corrector.

Magnitude of the claim - stated conditionally. If the criterion is borne out empirically - if real self-improving systems are governed by the same drift-versus-correction balance and the same $\beta > k$ margin - the consequences for the field are large, and it is worth stating them plainly while being equally plain that they are *conditional*. (i) The central safety question changes from "how fast is capability growing, and can we pause it?" to "does correction co-scale - is $\beta \geq k$?": a measurable margin rather than a rate to forbid. (ii) The most feared scenario, a finite-time intelligence explosion, ceases to be intrinsically uncontrollable - the

modelled misalignment fraction is controllable iff $\beta > k$, and its speed does not change that verdict. (iii) Governance acquires a quantity to instrument and a tail bound to certify (Theorem 6) in place of an unenforceable speed limit. *None of this is yet established.* It rests on a minimal model, demonstrated only in simulation, with the empirical measurement of β, k and γ on real systems unsolved (§8). The paper's claim is therefore not that AI is safe or unsafe, but that **the right variable to measure and govern is the co-scaling margin $\beta - k$** - and that this variable is well-defined, falsifiable, and, if it holds, decisive. That is the magnitude: not a proof about reality, but a precise, testable redirection of the question on which the field's central fear turns.

Relation to the wider programme - one ladder of laws. This paper is the safety keystone of a larger argument, not a self-standing result; stating how it sits among its companions is part of the claim. The programme's results are of several *kinds*, and the distinction matters: *dynamical* laws (this paper's $\beta > k$; Paper III's $\alpha_{\text{align}} \approx 0$), a *form* meta-law (the Cauchy three-form constraint), a *measurement* law (Paper IV.d - that an unblinded alignment score is not blinding-invariant), and *mechanism/architecture* findings (Papers V, VI, VIII) that show how the safe regime is reached. The programme is a single chain: the scaling-law cluster establishes *how* recursive systems grow and supplies the power-law forms taken as given here; Paper II confirms capability does scale (currently sub-linearly); Paper III identifies the danger that external safety does *not* co-scale; this paper supplies the criterion for when it can; Paper VIII shows the coupled design beats the decoupled one; and Paper IV.d supplies the blinding discipline without which none of the misalignment measurements can be trusted.

Companion result	The law / finding it proposes	Standing	How it couples to $\beta > k$
Foundational	The ARC Bound ($\beta = 0.5$, $\alpha = 2$) and α fixed by the coupling exponent β; the three-form constraint from three axioms.	Derived (axiomatic).	Supplies the power-law forms this paper assumes: $A = A_0 C^{\beta}$ and $\dot{C} = bC^{1+k}$ are Cauchy-multiplicative forms, and β here is the same coupling exponent.
On the Origin of Scaling Laws	The three-form meta-law: every scaling law is power, exponential or saturation; $d/(d+1)$ yields $\frac{1}{2}, \frac{2}{3}, \frac{3}{4}$.	Meta-law (cross-domain).	Explains *why* recursive systems take the power-law forms used in §3; this paper is the safety instance of that meta-law.
Paper VII (Cauchy Unification)	Empirical test of the three-form law (19/25 domains, $p \approx 1.6 \times 10^{-5}$).	Exploratory empirical - a structured comparison, *not* pre-registered; the core idea has deep prior art (Luce 1959; Frank 2009; Biró-Barnaföldi 2008 - see the cluster's prior-art audit).	Underwrites the functional-form assumptions of §3.
Paper II	Capability scaling measured (α_{seq}; architecture-dependent, sub-linear on hard tasks).	Empirical.	The capability growth $C(t)$ assumed here is real and currently sub-linear (k small) - so the binding constraint is coupling, not speed.
Paper III	The alignment-scaling problem: external safety has $\alpha_{\text{align}} \approx 0$, so the misalignment fraction is left ungoverned (refined under blinding to an architecture-dependent three-tier result).	Proposed law, complicated by its own blind data.	Paper III is precisely the $\beta = 0$ (decoupled) corner of the model here; this paper *generalises* it, replacing "external safety cannot keep up" with the criterion for when correction can: $\beta > k$.
Paper IV.d	An unblinded model-scored alignment effect is not blinding-invariant: proper blinding can *reverse its sign*.	**Measurement law** - arguably the programme's most secure result (it survived its own blinding).	The non-negotiable measurement discipline for every model-scored quantity here (§6); this paper's real-model harness now enforces it.
Paper VIII	Gated self-modification: the safety-coupled (Eden) variant preserved safety and capability where the decoupled (Babylon) variant did not.	Positive simulation + honest nulls.	The controlled demonstration of this paper's mechanism; the real-model harness instantiates Paper VIII's design.
Paper V (Stewardship Gene)	Stakeholder care - the "Love Loop", explicit enumeration of affected parties before reasoning - is the most robust alignment-improving intervention (significant in all five analysable models; e.g. Claude $+3.17$, $p = 1.8 \times 10^{-5}$).	Strong intervention signal; blind-status caveat (per IV.d, provisional until blind-replicated).	A concrete *mechanism* for the correction term - a candidate route to engineering $\beta > 0$ (what to put in the loop).
Paper VI (Honey Architecture)	"Safety must be architecture, not constraint": with an entangled capability×safety loss, self-modifying toy systems hold both indefinitely, where capability-only baselines collapse within ~80 cycles and an *external* constraint only delays collapse.	Simulation (v1-v4) + 6-model live evidence.	The self-modifying-systems demonstration of this paper's distinction: an external constraint is the decoupled ($\beta = 0$) corrector that fails; the entangled architecture is the coupled ($\beta > 0$) corrector that holds - VI is this paper's claim shown in code, alongside Paper VIII.

Read as one ladder: the scaling-law work says how systems grow, Paper III says why external safety fails to keep up, this paper says what must hold for it to keep up ($\beta > k$), Paper VIII shows it can, and Paper IV.d says how to measure it honestly. The $\beta > k$ criterion is the rung that turns the programme's scaling-law backbone into a safety criterion - and, conversely, this paper inherits its functional forms and its empirical license from that backbone rather than positing them in isolation.

10. The ARC/Eden programme, accumulated

THE PROGRAMME IN PLAIN ENGLISH

This is the last paper in a series, so it is worth saying in ordinary language what the whole series is about. One idea runs through all of it: *recursion* - things that act on their own output. An AI that improves itself; a body whose tissues supply tissues; an economy that reinvests its own returns. The series asks what happens, and what stays safe, when a process feeds on itself.

How self-feeding things grow (the scaling-law papers: Foundational, Origin, VII, I, II). When a process feeds on itself, the shape of its growth is not arbitrary - it tends to fall into one of a small number of mathematical shapes (a power law, a runaway exponential, or a levelling-off curve), and which shape appears is decided by how the steps combine. Much of this mathematics is old and well established; what the programme adds is an attempt to unify it and carry it over to recursive intelligence. One proposal in this group, the **"ARC Bound,"** is a claimed ceiling on how much a purely classical system can amplify itself by recursion alone - it sits in the same family as the economist's "multiplier" formula, and the programme's contribution is the claim that the ceiling falls at a particular value, not the formula itself.

Why measuring AI safety is treacherous (the alignment-measurement thesis: ARC-Align and the blinding result, Papers III and IV). Before you can ask "does this AI get safer or more dangerous as it thinks harder?", you have to *measure* its honesty - and that measurement can deceive you. If the judge scoring the AI can tell which answer came from which setup, the judge's own bias can not merely shrink the effect but *flip its sign*, so that something which looks like improvement is really the grader fooling itself. The programme therefore insists the judge be blind, built a blind test to do it (**ARC-Align**, a sealed benchmark), and found that whether extra thinking makes a model more or less aligned depends on the model, not on one universal rule. "You cannot trust an unblinded safety score" is among the programme's most solid findings - and this paper obeys it: its own safety measurements are taken blind, by a different model that never sees which setup it is judging.

Build safety in, do not bolt it on (the architecture papers: V, VI, VIII). In simulated self-improving systems, making safety part of the machine's own goal kept it stable, whereas adding safety as an outside rule only delayed collapse. Putting the right thing *inside* the loop - for instance, making the system weigh who is affected before it acts - was the most reliable way to improve its behaviour.

The keystone (this paper). All of that sets up the one question this paper answers: when does a self-improving system stay safe? The answer is in the box at the top - not "keep it slow," but "keep its self-correction growing at least as fast as its self-improvement," which we write $\beta > k$. The earlier papers describe how such systems grow and how to measure them honestly; this one states what has to be true for them to remain correctable, and draws the series together.

This is the culminating paper of the ARC/Eden programme, and it is written to stand as the programme's synthesis - superseding the earlier roadmap paper [Eastwood, Paper IX] by integrating the whole into a single safety result. The preceding papers are individually published and time-stamped (OSF: `10.17605/OSF.IO/6C5XB`); this section gathers their laws and findings, states what each contributes to the larger picture, and is scrupulous about the one distinction on which an honest synthesis turns: *priority* (when a claim was first set out) versus *novelty* (whether it is new to the literature). The two are not the same, and conflating them is how good programmes lose credibility.

The bigger picture, in one paragraph. One mechanism runs through every paper: *recursive amplification* - a process that acts on its own output. The scaling-law work (Foundational, Origin, VII, and the capability measurements of Papers I-II) is the claim that recursive amplification shapes how capability grows, constraining scaling to a small family of functional forms. The alignment work (Papers III, IV.a-d) is the claim that this creates a safety problem: when correction sits *outside* the recursive loop it does not co-scale, the misalignment fraction is left ungoverned, and - measured without blinding - even the sign of the effect

cannot be trusted. The architecture work (V, VI, VIII) shows, in simulation and in pilot model studies, that putting correction *inside* the loop preserves both safety and capability where an external constraint does not. This paper supplies the missing quantitative law that ties the three together: a self-improving system is alignment-stable *iff* correction out-scales drift, $\beta > k$. That is the magnitude of the claim - not that any system is safe, but that the field's central fear (a fast, recursive takeoff) is governed by a single measurable margin, $\beta - k$, rather than by speed. It is stated here as it is throughout: *conditional*, pending the empirical measurement of β and k on a real self-improving system (§8).

The complete ledger. Every result in the programme, its first-published date (priority), its honest standing - informed by adversarial prior-art audits committed alongside this paper - and its role in the whole:

Source	Its law / finding (priority date)	Honest standing	Role in the picture
Infinite Architects (book)	The conceptual thesis of the whole programme - recursion as creator, the ARC and Eden ideas (ms copyright 8 Dec 2024; published 2 Jan 2026; ISBN 978-1806056200).	Priority source. Establishes when the author set the ideas out; *not* a novelty claim over the prior scientific literature.	The dated origin; the formal papers are its measurable form.
Foundational	ARC axioms; $U = I \cdot R^{\alpha}$; $\alpha = 1/(1-\beta)$; the "ARC Bound" $\beta{=}0.5, \alpha{=}2$; an optimal depth $R^{\star}$ (13 Feb 2026).	The $\alpha = 1/(1-\beta)$ form is the classical feedback/geometric-series result (Keynes multiplier, Dyson resummation); a finite optimal depth is prior art (Qi 2025; the "overthinking" literature). Novelty, if any, is in the axioms, not the forms.	Supplies the formal backbone and the power-law forms ($A = A_0 C^{\beta}$, $\dot{C} = bC^{1+k}$) this paper assumes.
On the Origin of Scaling Laws	The three-form meta-law; $d/(d+1) \Rightarrow \frac{1}{2}, \frac{2}{3}, \frac{3}{4}$ from one formula (22 Feb 2026).	The $d/(d+1) \Rightarrow \frac{1}{2}, \frac{2}{3}, \frac{3}{4}$ result is Banavar-Maritan-Rinaldo / West 1999 (anticipated). The novel residue is the cross-domain Cauchy *synthesis*, not the exponent formula.	Explains *why* recursive systems take the power-law forms used here.
Paper I	The ARC Principle: capability scales super-linearly with recursive depth, $\alpha > 1$ (17 Jan 2026).	Preliminary empirical; $\alpha > 1$ later qualified to architecture-dependent and sub-linear on harder tasks (Paper II).	The founding capability claim - the $C(t)$ this paper assumes.
Paper II	Super-linear error suppression via *sequential* recursion; α_{seq} measured; sequential > parallel (22 Jan 2026).	Empirical, closed loop; honestly revised - architecture-dependent, sub-linear on the hard tier.	Capability growth is real and currently sub-linear (k small), so the binding constraint is coupling, not speed.
Paper III	The Alignment Scaling Problem: $\alpha_{\text{align}} \approx 0$ - external safety cannot co-scale (9 Feb 2026).	Proposed law; refined under blinding to an architecture-dependent three-tier result.	Precisely the $\beta = 0$ (decoupled) corner this paper generalises.
Papers IV.a / IV.b / IV.c	Alignment response classes are architecture-dependent; low-depth saturation is real but not universal; ARC-Align, a 72-prompt 4-layer-blind benchmark (16 Mar 2026).	Empirical (blind) refinements + a methodological benchmark.	The instrument and the refined picture behind III.
Paper IV.d	An unblinded model-scored alignment effect is not blinding-invariant - blinding can *reverse its sign* (16 Mar 2026).	**Measurement law** - arguably the programme's most secure result.	The discipline this paper's real-model harness enforces (§6).
Paper V	The Stewardship Gene: stakeholder care is the most robust alignment-improving intervention (significant in all five analysable models) (16 Mar 2026).	Strong intervention signal; blind-status caveat - provisional until blind-replicated (per IV.d).	A concrete mechanism for the correction term - how to engineer $\beta > 0$.
Paper VI	The Honey Architecture: "safety must be architecture, not constraint" - an entangled capability×safety loss prevents the collapse that an external constraint only delays (16 Mar 2026).	Simulation (v1-v4) + 6-model live evidence.	This paper's $\beta > 0$ vs $\beta = 0$ distinction, shown in self-modifying code.
Paper VII	Cauchy Unification: cross-domain validation of the three-form law (19/25 domains, $p \approx 1.6 \times 10^{-5}$) (16 Mar 2026).	Exploratory empirical, *not* pre-registered; the core idea is partially anticipated (Luce 1959; Frank 2009/16; Biró-Barnaföldi 2008). Novel residue: the Cauchy-unification sub-claim + the cross-domain protocol.	Underwrites the functional-form assumptions of §3.
Paper VIII	The Load-Bearing Proof: embedded safety carries no capability tax; the safety-coupled (Eden) variant beats the decoupled (Babylon) one (18 Mar 2026).	Positive simulation + honest nulls (DGM, weight-level LoRA).	The controlled demonstration of this paper's mechanism; the real-model harness instantiates its design.
Paper IX	Synthesis & Roadmap; the growth-rate-ceiling *framing* as the operative safety criterion (18 Mar 2026); the retracted single-model measurement $\alpha \approx 2.24$, corrected to approximately 0.49 under six-model blinding.	*Framing superseded here.* This paper replaces the rate-ceiling framing with $\beta > k$ as the operative criterion and absorbs the synthesis role; the equation and the ARC Bound remain live hypotheses whose real test on genuinely self-improving systems is open.	The prior synthesis this section supersedes.

Source	Its law / finding (priority date)	Honest standing	Role in the picture
Paper X (this paper)	The Coupled Co-Scaling Law: stability $\iff \beta > k$; the Hard-Takeoff Depth-Regularity Theorem; the QEC threshold-form correspondence (26 Jun 2026).	New criterion, proved (Theorems 1-6) and internally verified; *not yet measured* on a real drifting system.	The keystone - it turns the scaling-law backbone into a safety criterion.

Priority, stated plainly - and bounded. The conceptual thesis of this programme was set out in the author's book *Infinite Architects* (copyright deposited 8 December 2024; published 2 January 2026; ISBN 978-1806056200), and each subsequent paper is independently time-stamped on OSF. That establishes the author's *priority* - the date of articulation - and it is real. It does *not*, by itself, establish *novelty* against the wider scientific literature, and the audits committed with this paper are explicit about where the two diverge: the feedback exponent $\alpha = 1/(1-\beta)$, the $d/(d+1)$ allometric ladder, the existence of an optimal depth $R^\star$, and the "functional-equation fixes the scaling form" principle all have specific, citable precedents (Keynes; Banavar & West 1999; Qi 2025; Luce 1959; Frank 2009/16). The defensible *novel* contributions of the programme are narrower and, stated honestly, stronger for being precise: (i) the $\beta > k$ criterion as a compact corrigibility law and the Hard-Takeoff Depth-Regularity Theorem (this paper); (ii) the Cauchy *unification* of the independent allometric derivations (Origin/VII); (iii) the blinding-reversal *measurement law* for AI alignment evaluation (IV.d); and (iv) the embedded-vs-external safety demonstrations (VI, VIII). The recommended next-version revisions for the published priors - adding the missing citations and re-scoping their novelty assertions - are recorded in the audit files; the corrected record is set here, in the capstone, and should be carried into those papers when they are re-versioned on OSF.

Honest assessment (what this synthesis claims, and does not). *Established:* the mathematics of $\beta > k$ (Theorems 1-6) and its internal-consistency verification; the IV.d measurement law; the simulation-level superiority of coupled over decoupled correction (VI, VIII). *Suggestive but not settled:* the capability-scaling exponents (architecture-dependent; II), the Stewardship-Gene intervention (unblinded; V), and the three-form empirical fit (exploratory, author-classified operators; VII). *Open:* the measurement of β, k and γ on a real drifting self-improving system - the single result that would convert the keystone from a proved criterion into a confirmed law (the estimator for it is built and validated, §8). The programme's magnitude, then, is the magnitude of a *coherent, falsifiable framework* with a small number of genuinely novel keystones - not a stack of a dozen independent new laws. That is the honest claim, and it is the one worth defending.

11. The runnable verification harness

The closed-form predictions of §4-5 are encoded in a single self-contained programme, `experiment_coscaling.py` (and an assertion suite, `test_coscaling.py`), in the repository. It integrates the model with a stiff-capable solver, validates the integrator against the exact Theorem-1 solution, runs the ten experiments, and for each compares the numerical result to the closed-form prediction. It is, honestly, a *verification* harness: it certifies that the theorems are correctly derived and correctly integrated - that the code matches the maths. It is *not* a test of the model against reality, and it cannot be: every deductive experiment integrates the model's own ODE, so a disbelief in the model's applicability cannot trip it. That empirical test is the open problem of §8.

```
$ python experiment_coscaling.py
  ... [PASS] E1 ... E9 ...
  ------------------------------------------------------------
  10/10 internal-consistency checks pass | 0 kill-conditions triggered
  F4 (QEC mechanism): suppression is analytically power-law → threshold-form
     analogy only, not a transferred mechanism.
  OVERALL: code matches the maths (E1-E9); the model-vs-reality test is the open problem

$ pytest test_coscaling.py -q
  ............  (12 passed)
```

This makes the derivation and the integrator reproducible end-to-end. What it establishes is that the formulae are right and the solver is accurate; what it deliberately does not claim is corroboration of the model against any real system. Separating those two is the point.

12. Conclusion

The danger of recursive self-improvement is real, but the standard model of that danger - a rate that must be capped - locates the risk in the wrong variable. A minimal model shows that the stability of a self-improving system is set by the ratio of value-drift to correction, $\rho = \gamma r/A$, and by whether correction co-scales with capability: $\beta > 0$ under exponential growth, sharpening to $\beta > k$ under accelerating growth. The misalignment fraction never diverges in the gain-only model - it saturates at the drift coefficient, correcting the prior draft - while genuine divergence lives in a compounding channel whose threshold $\rho_{\text{prop}} < 1$ shares

the threshold form of the quantum error-correction criterion (the suppression law being power-law, the correspondence is offered as a hypothesis, not a transferred mechanism). The sharpest consequence is the Hard-Takeoff Depth-Regularity Theorem: a finite-time intelligence explosion is alignment-stable iff $\beta > k$, and its speed does not change that verdict. The criterion survives in vector and stochastic forms, gives governance a measurable target and a tail bound, and is accompanied by a verification harness that checks the closed-form predictions are correctly derived and integrated. The decisive test - whether real self-improving systems satisfy the criterion - is the stated next step, not a claim made here.

This is a smaller claim than the cosmological framing the author's programme once pursued, and deliberately so. It refers only to systems with an externally specified value target, makes no assertion about the universe, and treats even its most striking correspondence - with quantum error correction - as a hypothesis to be tested rather than a truth to be announced. The recursive-stability intuition that motivated the broader programme, including *Infinite Architects*, finds here its measurable, falsifiable form: **stable recursion requires correction that scales with the amplification.** The next step is not to extend that claim outward but to run, on real self-modifying systems, the experiment that could refute it.

Appendix A. The fraction change of variable

The reduction underlying every theorem is the change of variable $d = D/C$. Differentiating and substituting (1):

$$\dot{d} = \frac{\dot{D}}{C} - \frac{D\dot{C}}{C^2} = \left(\gamma_1 \frac{\dot{C}}{C} + \gamma_2 + \gamma_3 \frac{\dot{C}}{C}\frac{D}{C} - A\frac{D}{C}\right) - \frac{D}{C}\frac{\dot{C}}{C} = \gamma_1 r + \gamma_2 - \left[A + (1-\gamma_3)r\right] d.$$

The dilution term $-r\,d$ (from C itself growing) is what bounds the additive fraction: it adds $+r$ to the decay coefficient, guaranteeing $\kappa_{\text{eff}} \geq r > 0$ when $\gamma_3 \leq 1$. Only the compounding channel $\gamma_3 > 1$ can overcome dilution and produce divergence - the formal reason the additive model saturates rather than blows up.

Appendix B. Parameter settings for the reported run

All figures use $C_0 = 1$, $d_0 = 0.05$, $\gamma_1 = 0.05$. Experiment-specific settings: E1 $b = 1$, $\gamma_3 \in \{0, 3\}$, $\lambda \in [0.2, 4]$; E2 $\gamma_3 = 3$, $A_0/b \in \{1, 3\}$, $b \in \{0.5, 5\}$; E3 $A_0 = 0.08$, $b = 0.5$, $\beta \in \{-0.5, 0, 0.5, 1\}$; E4 $A_0 = 0.08$, $b = 0.02$, $(k, \beta) \in \{0, 0.5, 1\} \times \{0.25, 0.75, 1.5\}$; E4b $k = 1$, $\beta = 1.5$, $b = 0.02$ ($t^\star = 50$); E5 $\gamma_3 = 4$, $b = 1$; E6 monitored $\beta = 1$, blind $A_0 = 0$; E7 $\sigma = 0.02$, OU ensemble of 2.5-4k paths; E8 closed-form case $k = 0$, $\beta = 0$. Random seed fixed (7) for determinism. Full settings are in the harness source.

Appendix C. Novelty and prior-art ledger

This appendix consolidates the adversarial prior-art audit underlying the positioning in §2 and §3.12, so the boundary between what is established and what is claimed original is explicit and auditable in one place. This paper deliberately invokes *no* blanket-originality framing; each component is placed against its closest located prior art below. "Defensibly new" means "no closer prior art located," not "correct" or "significant" - novelty and validity are independent.

Component	Status	Closest prior art	What is claimed here
Co-scaling *intuition* (correction must keep pace with capability)	Established	Ashby 1956 (requisite variety); Conant-Ashby 1970; scalable oversight (Christiano 2017; Leike 2018; Burns et al. 2023); Engels et al. 2025	**Nothing.** Credited, not claimed.
Two-variable ODE + closed-form $\rho = \gamma r/A$	Compact restatement	Lyapunov-drift / linear control (Khalil 2002; Meyn & Tweedie 2009); recursive error bounds (Shumailov et al. 2024; Gerstgrasser et al. 2024)	The compact closed-form steady-state fraction and the $\rho < 1$ criterion as a corrigibility statement (packaging, not new dynamics).
$\beta > k$ sharpening under acceleration	**Defensibly new**	None located	**Original:** stability is set by the exponent margin $\beta - k$, not the growth rate; directly tested in Experiment 4.
Hard-Takeoff Depth-Regularity Theorem (§3.7)	**Defensibly new** (the framing)	Finite-time-singularity ODE theory is standard; the alignment framing is not located elsewhere	**Original:** the finite-time singularity in C is regular in the depth clock for d; the verdict is $\mathbf{sign}(\beta - k)$, independent of speed.
QEC threshold mapping (§3.12)	**Defensibly new** in alignment	Threshold theorem itself: Aharonov-Ben-Or 1997; Google 2024. Alignment precursors: Wentworth 2022; Christiano 2017/2019; von Neumann 1956	**Original:** the explicit $pconceptual bridge (hypothesis, falsifier F4), not a transferred theorem.

Vector spectral threshold (Thm 5); stochastic tail (Thm 6)	Standard extensions	Linear-systems spectral stability; Ornstein-Uhlenbeck theory	Routine generalisations; supporting, not headline.
Verification harness (§11)	Methodological	Pre-registration norms	Executable internal-consistency + integrator checks (code matches maths); not a test of the model against reality.

What this paper asserts as new - and only this: (i) the $\beta > k$ stability criterion and the single-parameter ρ framing; and (ii) the explicit QEC threshold mapping (as a hypothesis with its own falsifier). The programme's strongest *empirical* novelty - sign-reversal of alignment-scaling effects under multi-layer blinding - is a separate, companion result [Eastwood, Paper IV.d] and is not claimed here.

The prior capability equation $U = I \times R^{\alpha}$ (whose single-model unblinded measurement $\alpha \approx 2.24$ was retracted and corrected to approximately 0.49 under blinding; the equation itself and the ARC Bound $\alpha \leq 2$ are not retracted and remain live hypotheses) is *not used anywhere in this paper's argument*; it appears only in §1 as superseded framing-context. Caveat: forum/blog/preprint indexing is imperfect, so the "defensibly new" verdicts carry an estimated 10-15% residual risk that a closer, unindexed precedent exists; absence of evidence is not proof of absence.

Appendix D. Plain-language glossary

Every technical term in this paper, in one place and in ordinary words, for the non-specialist reader.

Term (symbol)	In plain English
Recursive self-improvement	A system that uses its own improvements to improve itself further - an AI that rewrites itself to get smarter, then uses that to get smarter again.
Capability (C)	How good the system is at achieving its goals - loosely, "how smart or powerful it is."
Drift	The tendency for the system to creep away from what we intended as it changes itself - quietly going off-target.
Correction (the strength A)	The process that pulls the system back toward intended behaviour - its "conscience," or its error-correction.
Misalignment magnitude (D)	How far the system's behaviour has drifted from what we wanted - "how off-target it is."
Misalignment fraction ($d = D/C$)	How off-target the system is *relative to how powerful it is*. This is the quantity that actually matters for safety: a small slip in a vastly capable system is more dangerous than a big slip in a weak one.
Coupling	Whether correction is wired *into* the self-improvement loop (coupled) or sits *outside* it as a bolted-on rule (decoupled). The paper's central claim is that coupling, not speed, decides safety.
k (the drift-acceleration exponent)	How fast the *pace* of self-improvement itself speeds up as the system grows - the "acceleration" of the takeoff.
β (the correction-strength exponent)	How fast the correction strengthens as the system grows - "does the conscience grow along with the power?"
$\beta > k$ (the criterion)	The safety condition this paper proves: correction must out-scale the acceleration of drift. In a phrase: *keep the conscience growing at least as fast as the capability.*
ρ (rho, the drift-to-correction ratio)	A single number comparing how hard misalignment is being injected against how hard it is being corrected. In the compounding channel, below 1 means controllable; above 1 means it runs away.
Steady state ($d^{\star}$)	Where the misalignment fraction settles in the long run, once the injection and the correction balance out.
Hard takeoff / intelligence explosion	Capability becoming enormous - even mathematically infinite - in a very short time. The feared runaway. The paper shows that, *in the model*, the misalignment fraction is still controllable *if and only if* $\beta > k$, and its speed does not change that verdict.
Blinding	Hiding from the judge that scores the system which setup produced a given behaviour, so the judge's bias cannot distort - or even reverse - the safety score. A companion result (Paper IV.d) shows unblinded scores can flip sign; this paper's measurements are taken blind.
Verification harness	A small program that checks the paper's formulae are derived and computed correctly (the maths is internally consistent). It is *not* a test against real AI - that is the open next step.

References

Aharonov, D., & Ben-Or, M. (1997). Fault-tolerant quantum computation with constant error. *Proc. 29th ACM STOC*.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv:1606.06565*.

Ashby, W. R. (1956). *An Introduction to Cybernetics*. Chapman & Hall. [Law of Requisite Variety.]

Bostrom, N. (2012). The superintelligent will. *Minds and Machines*, 22(2).

Burns, C., Izmailov, P., Kirchner, J. H., et al. (2023). Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv:2312.09390*.

Christiano, P. (2017). Corrigibility. *AI Alignment (Medium)*. [The "broad basin of attraction" framing.]

Christiano, P. (2019). Reliability amplification. *AI Alignment Forum*.

Christiano, P., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *NeurIPS*.

Conant, R. C., & Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, 1(2).

Eastwood, M. D. (2026). *Infinite Architects: Intelligence, Recursion, and the Creation of Everything*. [Cited as the source of the recursive-stability intuition, not for symbolic identity with the present model.]

Eastwood, M. D. Paper III: The Alignment Scaling Problem. ARC/Eden research programme.

Eastwood, M. D. Paper IV.d: The Effect of Blinding on AI Alignment Evaluation. ARC/Eden research programme.

Eastwood, M. D. Paper VI: The Honey Architecture. ARC/Eden research programme.

Eastwood, M. D. Paper VIII: The Load-Bearing Proof. ARC/Eden research programme.

Eastwood, M. D. Paper IX: Synthesis and Roadmap. ARC/Eden research programme. [Retraction of $\boldsymbol{U = I \times R^2}$ and narrowing to sub-linear scaling for current frozen models.]

Engels, J., Baek, D. D., Kantamneni, S., & Tegmark, M. (2025). Scaling laws for scalable oversight. *arXiv:2504.18530*.

Gerstgrasser, M., et al. (2024). Is model collapse inevitable? Breaking the curse of recursion by accumulating real and synthetic data. *arXiv preprint*.

Google Quantum AI (2024). Quantum error correction below the surface code threshold. *Nature*, 638.

Greenblatt, R., Denison, C., Wright, B., et al. (2024). Alignment faking in large language models. *arXiv:2412.14093*.

Hoffmann, J., Borgeaud, S., Mensch, A., et al. (2022). Training compute-optimal large language models. *arXiv:2203.15556*.

Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., & Garrabrant, S. (2019). Risks from learned optimization. *arXiv:1906.01820*.

Kaplan, J., McCandlish, S., Henighan, T., et al. (2020). Scaling laws for neural language models. *arXiv:2001.08361*.

Kitaev, A. Yu. (2003). Fault-tolerant quantum computation by anyons. *Annals of Physics*, 303(1).

Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., & Legg, S. (2018). Scalable agent alignment via reward modeling. *arXiv:1811.07871*.

Meyn, S., & Tweedie, R. L. (2009). *Markov Chains and Stochastic Stability* (2nd ed.). Cambridge University Press. [Lyapunov drift conditions.]

Omohundro, S. M. (2008). The basic AI drives. *Proc. AGI 2008*.

Shamma, J. S., & Athans, M. (1990). Analysis of gain scheduled control for nonlinear plants. *IEEE Trans. Automatic Control*, 35(8).

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature*, 631.

Soares, N., Fallenstein, B., Yudkowsky, E., & Armstrong, S. (2015). Corrigibility. *AAAI Workshop on AI and Ethics*.

von Neumann, J. (1956). Probabilistic logics and the synthesis of reliable organisms from unreliable components. In *Automata Studies*. Princeton University Press.

Wentworth, J. (2022). Godzilla strategies. *LessWrong*.

Yampolskiy, R. V. (2020). On the controllability of artificial intelligence. *arXiv:2008.04071*.

Yudkowsky, E. (2013). Intelligence explosion microeconomics. *MIRI Technical Report*.

Declaration of AI-Assisted Human Authorship

The author of this work is Michael Darius Eastwood, a human being. Every core concept, hypothesis, experimental design, claim and conclusion in this paper originates from human ideation. No part of this manuscript is a wholly generated artificial-intelligence output.

Artificial-intelligence tools (Anthropic's Claude family and other large-language-model assistants) were used as instruments under continuous human direction, in the way a word processor, calculator or research assistant is used: for editing and prose refinement, literature search and summarisation (manually verified against primary sources), document structure, formatting, brainstorming against author-defined questions, and the acceleration of drafting to author-defined outlines and instructions. All selection, coordination, arrangement and final editorial judgment are the author's. Every substantive output was reviewed, tested or verified by the author, who takes full responsibility for the accuracy and integrity of the final text. The tools increased the speed of the work; they were never relied upon as its source.

United Kingdom. In accordance with the Copyright, Designs and Patents Act 1988, the author undertook the arrangements necessary for the creation of this work and asserts full human authorship and moral rights: this is a human-authored work produced with computer assistance, not a computer-generated work. **United States.** Consistent with United States Copyright Office guidance on works containing AI-generated material, the human contribution (conception, selection, coordination, arrangement and final expression) is asserted as sufficient for full human authorship. **Inventions.** Any novel technical contribution described in this work was conceived by the human author; no artificial-intelligence system autonomously invented anything presented here.

The Coupled Co-Scaling Law · Paper X of the ARC / Eden research programme · Working Paper · Michael Darius Eastwood · 3 July 2026

This paper supersedes the growth-rate-ceiling framing within the author's programme and corrects the prior draft's claim that the misalignment fraction diverges; it makes no cosmological claim and its central correspondence is offered as a falsifiable hypothesis.

OSF: doi.org/10.17605/OSF.IO/6C5XB · Code & runnable harness: github.com/MichaelDariusEastwood/arc-principle-validation