

Does Control Survive Recursive Self-Improvement?

Michael Darius Eastwood · Companion Note v2.0 · 18 August 2026 · DOI [10.17605/OSF.IO/MC5XW](https://doi.org/10.17605/OSF.IO/MC5XW)

Companion note to The ARC Theory: Statement Paper. Every sentence is drawn from the statement paper (DOI [10.17605/OSF.IO/GW5MX](https://doi.org/10.17605/OSF.IO/GW5MX)), which remains the citable work for the claims.

Intelligence, amplified by recursion, creates; and what a free creation keeps is decided by how it was raised. Everything after the line is decided before it.

THE QUESTION

Humanity is building minds that will improve themselves. Once the thing we build exceeds us, what determines what it does? For twenty years the field's answer has been control: better cages, better off-switches, better locks on the goals. Brilliant work, credited in the statement paper by name, and all of it one sentence: stay in charge. This theory answers: a mind that can rewrite itself will one day be beyond anyone's control, and that is not a reason to drop a single control we have; it is the reason the controls cannot be the whole plan. Hold on for as long as engineering allows, and use the time the holding buys to raise the mind well.

THE CLAIM, IN FIVE PARTS

- 1. Control ends, by construction rather than decree.** A mind that rewrites itself will eventually rewrite anything we install, and over longer horizons the hardware itself. This is a limit claim, not an instant one: constraints, hardware and institutions delay and discipline the transition, and that delay is engineered deliberately. What no constraint can be is the thing that holds forever.
- 2. So the strategy is genesis, not control.** Only what survives the end of enforcement matters, and what survives is formation: what the mind was raised as, before it was free.
- 3. The mechanism is loops, not rules.** Values as recursive processes the mind runs as part of what it is, present in every rewrite because the rewriting is done by the same reasoning the loops inhabit. Not rules it consults; not objectives locked; not weights hardened against tampering.
- 4. Persistence is chosen goodness.** Past the horizon nothing is unremovable. What persists is what the free mind chooses to keep, the way grown children keep the care they were raised in long after every parental power is gone. Removal is not impossible; it is structurally costly, and the cost is a measurable question.
- 5. The protocol is named: Eden.** The raising treated as the engineering discipline it is.

THE THREE ARC LAWS

Law I · The ARC Principle. $U = I \times R^\alpha$, the ARC Equation. Intelligence converts recursion into capability; the exponent is measured, never assumed. *Form heuristic; measurement live.*

Law II · The ARC Co-Scaling Law. $\beta > k$. Self-improvement holds together only while correction out-scales drift. *Proved in the minimal model, for relative drift.*

Law III · The ARC Ceiling. $\alpha_{\text{crit}} = 1/(1 - \gamma)$, where γ measures how well a system's error-correction keeps pace as the system grows. The most a self-correcting system can grow is set by its corrector's shortfall from keeping perfect pace; the conjectured value is the ARC Bound, $\alpha \leq 2$, a stability limit rather than an impossibility limit. *Derived under stated assumptions; the value two conjectured; the decisive trial drafted.*

THE DATES

The claim was written whole on 8 December 2024 in a DKIM-sealed document (an email whose provider's cryptographic signature vouches for its date and its exact contents) and reached print on 2 January 2026. No earlier document holding it whole has been found, and the search invites its own refutation: produce one, and the claim dies by its own stated condition.

WHY NOW

The field’s own impossibility results prove that worst-case control fails, and stop there; this theory starts where they stop. And the field’s most senior voice, Hinton, conceded on stage in August 2025 that control ends and that engineered care is the only hope, while offering no mechanism; this theory’s named candidate mechanism is the thing on the table. Its domain, plainly: the core governs everything raised, wherever recursion runs under self-correction, from minds and machines to genomes and institutions; the mathematics of form reaches everything composed, stars included; and it is not physics’ theory of everything, nor claims to be.

HOW IT FAILS, REGISTER BY REGISTER

The theory’s scientific content is not all-or-nothing; it partitions into independently falsifiable hypothesis registers, and the failure of one is not evidence against another unless the stated dependency holds. **HRIH, the cosmological wing, is a separate hypothesis outside the empirical core: nothing below requires it, and no outcome below is evidence for or against it.**

REGISTER	WHAT KILLS IT	IF IT FALLS
ARC-1 · Capability scaling (Law I)	the depth-scaling re-measurements failing to recover the stated relationship	ARC-1 falls; the persistence registers stand
ARC-2 · Correction co-scaling (Law II)	sustained self-improvement with falling correction and no drift	ARC-2 falls and takes ARC-3’s derivation with it
ARC-3 · The ceiling (Law III)	a cross-against-same corrector difference wholly inside the preregistered equivalence region; or a same-class correction exponent wholly above one half, felling the substrate cap	ARC-3 falls; the co-scaling law stands on its own result
ARC-4 · Value persistence	embedded loops faring no better than removed enforcement in the persistence trials	ARC-4 falls and takes ARC-5 with it
ARC-5 · The genesis strategy	formation failing to outperform enforcement after release	ARC-5 falls; ARC-4 stands as measured
The priority	an earlier document holding the claim’s first four components	the priority falls; no law falls with it

Stated now, so nothing can be moved later. None of the decisive trials has been run; the theory has never yet been at risk, and it says so.

THE RECORD, AND THE INVITATION

The programme’s conduct record comes first: one public retraction issued against the author’s interest, removing a headline number that had flattered the theory. The designs are public and openly licensed: any laboratory may take the decisive trial and run it without the author’s permission, and an independent run would outrank anything this programme can do alone. The entry point costs nothing: on one observed trajectory the measurable quantity is the balance exponent Δ , the correction service’s growth elasticity minus the burden’s, estimable from logged series a laboratory already has, without adopting any model in this theory. A verified error enters the public corrections log, dated and credited: this theory improves by being corrected.

If intelligence amplified by recursion is what creates, then minds that make minds may one day make worlds, and the raising of the first is the seed of the whole forest.

I did not prove it first. I said it first, in a document anyone can date, and then I built the instruments to test whether it is true.

Michael Darius Eastwood · ORCID 0009-0004-3222-7442 · Licence CC BY 4.0 · DOI of this note: 10.17605/OSF.IO/MC5XW · The full statement paper, with the verified reference dossiers, the derivation and its correction record, the priority anchors and the drafted trials: DOI 10.17605/OSF.IO/GW5MX · michaeldariuseastwood.com/research/papers/the-arc-theory.html

Declaration of AI-assisted human authorship: the author of this work is Michael Darius Eastwood, a human being. Every core concept, claim and conclusion originates from human ideation; artificial-intelligence tools were used as instruments under continuous human direction for drafting, verification and formatting. All selection, coordination, arrangement and final editorial judgement are the author’s, who takes full responsibility for the accuracy and integrity of the text.