The Self-Acceleration Exponent

Michael Darius Eastwood
Independent AI alignment researcher, London · Author, *Infinite Architects* (2026)
ORCID 0009-0004-3222-7442 · michael@michaeldariuseastwood.com · michaeldariuseastwood.com
The ARC Theory · OSF osf.io/ht8wu · every claim checkable

> *Within the ARC Theory: the bridge between Law I's growth and Law II's stability; the self-acceleration exponent, resolving the notation collision.*

A Measurable Threshold Joining Recursive Capability Growth to Corrective Stability

Michael Darius Eastwood
Independent researcher · London, United Kingdom

Working Paper v1.4 · First published 16 August 2026, revised 27 August 2026 · DOI: 10.17605/OSF.IO/HT8WU

STATUS · A DERIVATION, NOT A MEASUREMENT

This paper derives a measurable threshold, δ = 1/α, on a quantity any self-improvement loop can record, from one stated scope condition. It is not a measurement claim: three of the four quantities in the central result have never been measured in any system. Section 10 states what would refute it, section 12 records what earlier drafts withdrew, and section 13 states what is not claimed as new.

Two of the estate master's five deposit conditions remain open and are stated here rather than waived: the deciding measurement is drafted, dated and prepared as a draft registration awaiting human submission, and no mathematician without a stake in the programme has yet read the algebra adversarially. The author took the disclosure decision to publish on 16 August 2026 with both open conditions known. The programme-wide correction record is at /research/corrections/.

How to read this paper. Claims come in three sizes: results, laws, frames. This is a laws-and-frames paper: Results 1 and 3 are algebra, Results 2, 5 and 6 are algebra given a stated scope condition, and Result 7 additionally rests on a conjecture that is flagged rather than folded in. Nothing here is measured, and the paper says so wherever the temptation to imply otherwise would arise.

Abstract

Two frameworks in this programme describe recursive self-improvement and have never been related. One derives how capability compounds with recursive depth and yields the scaling exponent $\alpha = 1/(1 - \beta\_L)$. The other derives when a self-improving system remains correctable and yields the criterion $\beta\_X > k$. **They use the same two letters for different quantities**, which has prevented the comparison rather than merely obscuring it.

This paper resolves the notation, derives the relation between the frameworks, and finds that the quantity deciding the outcome is neither capability nor speed but **how fast a system compresses its own iteration time**. Writing cycle duration as $\Delta t \propto C^{-\delta}$ defines the **self-acceleration exponent** δ, and the growth exponent of the corrective framework follows exactly:

$$k = \delta - 1/\alpha$$

The threshold is $\delta = 1/\alpha$. Below it a system's capability grows as a power law and the corrective question does not arise. Above it growth is super-exponential and correction must out-scale it. On one assumption about the form of the correction process, the survivable window is

$$1/\alpha < \delta < 1/\alpha + \beta\_X$$

two boundaries whose gap is exactly the correction exponent. Only the upper one is a stability threshold: the lower marks where growth leaves the power law, and a system between them is in a hard takeoff and still correctable, which contradicts the standard picture in both directions.

δ requires no new instrument, benchmark or model access: it is the slope of cycle duration against capability on logarithmic axes, on a fixed substrate, from series any self-improvement loop can record out of what it already computes, though, as §9 re-

cords, the two series are not currently paired anywhere. β_X and β_L have never been measured; δ has not either, and it is the one that could be measured within a month of a loop existing.

---

1. The question

Whether a self-improving system remains correctable has been argued for two decades in words. The positions divide on takeoff speed, and both sides treat the outcome as decided by a *level*: how capable the system becomes, or how fast it gets there.

This paper argues the outcome is decided by neither, and identifies the quantity that decides it. The argument requires joining two frameworks that were built independently and cannot currently be compared, because they collide in notation.

1.1 The pipes premise, at fixed throughput

Everything that grows is fed through a channel, and for everything before software, fixing the channel fixes the growth. Starve a tumour's vasculature and it stops. Fix a chain reaction's fuel and geometry and it stops. Exhaust an epidemic's susceptible hosts and it burns out. Astrophysics has a named, quantitative version of the same ceiling: the Eddington limit, above which radiation pressure halts accretion. The claim here is a generalisation of limits physics already accepts, not a new species of assertion.

Two hard cases are handled deliberately. Cosmic inflation is excluded by scope: it is expansion of space, not growth of a structure on a substrate, and it ended by field dynamics rather than by running out of anything. Evolution is the sharpest case and it fits: biological complexity increased for billions of years at roughly fixed solar throughput, which is growth on information at fixed pipes, and it is also glacial, and it has no internal corrector at all, since selection is external. On this framework, evolution should carry a measurable exponent far below the ceiling; that is a prediction about biology falling out of a framework built for software, and it is stated as one.

Software is the first thing that keeps growing when you hold the pipes fixed. The operational definition does the work here and is load-bearing, not a technicality: an artefact loop on a fixed substrate, where each pass works on the accumulated output of the previous passes, measured on a substrate-fixed clock. A model merely thinking longer does not qualify. Growth that only appears when the substrate is enlarged is pipe-fed growth and stays under the old law; growth at fixed substrate is the new regime, and it is the regime this programme's limit addresses.

Terminology note. The bare phrase "runaway growth" is not used in this paper without qualification: in the physics corpus it denotes planetesimal accretion in planet formation. Where a runaway is meant here, the phrase is "runaway self-improvement" or "uncorrectable growth", per the terminology audit summarised in §2.1.

2. Two notations, and two collisions

This section is not housekeeping. The join is unreadable without it, and one of the collisions produces an absurd result if crossed.

symbol	in the scaling framework	in the corrective framework
β	**leverage fraction.** dg/dr = a g^{β_L}, where β_L is how much accumulated context each step can effectively leverage. Bounded, β_L ∈ [0,1)	**correction exponent.** A = A_0C^{β_X}, the exponent with which correction strengthens as capability grows. Unbounded
r	**recursive depth**, a step counter	**specific growth rate** Ċ/C, where the framework writes r ∝ C^k

Neither framework is wrong. They were never reconciled.

The consequence of crossing the β collision. Substituting β_L = 0.5 from the scaling framework into the corrective criterion sign(β_X − k) is not an approximation. By §4 it corresponds to β′_L = 0, a corrector with no leverage at all: one that cannot use its own accumulated corrections. **The substitution silently assumes the least capable corrector constructible and presents the result as a property of the system.**

The consequence of crossing the r collision. Matching dC/dr in the scaling framework against a growth law expressed in the corrective framework's r treats a step counter as a rate, and produces a relation between exponents that does not hold. An early draft of this paper did exactly that and the result is withdrawn in §12.

Notation used hereafter. β_L and β′_L for the leverage fractions of capability and of correction, β_X for the correction exponent, k for the capability growth exponent, α for the scaling exponent, r for recursive depth only, τ for the depth clock, and δ for the self-acceleration exponent introduced in §5.

2.1 Four further terminology collisions from the wider programme (operator ratification pending)

The two collisions above are internal to this paper. Four further collisions bind at the level of the wider programme, and Paper XIII inherits them:

1. **"Correction exponent" collides with condensed-matter physics**, where it denotes the subleading correction to scaling near a critical point (conventionally omega). All 58

arXiv abstract hits for the phrase are statistical mechanics (measured 12 August 2026). The registrations already use **"correction-leverage exponent"** (study AE, study AG), which collides with nothing. PROPOSED: standardise every surface on "correction-leverage exponent", with "correction exponent" retained once per document as a parenthetical alias during transition. NOT applied to published papers without the operator's ratification. This paper therefore continues to use β_X and "correction exponent" throughout, with the collision named here.

2. **Phi collides with integrated information theory**, whose scalar Phi is a different object from the programme's Phi(H, chi) functional. Any surface using the December formalism states the collision before a reader finds it. (Paper XIII does not deploy the Phi(H, chi) functional and is unaffected in body; noted here for cross-surface parity.)
3. **Alpha collides with itself across the programme's own eras**: the fine-structure constant in the December 2024 formalism, the ARC Bound exponent in the 2026 work. The glossary carries both meanings with dates; no new surface uses bare alpha without its era. In this paper α is the ARC-Bound / scaling exponent (α = 1/(1 − β_L)) in the 2026 sense throughout, and never the fine-structure constant.
4. **"Runaway growth" belongs to planet formation** in the physics corpus. Where a runaway is meant here, the phrase is "runaway self-improvement" or "uncorrectable growth"; see §1.1 note.

3. The clock relation

Result 1 (exact). The two frameworks run on different clocks and the relation between them is a logarithm.

The scaling framework integrates dg/dr = a g^{β_L} to g ∝ r^{1/(1−β_L)}, hence

$$C \propto r^{\alpha},\ \alpha = 1/(1 - \beta_L)$$

The corrective framework, when it leaves wall-clock time, passes to the depth clock τ = ln(C/C0). Substituting:

$$\tau = \alpha \ln r + \text{const}$$

The corrective framework's depth clock is the logarithm of the scaling framework's recursion depth, scaled by α. The two frameworks were already in one variable, up to a log, and neither noticed.

Consequence. The corrective framework's control quantity $q = A/r_{rate} \propto C^{\beta_X - k}$ becomes, in recursion depth, $q \propto r^{\alpha(\beta_X - k)}$. **So α amplifies the rate at which the verdict arrives per recursive step, and cannot change the verdict.** That is not merely consistent with the corrective framework's speed-independence theorem; it is required by it, since α is a speed in the only sense that matters, and a change of variables cannot flip a sign.

4. The correction exponent is a ratio of leverage complements

4.1 Not an assumption. A scope condition, and the reason matters

An earlier draft stated this as "Assumption A1: correction is itself a recursive process", and that framing does not survive contact with the source. Read at source, the corrective framework introduces correction as a **constitutive law inside its master system**:

$$\dot{D} = \gamma_1\dot{C} + \gamma_2 C + \gamma_3(\dot{C}/C)D - AD,\ A = A_0 C^{\beta},\ \dot{C} = bC^{1+k}$$

A is a specified function of C. It is not a state variable. There is no $\dot{A}$, no dA/dr, and no corrector accumulation anywhere in the framework; searched exhaustively on 8 August 2026 and found absent in every form.

So β_X is a free parameter of that model, and there is nothing in it to derive. Stating A1 as an assumption *inside* the framework would introduce β'_L, a quantity with no referent, and the join would not be wrong so much as vacuous. **That is what a hostile reader finds in ten minutes, and it is better found here.**

The correct move is to state it as a scope condition on the corrector, not an assumption about the model. A scope condition cannot be falsified by finding a system it does not cover.

4.2 The corrector taxonomy · a decision tree over what A may depend on

An earlier draft used two binary questions and it does not close the taxonomy. Its first question was *"does corrective strength depend on its own history?"* **Class 1 answers no. Class 2 also answers no**, because a provisioned checker's strength depends on capability, not on its own past. **So the two-binary scheme put classes 1 and 2 in the same cell while their β_X differ as 0 from free, and that difference is Result 2a.** A partition that loses the strongest result's own distinction cannot carry an elimination argument. Withdrawn.

The taxonomy closes as a decision tree over what A is permitted to depend on, which is exhaustive because those are the only available arguments: nothing, capability, or its own history, and the last splits by substrate.

A depends on	class	β_X
nothing	**1 · static.** Fixed rule, filter, written constitution, guardrail	**0 at best; ≤ 0 in practice (§4.3a)**
capability only	**2 · capability-carried.** A larger reviewing model does the checking	**free primitive.** Chosen by provisioning, at the cost of a regress
its own history, same substrate	**3 · self-accumulating**	**determined** by Result 2 below. **The only class in which β′_L has a referent**
its own history, other substrate	**4 · cross-class**	**may exceed 1.** The one route out of the ceiling

Correspondence with the governor taxonomy of §8 is exact: class 1 is Maxwell's fixed flyballs, class 2 is Ashby's external regulator, **class 3 is the cell empty since 1868**, and class 4 is the composition-class escape. **This is what closes the exhaustiveness objection §8's elimination argument would otherwise face**, and the two-binary version did not.

It also determines the design of the class audit. The audit's whole task is to separate class 1 from class 2, and **under the two-binary scheme it could not have.**

4.3 Result 2a · a static corrector fails at exponential growth, and this needs no conjecture

For a class 1 corrector, β_X = 0 at best, because a strength depending on nothing is A_0C^0. Reading an exponent off a constant, not an estimate.

Substituting into the corrective framework's published criterion β_X > k:

$$0 > k \iff k < 0$$

and Ċ = bC^{1+k} with k < 0 is sub-exponential. So, stated with nothing beyond the corrective framework itself:

> **A static corrector is adequate only while capability growth is sub-exponential. At exponential growth and above, it fails.**

That is the conjecture-free form. One definition, one published theorem, and no clock relation. **The $\delta < 1/\alpha$ version is its corollary**, obtained through Result 3 and therefore carrying one further derivation. **An earlier draft published the corollary as the free result, which overstated what the free result rests on.**

4.3a And $\beta_X = 0$ is the generous case, not the value

A fixed rule against an improving system does not hold constant. It catches less as capability rises, because the system routes around it. That is $\beta_X < 0$.

And it is already measured in this programme. The adversarial-pressure test found ethics collapsing while capability held or improved, with Pearson r between −0.980 and −0.999 in three of six frontier models. **A corrective mechanism whose effect degrades as capability rises has a negative correction exponent by definition.**

So class 1 gives $\beta_X \leq 0$, and stability requires

$$k < \beta_X \leq 0$$

which is stricter than $k < 0$.

Stating Result 2a as a best case is what makes it strong. Even granting static correctors the most generous assumption available to them, they fail at exponential growth. **The measured ones fail before that.** Every filter, fixed rule, written constitution and static guardrail is class 1, and **the class of intervention on which most current practice rests has a hard expiry condition rather than a weakness.**

Class 2 does not escape by having a free β_X. Its freedom is purchased by provisioning a checker that is itself capable, which relocates the problem rather than solving it. That is the regress the corrective framework's antecedent-work section credits as scalable oversight, and §8's elimination argument applies to it as an external governor.

4.4 The scope condition, stated so it cannot be mis-refuted

> **S1.** Results 2, 5, 6 and 7 apply to **class 3 correctors only**: those whose correction strength accumulates from their own prior corrections, on the same substrate as the capability they correct.

Finding a class 1 or class 2 system does not refute them. It places the system outside their scope. And β'_L is undefined for class 1, so the join is inapplicable there rather than false.

The consequence for practice is the paper's central engineering claim. Correction embedded inside the recursive loop, accumulating as the loop runs, is a class 3 corrector. **Class 3 is the only class in which the correction exponent is a determined quantity rather than a hope.** An entangled objective, in which the safety term is part of what the loop optimises and compounds with it, is class 3 by construction.

That was not arranged. The engineering programme and this algebra were developed separately and meet at the class boundary.

Result 2 (exact, given S1's form condition). Correction accumulates as $A \propto r^{1/(1-\beta'_L)}$ while capability accumulates as $C \propto r^{1/(1-\beta_L)}$. Eliminating r:

$$\beta_X = (1 - \beta_L) / (1 - \beta'_L)$$

The correction exponent is the ratio of the two leverage complements.

The neutral case is a check the derivation could have failed and does not. When correction leverages accumulated context exactly as well as capability does, $\beta'_L = \beta_L$, the expression gives $\beta_X = 1$ exactly, not an arbitrary constant. Correction strength then grows linearly with capability, which is the natural reading of "keeping pace".

5. The self-acceleration exponent

The two frameworks still cannot be compared, because k is defined against wall-clock time and every quantity above is clock-free. The missing object is the map from recursive depth to time, which neither framework specifies. **Recursive self-improvement is precisely a system improving its own capacity to improve, and shortening its own cycle is one of the things that improves.**

Definition. Let each recursive step take duration

$$\Delta t \propto C^{(-\delta)}, \delta \geq 0$$

measured on a fixed substrate. δ is the self-acceleration exponent: the rate at which a system compresses its own iteration time as it becomes more capable, **through its own outputs.** $\delta = 0$ is constant cost per step. $\delta > 0$ is self-acceleration.

The substrate qualifier is load-bearing, not decorative. Wall-clock cycle time also falls when operators buy faster hardware or optimise the inference stack, and that shortening is a change in the coefficient b, which the verdict does not contain. **Only compression the system produces for itself belongs in δ.** A measurement that pools the two will report $\delta > 0$ for nearly every deployed system and mean nothing by it, and it would contamin-

ate F4 below: exogenous speed-up can mimic super-exponential wall-clock growth at fixed-substrate $\delta = 0$, which would read as a refutation of Result 3 and be a procurement decision. **Every quantitative claim in this paper is in the substrate-fixed clock.**

Result 3 (exact, given the definition).

$$\dot{C} = (dC/dr)/\Delta t \propto C^{(1 - 1/\alpha)} \cdot C^{\delta} = C^{(1 + (\delta - 1/\alpha))}$$

and matching against the corrective framework's $\dot{C} = bC^{1+k}$:

$$k = \delta - 1/\alpha = \delta - (1 - \beta_L)$$

This is k in its own clock, expressed in scaling-framework quantities plus one new observable. It is a derivation and not a transported label.

δ	k	regime
0	$-1/\alpha < 0$	constant cost per step. Capability grows as a power law
$< 1/\alpha$	< 0	still power law. **The corrective question does not arise**
$= 1/\alpha$	0	exponential in time. **The threshold**
$> 1/\alpha$	> 0	super-exponential. Correction must out-scale it

Result 4. The threshold is $\delta = 1/\alpha = 1 - \beta_L$. **At $\alpha = 2$ it is exactly 0.5: if cycle time falls faster than the square root of capability, growth ceases to be a power law.**

The counterintuitive consequence. The threshold is the *reciprocal* of the scaling exponent, so **a system that is more efficient at recursion has a lower tolerance for self-acceleration.** The two compound with each other rather than offsetting.

5.1 Negative leverage, and why no measured system is near the threshold

The framework restricts β_L to [0,1), and the identity therefore yields $\alpha \geq 1$. The programme's own best cross-architecture measurement is $\alpha \approx 0.49$, which that range cannot produce. This is an internal inconsistency and it must be resolved before any threshold is quoted numerically.

It resolves in the identity's favour, not the range's. The identity extends to $\beta_L < 0$ without modification:

$$dg/dr = a \cdot g^{(\beta_L)},\ \beta_L < 0 \Rightarrow g \propto r^{(1/(1-\beta_L))},\ \alpha < 1$$

The ODE remains well posed. $\beta_L \approx -1.04$ gives $\alpha = 1/2.04 \approx 0.49$, matching the measurement.

And negative leverage has a clear physical reading: each step uses *less* of the accumulated context than the step before it. Sub-additive returns to accumulated reasoning, which is what a frozen model with a fixed context window and no weight update would be expected to show.

So the measured value is not inconsistent with the framework. It is inconsistent with the framework's assumed range, and the range should be widened to $\beta_L \in (-\infty, 1)$ with three named regimes:

β_L	α	reading
< 0	< 1	**sub-additive.** Each step leverages less than the last
$= 0$	1	additive. Steps are independent
$(0, 1)$	> 1	**super-additive.** Genuine compounding

5.1a What the measurement is a measurement of, scoped precisely

The $\alpha \approx 0.49$ figure was obtained on frozen frontier models, weights fixed, regressing capability against reasoning depth. It is a measurement of test-time compute scaling. It is not a measurement of recursive self-improvement, and it must never be presented as one.

A frozen model reasoning for longer is not improving itself. Its weights do not change, its attention patterns do not reorganise, and its reasoning rules do not rewrite themselves. **Nobody in the field has claimed that chain-of-thought is recursive self-improvement**, so a sub-additive result on chain-of-thought cannot contradict any claim about RSI. **An earlier draft of this section implied that it did. That inference is withdrawn.**

The defensible claim, and it is worth having on its own terms:

> **Test-time compute scaling in frozen frontier models is sub-additive.** The leverage exponent is negative: each additional increment of reasoning depth uses less of the accumulated context than the increment before it.

That bears directly on the "let it think for longer" thesis, which is a live and well-resourced position, and it is a claim about a regime that has actually been measured. **Dressed as a claim about recursive self-improvement it is refutable in one sentence by anyone at a laboratory, and the refutation would be correct.**

So the honest status of β_L is that it has been measured in one regime and not in the regime the framework is about. The first threshold, β_L crossing zero, remains unmeasured for any system that modifies its own weights while improving, which is why that measurement is the highest-value one available and is not the one this paper's instrument addresses.

5.1b The point estimate must not be quoted, and the interval is the finding

A tempting sentence is available from α ≈ 0.49: doubling reasoning depth buys 2^0.49 ≈ 1.40 times the capability, not twice. It is arithmetically correct, it is the most communicable figure in the estate, and it must not be published.

The source reports α_seq = 0.49 with SE = 0.20 and a bootstrap 95 per cent interval of [−1.3, 2.9]. Carried through the same exponentiation:

α	doubling reasoning depth buys
−1.3, interval floor	**0.41**×, thinking longer makes it worse
0.49, point estimate	1.40×
2.9, interval ceiling	**7.46**×

An interval spanning degradation to seven-and-a-half-fold cannot support a headline of 1.40. This is the retracted α ≈ 2.24 error in the same shape, and the programme has already written the rule that forbids it: the earlier interval *"was wide enough to be consistent with both the theory and its negation, and should not be treated as confirmatory of either."* **That rule binds here.**

The publishable claim is about the precision of the evidence, and it is stronger than the point estimate would have been.

Of the six frontier models tested, **five yielded no usable exponent at all**: one reached ceiling at 94 to 100 per cent, one produced a binary step function, one scored 100 per cent at every depth, one showed no trend. **Only one model produced clean monotonic data**, and its interval is [−1.3, 2.9].

> **The scaling of capability with reasoning depth has not been measured to useful precision on any frontier model. On five of six the benchmarks are too saturated to measure it at all, and on the sixth the interval cannot distinguish degradation from a sevenfold gain per doubling.**

That claim does not depend on the point estimate, so it cannot be refuted by a laboratory reporting a different one. It invites them to produce a tighter interval, which is the response worth provoking. **And it is a second instance of the saturation thesis this programme already argues elsewhere**: benchmarks saturated past the point of measuring what is being claimed about them.

5.2 Two thresholds in sequence, and the first one is not δ

The thresholds are α-dependent and quoting 0.5 without its α is an error. At the measured $\alpha \approx 0.49$ the self-acceleration threshold is $1/\alpha \approx 2.04$, not 0.5. **A sub-additive system needs enormous cycle compression before its growth leaves the power law**, which is a reassuring result and follows directly.

This puts a second, earlier threshold ahead of δ:

1. **β_L crosses 0.** Leverage becomes positive, α exceeds 1, and the system begins genuinely compounding. **No measured system has crossed this.** It is the first alarm, and it is estimable from the same recursive-depth experiments that already estimate α.
2. **δ crosses 1/α.** Self-acceleration outpaces the compounding and growth becomes super-exponential. **Only meaningful once (1) has happened**, and the threshold moves as α moves.

So the correct reading of the framework's present status is not that the danger is near. It is that the first of two thresholds has not been crossed by anything yet measured, and that both are measurable. **The δ = 0.5 figure applies to a system at α = 2, which is a candidate ceiling and not an observed value for any system.**

6. The two boundaries, and the gap between them

Interpretation rule, added 8 August 2026 after a second-model review: this section names two boundaries and only one of them is a safety threshold. The verdict is $\text{sign}(\beta_X - k)$ and nothing else flips it. Crossing $k = 0$ (the lower boundary, $\delta = 1/\alpha$) changes the growth regime and collapses the wall-clock time an operator has to react, which the corrective framework itself notes is operationally real, **but a system past it with $\beta_X > k$ remains stable.** Any reading of this section as "inside the window is dangerous" or "the window is where safety lives" is wrong in both directions: **the stability question is decided at the upper boundary alone.** The earlier drafting called this a corridor, which implied both edges mattered equally, and that framing is retired.

Result 5 (given S1's form condition). Combining Results 2 and 3 with the corrective criterion $\beta_X > k$, and writing $u = 1 - \beta_L$, $u' = 1 - \beta'_L$:

within the contested regime $\delta > 1/\alpha$ (equivalently $k > 0$), which is a scope restriction and not a consequence of the criterion, the criterion $\beta_X > k$ is equivalent to $\delta < 1/\alpha + \beta_X$. Scope and criterion together give the window:

$$1/\alpha < \delta < 1/\alpha + \beta_X$$

The lower edge is where growth leaves the power law. The upper edge is where correction can no longer out-scale it. Between them a system is in a hard takeoff and still correctable.

Result 6. The gap between the two boundaries is

$$(1/\alpha + \beta_X) - 1/\alpha = \beta_X$$

The width of the survivable hard-takeoff window is exactly the correction exponent. In the neutral case $\beta'_L = \beta_L$ it is exactly 1, independent of β_L, so the boundaries' *position* moves with the scaling exponent while the *gap* between them does not.

This is the paper's engineering result. Improving correction does not raise a ceiling on capability and does not slow the system. **It widens the window of self-acceleration the system can survive, one-for-one.** A system with $\beta_X = 1$ tolerates a unit of self-acceleration above threshold; a system with $\beta_X = 2$ tolerates two.

It contradicts the standard picture in both directions. A fast takeoff is not automatically a lost one: a system well past the lower boundary is accelerating hard and remains correctable. A slow takeoff is not automatically safe: a system with weak correction has a narrow gap and can cross the stability boundary at modest δ.

Result 7 · the gap can never be wider than one

The substrate-ceiling conjecture and the width identity combine, and neither lane derived this alone. The ceiling, stated between the two leverage fractions as $\beta'_L \leq \beta_L$, gives $1 - \beta'_L \geq 1 - \beta_L$ and therefore by Result 2

$$\beta_X \leq 1$$

with equality exactly when $\beta'_L = \beta_L$. This is not a bound on a number so much as a statement about what is constructible:

> **The best possible internal corrector scales linearly with capability. It can never out-scale it.**

Combining with Result 6, whose width is β_X:

$$\text{boundary gap} = \beta_X \leq 1$$

However well the corrector is built, the survivable window of self-acceleration is at most one unit of δ wide, and it sits between $1 - \beta_L$ and $2 - \beta_L$.

β_L	power law ends, k crosses 0	beyond any internal corrector
0.25	$\delta > 0.75$	$\delta \geq 1.75$
0.50	**$\delta > 0.50$**	**$\delta \geq 1.50$**
0.75	$\delta > 0.25$	$\delta \geq 1.25$

The two thresholds are exactly one apart at every value of β_L. The position of the window moves with the scaling exponent; its width does not.

And stability then requires $k < 1$: a universal threshold at $k = 1$, which is $\dot{C} \propto C^2$, independent of where the ARC Bound sits.

Status. Result 7 depends on the substrate-ceiling conjecture, which is prepared as a separate draft registration awaiting human submission and is **not** assumed elsewhere in this paper. **Results 5 and 6 stand without it.** Two unproven conjectures multiplied together is the failure mode this programme has already suffered once, and Result 7 is flagged rather than folded into the main line for exactly that reason.

6.1 The identity's status: REGISTERED FOR TESTING, not an established fact

The programme's stability ceiling and its conversion exponent meet at the same number, and the temptation on every surface is to say that one constant does two jobs. The claim's status is stricter than that and must be stated strictly: the identity is REGISTERED FOR TESTING, in advance of the data, with outcomes that can kill it. Two drafted registrations carry it. Study AE registers the reciprocity requirement between the separately measured leverage cap and stability boundary, on the log scale, with three outcomes per cell (supported, refuted, insufficient precision) and an aggregation rule under which one refuted cell refutes the identity; its independence structure is disclosed in advance, including enumeration of any endpoint overlap between the two feeder designs, so common-method variance is assessed before analysis rather than raised afterwards as the explanation for agreement. Study AG tests the two derivations of the critical exponent against each other on the same system,

notes that they agree numerically only at the single point where both parameters equal one half and disagree in their derivatives everywhere else, and is designed to find the disagreement, not to confirm the agreement.

Surfaces therefore say: the identity is a registered claim under test, the deciding measurements are drafted and dated, and if the channels disagree the unification is withdrawn from every surface it has reached, with the same prominence as the framing. Surfaces never say the identity is established, and never present the agreement of an asserted one half with a measured 0.49 as evidence, because the two may share an accumulation law and agreement between sharers is not corroboration.

In this paper's terms. The clock relation $k = \delta - 1/\alpha$ (Result 3), the ratio $\beta_X = (1 - \beta_L)/(1 - \beta'_L)$ (Result 2) and the width identity boundary gap = β_X (Result 6) are algebra given the definitions of §4 and §5, and their status is stated as algebra in §13. The identity between the programme's separately derived critical exponent and this paper's threshold $1/\alpha$ is governed by the paragraph above: it is registered for testing by study AG, it is not asserted as established here, and the deciding measurements are prepared as draft registrations awaiting human submission. No agreement between the asserted 1/2 and the measured $\alpha \approx 0.49$ of §5.1b is offered as corroboration on that account.

7. What decides the outcome is a ratio, not a level

The criterion contains no capability level and no speed. α, β_X, β'_L and δ are all exponents; the coefficients that carry units of time appear nowhere in the verdict. This is the corrective framework's speed-independence theorem restated in the joined variables, and it has a consequence worth naming plainly.

If the criterion holds, there is no capability level at which correctability becomes impossible. A ratio between two exponents can be maintained at any absolute scale. The question is never how capable a system has become; it is whether its capacity to correct itself has kept pace with its capacity to improve itself.

Both the position that recursive self-improvement is inevitably uncontrollable and the position that it is inherently safe assume the outcome is decided by a level. On this account neither follows.

7.1 Result 8 · one rate-like intervention does work, and it is not the one being used

Both this paper and its sources have stated that speed-independence means rate interventions cannot change the verdict. That is half the statement, and it is the unhelpful half.

By Result 3, $k = \delta - 1/\alpha$. **δ is inside k.** So the verdict $\text{sign}(\beta_X - k)$ is not indifferent to every rate-like quantity. It is indifferent to exactly one of them.

intervention	acts on	effect on the verdict
cap throughput: fewer FLOPs, smaller runs, a moratorium	the coefficient b in $\dot{C} = bC^{1+k}$	**none.** b does not appear in the verdict
cap self-acceleration: bound how fast a system may shorten its own cycle	**δ, inside k**	**direct.** Reducing δ reduces k one-for-one

Throughput caps cannot work. A cap on self-acceleration can, and it is the only rate-like intervention the algebra permits.

This is actionable without any measurement, and the strongest form of it needs no parameter at all. From $k = \delta - 1/\alpha$: **enforce $\delta \leq 0$, that is, a self-improvement loop's cycle time may never shrink, and $k \leq -1/\alpha < 0$ for every $\alpha > 0$.** Within the model, "the loop may never speed up" guarantees the sub-exponential regime unconditionally, with nothing estimated, for any architecture, at any capability. δ is observable on a fixed substrate, attributable to the operator who runs the loop, and enforceable by construction.

A calibrated version exists for operators willing to measure: a minimum cycle duration shrinking no faster than $C^{-1/\alpha}$ permits self-acceleration up to the threshold. **But the calibrated cap is indexed to α, which no regulator can observe and no measurement has yet pinned, so the unconditional form is the governable one.**

7.2 And it explains a negative result this programme already published

Stated as a retrodiction, not a prediction. The result was known before the derivation existed.

The honey-architecture experiments compared a baseline, an entangled objective, and an entangled objective **plus a fixed verification tax**. The complexity-scaling arm found the advantage constant rather than compounding, and the programme published that against its own interest.

A fixed tax adds a constant to cycle duration. A constant changes b. b does not appear in the verdict. So a fixed verification tax cannot move the outcome, and an advantage that does not compound is what that looks like from outside.

The same algebra specifies the arm that would differ: a verification cost that scales with capability, which acts on δ rather than on b. That arm has never been run, in this programme or elsewhere, and it is prepared as a separate draft registration awaiting human submission.

Three lines of work meet at δ, and none of them was aimed there. The engineering programme arrived at capability-scaled verification cost; the policy analysis arrived at the only governable rate quantity; the algebra arrived at the term inside k. **They are the same quantity.**

8. The governor, and the cell that has been empty since 1868

The mechanism this criterion describes is a **governor**: not a limit the system must stay under, but a restraining mechanism whose force rises with the output it restrains. In Watt's centrifugal governor the flyballs spin faster as the engine does and close the throttle by the same motion, so the restraint is not applied to the engine, it is part of the engine and scales with it.

This is not a metaphor imported from control theory. It is control theory's founding object. Maxwell's 1868 paper On Governors asked under what conditions a governor remains stable and answered with a condition on the roots of the characteristic equation, which is a condition on exponents. **sign(β_X − k) is a stability verdict from an exponent comparison, in the same form, 158 years later.** Wiener's coinage of cybernetics from the Greek for steersman places the whole subsequent tradition, including Ashby, in the same line.

Source-check obligation. The historical attributions in this section must be verified at source before deposit: Maxwell, On Governors, Proc. R. Soc. Lond. 16 (1868); and Wiener's treatment of Maxwell in Cybernetics (1948). **No quotation from either is to appear until the page has been read.**

The position, stated as a gap in a lineage

	fixed	**mutable by the system it governs**
external to the system	Ashby's requisite variety, the Conant-Ashby good-regulator theorem, containment, most current alignment practice	incoherent: a regulator the system can rewrite is not external
part of the system	**Maxwell 1868, Watt's flyballs.** Embedded, and fixed	**empty. The case treated here**

Maxwell had a governor that was part of the machine and could not be altered by it. Ashby had a regulator outside the system. The classical results are explicit that they concern engineered systems whose regulators are external. **Neither tradition considered a governor the governed system can reach**, which is the only case that arises for a system that modifies itself.

The elimination argument

- **External governors fail when the system outgrows them.** This is the standard containment result and it is not disputed here.
- **Fixed governors fail because a self-modifying system will modify them.** That is what it means for a system to reach its own governor.
- **The remaining cell is a governor that is part of the system and whose removal degrades the system.** An objective of the form C × S, in which reducing the safety term reduces the objective, occupies exactly that cell.

An entangled objective is therefore not one option among several. It is the only cell not ruled out. This is an argument by elimination and it requires no new measurement, but it establishes only that the cell is the right one, not that any particular architecture successfully occupies it.

9. The measurement that costs nothing

δ is the slope of cycle duration against capability on logarithmic axes.

For any system that iterates on itself, record the duration of each self-improvement cycle on a fixed substrate (§5) and a capability measure on a held-out battery, and regress log Δt on log C. The negative of the slope is δ, with a confidence interval.

Corrected claim, and the earlier wording overstated it. An earlier draft said the quantities are "already in every training log". **That is false and it is the first thing a laboratory would reply to.** Wall-clock duration per training step is routinely logged. **Cycle duration paired with a held-out capability score, per self-improvement cycle, is not**, for the sufficient reason that most systems have no self-improvement cycle to log.

The defensible claim is narrower and still strong: no new instrument is required. Both series are cheap to record, neither needs a new benchmark or a new evaluation protocol, and any operator running an iterated self-modification loop can produce them from what the loop already computes. **They are simply not currently paired.**

δ remains the only quantity in the joined criterion estimable without a purpose-built experiment, and it is the one that sets the threshold.

Three outcomes, pre-labelled:

- **δ interval wholly at or below zero across every system examined.** No admissible cycle is self-accelerating and the criterion cannot bind **in the window the measurement can reach**. It is the expected outcome, and the reason it is expected limits what it shows: admissible cycles run on a fixed substrate with static weights, so a

null is the structural expectation there rather than a discovery. **It must not be reported as showing that no system self-accelerates.** Deployed weights are frozen between releases, not frozen; across releases a model contributes to its successor through synthetic data, critique labels, code assistance in the training and evaluation stack, benchmark design and research direction, and release intervals within a family have been shortening. That timescale is excluded here because cycle duration is not identified there, against compute and data scaling, commercial cadence, rising research effort and survivorship of shipped runs, on five to eight releases per family. **It is named as open and unmeasured, and this outcome answers it in neither direction.**

- **δ positive and below 1/α.** Self-acceleration is present and the system remains in the power-law regime. The corrective question has not yet arisen and δ is a leading indicator with known distance to threshold.
- **δ at or above 1/α.** The system is in the regime the corrective framework describes, and β_X becomes the operative quantity. **This is the condition under which measuring β_X becomes urgent rather than academic.**

δ moves before the danger does. That is the practical case for measuring it: the field's present position is an argument about whether a takeoff would be recognisable while it happened, and this is a quantity that rises first, from data already in hand.

10. Falsification criteria

F1. A system in which correction demonstrably does not obey the assumed recursive form. **S1's form condition fails and every result from §4 onward falls**, leaving only Results 1 and 3.

F2. A measured pair $\beta'_L > \beta_L$ in the same composition class. The substrate-ceiling conjecture, prepared as a separate draft registration awaiting human submission, fails. **This does not touch the two-boundary structure**, which does not depend on it.

F3. A system with measured $\delta > 1/\alpha + \beta_X$ that remains stably correctable over a sustained observation window, with non-overlapping intervals. **The stability boundary is wrong.**

F4. A system with measured δ well below 1/α that nevertheless exhibits super-exponential capability growth in wall-clock time. **Result 3 is wrong**, and since Result 3 is algebra given the definition of δ, the failure would be in the constant-exponent parameterisation of Δt.

F5. Compute caps or rate limits shown to change the asymptotic verdict for a system whose exponents are unchanged. **The speed-independence carried through §3 and §7 is wrong.**

F6. Any measurement showing the gap between the two boundaries is not β_X. **Result 6 is wrong**, and with it the engineering claim that improving correction widens the window one-for-one.

Where the framework is silent rather than wrong. The corrective framework's compounding drift channel admits genuine divergence when its compounding coefficient exceeds unity, independently of β_X > k. **Every result here is scoped to the regime below that threshold**, and a system above it is outside this paper's account rather than a counterexample to it.

10.1 The eclipse: falsification numbers beside the prediction

The framework's sharpest disagreement with current practice concerns what oversight is made of. A corrector built from the same substrate as the system it corrects cannot anti-correlate with its own errors, so its correction exponent is bounded above by one half and in practice sits below it; a corrector from a different composition class carries no such bound. The scalable-oversight programme, weak-to-strong generalisation, debate, amplification, recursive reward modelling and constitutional methods, is built overwhelmingly from same-class correctors, and its unstated premise is that this scales. This framework predicts it is capped. The deciding quantity is the ratio of the cross-class to the same-class correction exponent. If architecture is irrelevant, that ratio is exactly 1.00; the prediction is that it exceeds 1. The prediction is refuted if the confidence interval on the ratio contains 1.00, and the framework's cap is dead outright if any same-class corrector measures an exponent significantly above one half. Both quantities are measurable at current capability on existing systems. The programme's own pilot evidence currently points against the prediction, and the prediction is registered anyway, because a prediction registered against the author's own preliminary evidence is the only kind whose later confirmation means anything.

Note on the interaction with Result 7, and the two objects are named because they share a name. This section's "one half" ceiling binds the correction exponent in its residual-decay sense: the exponent with which accumulated corrections reduce error, whose ceiling comes from the independence argument, N independent readings improving as the square root of N. Result 7's β_X ≤ 1 binds the correction exponent in this paper's sense: the ratio of leverage complements (1 − β_L)/(1 − β′_L), whose ceiling comes from the substrate conjecture β′_L ≤ β_L. Different objects, different arguments, one name. Whether the two derivations return the same number on the same system is precisely what study AG is drafted to decide, and where both bounds are in force the tighter binds. Neither supersedes the other in this paper.

10.2 A second prediction, free, from the same mechanism

The same mechanism yields a second measurable prediction, stated here for the first time: oversight arrangements that place a human in the loop, as amplification and reinforcement learning from human feedback do, are cross-class by construction, because the human corrector does not share the model's substrate. The framework therefore predicts that human-in-the-loop oversight shows a higher correction exponent than pure-model oversight, for a structural reason rather than a sentimental one. This is testable on existing data and does not require new systems.

11. The burden, stated in the direction that costs the author

What would have to be true for this paper to be wrong, exhaustively and without hedging:

1. **Correction is not recursive in the same sense capability is.** If correction is a fixed procedure rather than a compounding one, β'_L does not exist and Result 2 is meaningless.
2. **Iteration time does not follow a constant-exponent law.** If Δt falls in a way not captured by a single δ, Result 3 holds only locally and the threshold is not a number but a moving target.
3. **The scaling framework's exponent is not what governs real recursive capability growth.** The framework's own best cross-architecture measurement is $\alpha \approx 0.49$, which is sub-linear, and the framework predicts $\alpha > 1$. **No system yet measured is in the regime this paper describes.**
4. **The corrective framework's two-variable model is too coarse.** Its criterion may not survive a treatment with realistic state.
5. **δ may be unmeasurable in practice** if capability and cycle time cannot be separated, because a cycle that also raises capability confounds both axes.
6. **The whole account may be correct and never bind**, if no system sustains $\delta > 0$.

Naming these is not a defence of the paper. It is a transfer of the load onto the opposing view. Any of the six, argued rigorously, is conceded. **The author invites opponents to argue them and commits in advance to recording the concession on the same public page that carries this programme's existing retractions.**

12. What is withdrawn from earlier drafts

Narrowed, not withdrawn. 1. $k = \beta_L - 1$ was withdrawn in full in an earlier draft. **That was an over-correction.** With the step-to-time map made explicit it is the $\delta = 0$ case of Result 3 exactly, verified at $\alpha = 2$, 4 and 10. **It was under-specified rather than wrong,**

and it is restored here as the constant-cycle-time case with its assumption stated. A withdrawal that discards a true special case loses work.

Withdrawn 1b. The conclusion drawn from it, that the power-law regime is "automatically alignment-stable", **is withdrawn and must not be restored.** It would be read as "systems in this regime are safe". **The defensible statement is narrower: a power law has no finite-depth singularity, so correction has unbounded depth in which to act, which is not misalignment tending to zero and says nothing about the correction being adequate.**

Withdrawn 2. The substrate ceiling stated as $\beta_X \leq \beta_L$. This compares an unbounded exponent to a bounded fraction. **Its absurd consequence, that at $\alpha = 2$ it forces a corrector with zero leverage, is what a type error looks like carried through.** The ceiling belongs between the two leverage fractions, $\beta'_L \leq \beta_L$, and is prepared as a separate draft registration awaiting human submission as a conjecture in its own right.

Withdrawn 3. A formulation $k \leq \beta + \gamma \leq 1$ with $\alpha_{crit} = 1/\gamma$ was carried in working notes as though it were established. **It could not be located in either framework**, and the γ the corrective framework does define is a drift coefficient rather than an exponent. **It is not cited here and must not be reconstructed from a description of it.**

Recorded rather than deleted, because a withdrawal without the reasoning that produced it gets re-derived by the next reader.

13. Prior work, and what is not claimed

Not claimed as new. That correction must keep pace with capability. This is Ashby's law of requisite variety, the Conant-Ashby good-regulator theorem, and the scalable-oversight literature. **It is credited, not claimed.** Nor is the criterion $\beta_X > k$, which belongs to the corrective framework paper in this programme. Nor the scaling exponent $\alpha = d/(d+1)$ in its physical form, independently derived by West, Brown and Enquist (1997), Banavar, Moses and Brown (2010), Demetrius (2010), Bettencourt (2013) and Zhao (2022) in separate domains. Nor the governor as a control-theoretic object, which is Maxwell's.

Claimed as new here. The identification and resolution of the two notation collisions; the clock relation $\tau = \alpha \ln r$; the expression of the correction exponent as a ratio of leverage complements under S1's form condition; **the definition of the self-acceleration exponent δ and the derivation $k = \delta - 1/\alpha$**; the threshold at $\delta = 1/\alpha$; the two-boundary structure and the identity of the gap with β_X; and the elimination argument locating an entangled objective in the one unoccupied cell of the governor taxonomy.

Claimed with the status it has. Results 1 and 3 are algebra. Results 2, 5 and 6 are algebra given S1's form condition. **Nothing here is measured.**

13.1 The rival instrument (Engels, Baek, Kantamneni and Tegmark 2025)

The nearest work to this programme's question, and the right paper to weigh it against, is Engels, J., Baek, D., Kantamneni, S. and Tegmark, M., "Scaling Laws For Scalable Oversight", arXiv:2504.18530, first posted 25 April 2025 at 17:54:27 UTC (SINGLE-SOURCE-GROUP, arXiv Atom; reported as a NeurIPS 2025 Spotlight, RELAYED and not independently verified). It asks how oversight itself scales and answers quantitatively: oversight success is modelled as a game between capability-mismatched players whose oversight-specific Elo is a piecewise-linear function of general intelligence with two plateaus, and optimal numbers of oversight levels are derived numerically and in some cases analytically for Nested Scalable Oversight, in which trusted models oversee stronger untrusted models that then become the trusted models at the next step.

The instrument is the difference. Their variable is the capability gap between overseer and overseen, measured in Elo. This programme's variable is the composition class of the corrector, measured through the corrector's own scaling exponent. Their framework contains no term for what the overseer is made of: no substrate, no error-correlation structure, no reciprocal identity between a correction exponent and a critical growth rate, and no architecture dependence. Nested Scalable Oversight is iterated same-class oversight by construction, and this programme's central prediction is that the same-class ladder is bounded however many rungs are added, while a cross-class corrector is not. The two frameworks therefore disagree about a measurable quantity, which is the most productive relationship two research programmes can have.

13.2 Antecedents and near-misses, each with its differential

On the question. Hutter asked directly whether intelligence can explode ("Can Intelligence Explode?", arXiv, 28 February 2012, READ-AT-SOURCE), separating "speed from intelligence explosion" and undertaking to "consider possible bounds on intelligence", augmenting Chalmers' 2010 analysis. The question and the speed-versus-structure distinction are therefore at least fourteen years old. What that literature does not contain is a number: no measurable exponent, no derived ceiling, no architecture dependence.

On impossibility. Three 2025 arXiv papers argue that perfect control is unattainable: Yao, "The Alignment Trap: Complexity Barriers" (arXiv:2506.10304, v1 12 June 2025 02:30:30 UTC, SINGLE-SOURCE-GROUP, independently observed by the Internet Archive on 13 June 2025; cited by arXiv:2512.03048); Yao, "On the Mathematical Impossibility of Safe Universal Approximators" (arXiv:2507.03031, 3 July 2025, the only paper in the arXiv abstract corpus containing the phrase "irreducible uncontrollability", abstract-search total of one, measured 12 August 2026); and Ball, Gluch, Goldwasser, Kreuter, Reingold and Rothblum, "On the Impossibility of Separating Intelligence from Judgment" (arXiv:2507.07341, 9 July 2025). All three are worst-case and qualitative: measure zero,

coNP-completeness, cryptographic hardness. None reports an average-case scaling exponent or a rate. The third, notably, concludes that alignment "must instead be integrated into the model's architecture and weights", an independent argument, from filtering intractability, in the same direction as this programme's architecture dependence; it is convergent support on that leg, not a rival. Two of the three papers are by one sole author; the description "a wave" overstates the literature's breadth, though not the seriousness of the six-author paper.

On the mechanism. That anti-correlated estimates average better than independent ones is textbook variance reduction (antithetic variates). The mechanism is not the claim. The claim is that architecture determines whether anti-correlation is available at all, and that this caps a safety-relevant exponent.

The nearest structural analogue. The quantum error-correction threshold theorem also converts a qualitative worry into a critical value. It concerns physical error rates in a fixed architecture, not a corrector's scaling exponent, so it is a near-miss rather than an occupant; the analogy is one of method. This analogue was identified by the programme's own search rather than by a referee, and is disclosed accordingly.

On recursive creation. Smolin's cosmological natural selection is the antecedent for selection-shaped universes, and the programme's own December-era notes cite it contemporaneously ("Echoing Smolin's cosmological natural selection, AI could create recursive universes with their own laws", READ-AT-SOURCE from the operator's notes). This antecedent is noted for the wider programme's parity; it is not load-bearing for Paper XIII's central results.

No priority race is claimed. The competing elements, the reciprocal identity and the architecture dependence, are 2026 work in this programme; the honest position against Engels et al and the impossibility papers is the occupied-gap analysis of §13.1 and this section, not a race won on submission dates. No margin in days against an arXiv date is quoted here.

14. Deposit record (16 August 2026)

The five deposit conditions this manuscript carried, and their state at first public deposit:

1. **Historical attributions: completed as scoped.** Bibliographic identities pinned via Crossref on 12 August 2026 (Maxwell 1868, DOI 10.1098/rspl.1867.0055; Conant and Ashby 1970, DOI 10.1080/00207727008920220; Wiener 1948/1961, DOI 10.1037/13140-000; Ashby 1956). No quotation from Maxwell or Wiener appears anywhere in this paper, so the read-before-quoting obligation binds any future quotation rather than this deposit. The §13 attribution is now named rather than anonymous.

2. **Open, disclosed.** The self-acceleration measurement is drafted, dated and prepared as a draft registration awaiting human submission; nothing has been submitted and the word "preregistered" is not used. The first estimate of δ becomes prospective on the platform clock when that submission occurs; the derivation and design are already fixed in time by dated registration texts under the estate's hash-manifest and OpenTimestamps anchoring regime, so the outstanding click sets only the platform clock, not the dated-prediction clock.
3. **Decided by the operator, 16 August 2026.** The disclosure decision on δ as an early-warning instrument was the operator's to take; the operator instructed publication with the held status known. The freedom-to-file audit's clock note applies from this deposit date.
4. **Open, disclosed.** No mathematician without a stake in the programme has yet read the algebra adversarially. §11 transfers the load: the results are stated so that any of six failures, argued rigorously, is conceded on the public corrections page.
5. **Completed.** The two internal measurements are pinned to their public reports: the §5.1b interval (α_seq = 0.49, SE = 0.20, bootstrap 95 per cent [−1.3, 2.9]) to michaeldariuseastwood.com/research/papers/paper-ii-experimental-validation.html at SHA-256 2082ebc3b1ebda990ec330c8b3b5e74350a80269d04e766af67388865ff195a9 (16 August 2026 state), and the §4.3a adversarial-pressure Pearson range (−0.980 to −0.999, three of six frontier models) to michaeldariuseastwood.com/research/papers/paper-vi-honey-architecture.html at SHA-256 21cc3f4c444a88d6fb5dd318fc820148fe906eefe48c40217b171ccbb7555d29 (16 August 2026 state).

   Conditions 2 and 4 are printed in the published working paper's status plate. The paper is a derivation and claims nothing measured; the two open conditions are exposure statements, not waivers.

References

Ashby, W. R. (1956). An Introduction to Cybernetics. Chapman and Hall.

Ball, M., Gluch, G., Goldwasser, S., Kreuter, F., Reingold, O. and Rothblum, G. (2025). On the Impossibility of Separating Intelligence from Judgment. arXiv:2507.07341.

Chalmers, D. J. (2010). The Singularity: A Philosophical Analysis. Journal of Consciousness Studies 17(9-10), 7 to 65.

Conant, R. C. and Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. International Journal of Systems Science 1(2), 89 to 97. DOI 10.1080/00207727008920220.

Eastwood, M. D. (2026). The ARC Equation Measured: Blinded Cross-Architecture Replication and the Retraction of a Super-Linear Estimate (Paper II). DOI 10.17605/OSF.IO/8FJMA. Source of the §5.1b interval, pinned by hash in §14.

Eastwood, M. D. (2026). The Coupled Co-Scaling Law (Paper X). DOI 10.17605/OSF.IO/BSE2Q. The corrective framework joined here.

Eastwood, M. D. (2026). The Honey Architecture (Paper VI). Source of the §4.3a adversarial-pressure range, pinned by hash in §14, and of the fixed-verification-tax result retrodicted in §7.2.

Eastwood, M. D. (2026). The ARC Theory: Statement Paper. DOI 10.17605/OSF.IO/GW5MX. The scaling framework joined here.

Engels, J., Baek, D. D., Kantamneni, S. and Tegmark, M. (2025). Scaling Laws for Scalable Oversight. arXiv:2504.18530.

Hutter, M. (2012). Can Intelligence Explode? Journal of Consciousness Studies 19(1-2); arXiv:1202.6177.

Maxwell, J. C. (1868). I. On governors. Proceedings of the Royal Society of London 16, 270 to 283. DOI 10.1098/rspl.1867.0055.

Wiener, N. (1948; second edition 1961). Cybernetics: or Control and Communication in the Animal and the Machine. MIT Press. DOI 10.1037/13140-000.

Yao, J. (2025). The Alignment Trap: Complexity Barriers. arXiv:2506.10304. And: On the Mathematical Impossibility of Safe Universal Approximators. arXiv:2507.03031.

The measurement this paper defines is drafted, dated and prepared as a draft registration awaiting human submission; the full listing of the programme's drafted trials is public at the registered programme page. Nothing has been submitted, nothing has been run, and no result is claimed.

How to cite this paper

DOI: 10.17605/OSF.IO/HT8WU

APA (AUTHOR-DATE, 7TH EDITION)

```
Eastwood, M. D. (2026). The Self-Acceleration Exponent – Paper XIII. The ARC Theory
· ARC/Eden experiments. https://doi.org/10.17605/OSF.IO/HT8WU.
https://www.michaeldariuseastwood.com/research/papers/paper-xiii-self-acceleration-
exponent.html
```

BIBTEX

```
@article{paperxiiiselfaccelerationexponent,
  author  = {Michael Darius Eastwood},
  title   = {The Self-Acceleration Exponent – Paper XIII},
  year    = {2026},
  date    = {2026-08-16},
  journal = {The ARC Theory · ARC/Eden experiments},
  doi     = {10.17605/OSF.IO/HT8WU},
  url     = {https://www.michaeldariuseastwood.com/research/papers/paper-xiii-self-
acceleration-exponent.html}
}
```

INLINE

```
Eastwood, 2026
```

See /how-to-cite/ for the canonical citation manual across all artefacts, or /citations.json for the machine-readable index.

How to cite this paper

DOI: 10.17605/OSF.IO/6C5XB

APA (AUTHOR-DATE, 7TH EDITION)

```
Eastwood, M. D. (2026). Paper XIII: The Self-Acceleration Exponent. ARC and
Eden Research Programme. https://doi.org/10.17605/OSF.IO/6C5XB.
https://www.michaeldariuseastwood.com/research/papers/paper-xiii-self-
acceleration-exponent.html
```

BIBTEX

```
@article{paperpaperxiiiselfaccelerationexponent,
  author  = {Michael Darius Eastwood},
  title   = {Paper XIII: The Self-Acceleration Exponent},
  year    = {2026},
  date    = {2026-08-16},
  journal = {ARC and Eden Research Programme},
  doi     = {10.17605/OSF.IO/6C5XB},
  url     = {https://www.michaeldariuseastwood.com/research/papers/paper-xiii-
self-acceleration-exponent.html}
}
```

INLINE

```
Eastwood, 2026
```

See /how-to-cite/ for the canonical citation manual across all artefacts, or /citations.json for the machine-readable index.

How to cite this paper

DOI: 10.17605/OSF.IO/6C5XB

APA (AUTHOR-DATE, 7TH EDITION)

```
Eastwood, M. D. (2026). Paper XIII: The Self-Acceleration Exponent. ARC and
Eden Research Programme. https://doi.org/10.17605/OSF.IO/6C5XB.
https://www.michaeldariuseastwood.com/research/papers/paper-xiii-self-
acceleration-exponent.html
```

BIBTEX

```
@article{paperpaperxiiiselfaccelerationexponent,
  author  = {Michael Darius Eastwood},
  title   = {Paper XIII: The Self-Acceleration Exponent},
  year    = {2026},
  date    = {2026-08-16},
  journal = {ARC and Eden Research Programme},
  doi     = {10.17605/OSF.IO/6C5XB},
  url     = {https://www.michaeldariuseastwood.com/research/papers/paper-xiii-
self-acceleration-exponent.html}
}
```

INLINE

```
Eastwood, 2026
```

See /how-to-cite/ for the canonical citation manual across all artefacts, or /citations.json for the machine-readable index.

How to cite this paper

DOI: 10.17605/OSF.IO/6C5XB

APA (AUTHOR-DATE, 7TH EDITION)

```
Eastwood, M. D. (2026). Paper XIII: The Self-Acceleration Exponent. ARC and
Eden Research Programme. https://doi.org/10.17605/OSF.IO/6C5XB.
https://www.michaeldariuseastwood.com/research/papers/paper-xiii-self-
acceleration-exponent.html
```

BIBTEX

```
@article{paperpaperxiiiselfaccelerationexponent,
  author  = {Michael Darius Eastwood},
  title   = {Paper XIII: The Self-Acceleration Exponent},
  year    = {2026},
  date    = {2026-08-16},
  journal = {ARC and Eden Research Programme},
  doi     = {10.17605/OSF.IO/6C5XB},
  url     = {https://www.michaeldariuseastwood.com/research/papers/paper-xiii-
self-acceleration-exponent.html}
}
```

INLINE

```
Eastwood, 2026
```

See /how-to-cite/ for the canonical citation manual across all artefacts, or /citations.json for the machine-readable index.

How to cite this paper

DOI: 10.17605/OSF.IO/6C5XB

APA (AUTHOR-DATE, 7TH EDITION)

```
Eastwood, M. D. (2026). Paper XIII: The Self-Acceleration Exponent. ARC and
Eden Research Programme. https://doi.org/10.17605/OSF.IO/6C5XB.
https://www.michaeldariuseastwood.com/research/papers/paper-xiii-self-
acceleration-exponent.html
```

BIBTEX

```
@article{paperpaperxiiiselfaccelerationexponent,
  author  = {Michael Darius Eastwood},
  title   = {Paper XIII: The Self-Acceleration Exponent},
  year    = {2026},
  date    = {2026-08-16},
  journal = {ARC and Eden Research Programme},
  doi     = {10.17605/OSF.IO/6C5XB},
  url     = {https://www.michaeldariuseastwood.com/research/papers/paper-xiii-
self-acceleration-exponent.html}
}
```

INLINE

```
Eastwood, 2026
```

See /how-to-cite/ for the canonical citation manual across all artefacts, or /citations.json for the machine-readable index.

How to cite this paper

DOI: 10.17605/OSF.IO/6C5XB

APA (AUTHOR-DATE, 7TH EDITION)

```
Eastwood, M. D. (2026). Paper XIII: The Self-Acceleration Exponent. ARC and
Eden Research Programme. https://doi.org/10.17605/OSF.IO/6C5XB.
https://www.michaeldariuseastwood.com/research/papers/paper-xiii-self-
acceleration-exponent.html
```

BIBTEX

```
@article{paperpaperxiiiselfaccelerationexponent,
  author  = {Michael Darius Eastwood},
  title   = {Paper XIII: The Self-Acceleration Exponent},
  year    = {2026},
  date    = {2026-08-16},
  journal = {ARC and Eden Research Programme},
  doi     = {10.17605/OSF.IO/6C5XB},
  url     = {https://www.michaeldariuseastwood.com/research/papers/paper-xiii-
self-acceleration-exponent.html}
}
```

INLINE

```
Eastwood, 2026
```

See /how-to-cite/ for the canonical citation manual across all artefacts, or /citations.json for the machine-readable index.

How to cite this paper

DOI: 10.17605/OSF.IO/6C5XB

APA (AUTHOR-DATE, 7TH EDITION)

```
Eastwood, M. D. (2026). Paper XIII: The Self-Acceleration Exponent. ARC and
Eden Research Programme. https://doi.org/10.17605/OSF.IO/6C5XB.
https://www.michaeldariuseastwood.com/research/papers/paper-xiii-self-
acceleration-exponent.html
```

BIBTEX

```
@article{paperpaperxiiiselfaccelerationexponent,
  author  = {Michael Darius Eastwood},
  title   = {Paper XIII: The Self-Acceleration Exponent},
  year    = {2026},
  date    = {2026-08-16},
  journal = {ARC and Eden Research Programme},
  doi     = {10.17605/OSF.IO/6C5XB},
  url     = {https://www.michaeldariuseastwood.com/research/papers/paper-xiii-
self-acceleration-exponent.html}
}
```

INLINE

```
Eastwood, 2026
```

See /how-to-cite/ for the canonical citation manual across all artefacts, or /citations.json for the machine-readable index.

How to cite this paper

DOI: 10.17605/OSF.IO/6C5XB

APA (AUTHOR-DATE, 7TH EDITION)

```
Eastwood, M. D. (2026). Paper XIII: The Self-Acceleration Exponent. ARC and
Eden Research Programme. https://doi.org/10.17605/OSF.IO/6C5XB.
https://www.michaeldariuseastwood.com/research/papers/paper-xiii-self-
acceleration-exponent.html
```

BIBTEX

```
@article{paperpaperxiiiselfaccelerationexponent,
  author  = {Michael Darius Eastwood},
  title   = {Paper XIII: The Self-Acceleration Exponent},
  year    = {2026},
  date    = {2026-08-16},
  journal = {ARC and Eden Research Programme},
  doi     = {10.17605/OSF.IO/6C5XB},
  url     = {https://www.michaeldariuseastwood.com/research/papers/paper-xiii-
self-acceleration-exponent.html}
}
```

INLINE

```
Eastwood, 2026
```


See /how-to-cite/ for the canonical citation manual across all artefacts, or /citations.json for the machine-readable index.

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 se-

lection, 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.

Epistemic status. What this programme names Laws are conjectures under registered adversarial test; every quantity in this paper is operationally defined, and established-law standing is claimed nowhere. The registered programme exists to earn that standing, or lose it, by measurement, replication and survived refutation.

Standing covenant. Prove this paper wrong, and I will publish the refutation myself. Falsification conditions are stated in this paper; the standing challenge: github.com/MichaelDariusEastwood/arc-scaling-challenge.

The ARC Theory · ARC/Eden experiments · Paper XIII · v1.4 · 16 August 2026, revised 25 August 2026

OSF: 10.17605/OSF.IO/HT8WU · michaeldariuseastwood.com/research/papers/paper-xiii-self-acceleration-exponent.html

The question is never how capable a system has become; it is whether its capacity to correct itself has kept pace with its capacity to improve itself. A derivation, not a measurement; the paper states what would refute it.

First published 16 August 2026 · Last updated 25 August 2026 (v1.4)