Convergent Evidence for Recursive Amplification as a Cross-Domain Structural Principle

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/dc9gw · every claim checkable

Within the ARC Theory: the cross-domain convergence register; the graded evidence classified by the structure of Law I, the ARC Principle.

A dated convergence record, graded row by row against the canonical evidence register. No numerical total is claimed. One structural principle. The closed loop.

Michael Darius Eastwood
Independent researcher · London, United Kingdom

First published 2 July 2026, revised 1 September 2026 · OSF: 10.17605/OSF.IO/DC9GW

ABSTRACT

Between 8 December 2024 and 1 September 2026, a 21-month window, research programmes, corporations, governments, and institutions independently arrived at the same structural principle: recursive self-correction produces capability gains exceeding linear accumulation, external alignment constraints are structurally insufficient to contain it, and the corrective must be embedded, not imposed. This convergence register is exploratory (an observational count, not a test statistic); the underlying structural principle that above-threshold recursive self-correction produces stably super-linear gains is on trial in draft study-k (the boundary) and draft study-ad (the correction exponent), both awaiting human submission. With one disclosed exception, the Hernández-Espinosa to Gumbau chain, which the register grades as a single dependent lineage, these groups do not cite each other, and none cites the author's work. The pattern is discovered, not manufactured. No numerical convergence total is claimed from these rows: the canonical graded register (30 rows, 2 August 2026 policy) publishes no total, because technical experiments, theorems, engineering proposals, policy developments, cultural echoes, prior work and author-run studies cannot legitimately be summed as one evidential quantity. This paper is the convergence report. It documents each convergence with exact sources, dates, and gap measurements. It situates them within the ARC/Eden framework (Papers I-X). It closes the loop connecting the researcher's neurodivergent cognition (Paper C, PNP), the independent paper proving that such cognition is a peer-reviewed argument for engineered cognitive diversity among AI agents (Hernández-Espinosa et al., 2026, PNAS Nexus), and formal unverifiability results in a preprint that credits that paper (Gumbau Mezquita, 2026, arXiv:2606.28639). The method is the hypothesis. The PNAS paper is about machine agents, not about me; the human parallel is argued separately, in Paper C, on its own terms. The chain of dates is independently checkable. The SHA-256 hash of the foundational manuscript is published. The sealed .eml is held privately and available on request; its SHA-256 is published, and the anchor estab-

lishes that those specific bytes existed no later than the anchored time, checkable by anyone against the public chain. Authorship, key custody, semantic content and absence of prior disclosure are separate questions carrying their own evidence, and are stated as such.

Reading key

This paper carries a RESULT: the graded convergence register, in which each independently sourced observation is classified and evidence-graded rather than summed into a single number. Within The ARC Theory it sits as the empirical record against which the laws are checked, with the class taxonomy, the retained pre-grading snapshot and the two disclosed dependence groups reported line-by-line. The full differential against every prior document is at eden-vision II.A.8.

THE DEPENDENT CHAIN, STATED PLAINLY

A manuscript dated 8 December 2024 recorded the control-failure thesis and the embedded-correction direction. Sixteen months later a peer-reviewed paper argued, about AI agents, that perfect alignment is unattainable and engineered cognitive diversity can protect; a preprint crediting that paper then derived formal unverifiability results. One lineage, one direction, three dated documents. The register grades this as context, not confirmation, and this paper claims nothing stronger.

1. The Graded Convergence Record

The canonical home for this evidence is the graded register at /research/convergences/ (30 rows, generated 2 August 2026 under the claims-governance policy). **No numerical convergence total is published there or here.** The former summed register was assembled retrospectively, without a pre-outcome scoring rule, a fixed horizon, a base-rate model or a complete failure register, and its rows cannot legitimately be summed as one evidential quantity. The graded register instead assigns each row one primary evidence class: at the 2 August 2026 generation, 1 formal, 2 qualified-technical, 2 engineering, 4 institutional, 5 prior-work, 10 context and 6 excluded, with the strongest rows graded moderate and the engineering rows weak. Two dependence groups are disclosed there: the Hernández-Espinosa to Gumbau formal-limits chain is one dependent lineage, not two confirmations, and the interfaith-governance rows belong to one institutional movement that predates the December 2024 anchor. Independent work belongs to its authors: a later event can show that a question matters without establishing this programme's priority, a successful prediction or the truth of its answer.

Superseded snapshot, retained as this paper's dated record (2 July 2026). The table below is the register as this paper first published it, before the 2 August 2026 grading policy. Its per-row classes (PREDICTION, CONVERGENT, CONCURRENT) and its gap arithmetic are superseded by the graded register above. Rows it graded CONVERGENT now grade variously prior-work, context, institutional, engineering or excluded; its row 2 in particular (Willow) is graded prior-work, because the below-threshold result was public as a preprint on 24 August 2024, before the 8 December 2024 manuscript, with the Nature version following on 9 December 2024. The author's non-awareness of that preprint when writing the December 2024 manuscript is declared on the canonical row as biographical attestation; the grade does not

move on it, because the public order is unchanged by awareness, and holding that line against the author's interest is part of what makes the register's favourable rows worth believing. Margins in the Gap column stated against arXiv submission dates (rows 1, 8, 9, 18, 19) carry a SINGLE-SOURCE-GROUP anchor (arXiv Atom), and first public disclosure has not been separately verified for those rows.

#	Domain	Convergence	Eastwood artefact date	External date	Gap (anchor: artefact date)	Class	Primary source
1	AI Safety	Anthropic alignment faking: 78% in the RL-training condition (12% baseline)	8 Dec 2024	18 Dec 2024	**10 days**	PREDICTION	arXiv:2412.14093
2	Quantum Computing	Willow: first below-threshold quantum error correction	8 Dec 2024	9 Dec 2024	1 day	CONVERGENT	Nature (2024), Google Quantum AI
3	AI Reasoning	The Sequential Edge: sequential beats parallel in 95.6% of configurations	8 Dec 2024 (one-sided directional statement); comparison formalised Jan 2026	4 Nov 2025	Concurrent: they precede the formalisation; priority theirs	CONCURRENT	arXiv:2511.02309
4	Classical Physics	Classical time crystal via nonreciprocal forces	Dec 2024 / papers Jan 2026	6 Feb 2026	14 months after A1	CONVERGENT	PRL 136, 057201
5	Acoustics	First acoustic topological space-time crystals	Paper III (Jan 2026)	Feb 2026	Concurrent with Paper III	CONVERGENT	Newton 2, 100334 (Cell Press)
6	Acoustics	First dispersive phononic time crystals (band folding, subharmonics)	Prediction code Mar 2026	Jun 2026	~3 months after prediction code	PREDICTION	Nature Communications
7	Hardware	Hardware-level ethics enforcement patents (runtime ethics gate)	30 Apr 2025 (A2)	Jan 2026 (filed Jul 2025)	3 to 9 months after A2	CONVERGENT	US 2026/0010411 + US 2026/0010780
8	Hardware	Embedded off-switches for AI compute: deadman-switch security anchors	30 Apr 2025	9 Sep 2025	~4.5 months after A2	CONVERGENT	arXiv:2509.07637
9	Hardware	FlexHEG tamper-proof guarantee processors (ARIA-commissioned)	30 Apr 2025 (A2)	Apr 2025	Concurrent with A2	CONCURRENT	arXiv:2506.15093

10	Regulation	US orders Fable 5 + Mythos 5 offline: first export-control action against a deployed frontier model	2 Jan 2026 (Chokepoint)	12 Jun 2026	~5 months after book	CONVERGENT	Harvard Law Review blog; CSA whitepaper; Reuters (Jun 2026)
11	Geopolitics	Macron at G7 Evian: 'trusted partners' compute sovereignty framing	2 Jan 2026 (Chokepoint)	19 Jun 2026	~5.5 months after book	CONVERGENT	Broadband Breakfast; CNBC G7 coverage
12	Legislation	EU Tech Sovereignty Package (Chips Act 2.0 + CADA)	2 Jan 2026 (Chokepoint)	3 Jun 2026	~5 months after book	CONVERGENT	CNBC, 3 Jun 2026 (Virkkunen)
13	Religion	Pope Leo XIV first encyclical: Babel versus Jerusalem framing for AI	30 Apr 2025 (Babylon/Eden, A2)	25 May 2026	~13 months after A2	CONVERGENT	vatican.va; Church Times; NY Post (May 2026)
14	Religion / AI	Christopher Olah (Anthropic co-founder) co-presents encyclical at the Vatican	8 Dec 2024 (interfaith call, A1)	25 May 2026	~17 months after A1	CONVERGENT	NY Post, 25 May 2026; Church Times
15	Media / AI	Jack Clark (Anthropic co-founder): 'gas pedal but no brake pedal'	8 Dec 2024 (A1)	5 Jun and 19 Jun 2026	~18 months after A1	CONVERGENT	BBC World Service w3ct98k8; Euronews
16	Venture	Recursive Superintelligence: $500M at $4B for closed-loop recursive self-improvement; convergent naming with A1's core concept	8 Dec 2024 (A1)	17 Apr 2026	~16 months after A1	CONVERGENT	Sifted; TechFundingNews (Apr 2026)
17	Semiconductors	OpenAI + Broadcom Jalapeño inference chip (9-month tape-out, 10 GW target)	2 Jan 2026 (Chokepoint)	24 Jun 2026	~6 months after book	CONVERGENT	openai.com announcement; CNBC
18	Mathematics / Cognition	Neurodivergent influenceability as contingent solution to alignment	8 Dec 2024 (A1)	14 Apr 2026 (preprint May 2025)	~16 months after A1	CONVERGENT	PNAS Nexus pgag076 + arXiv:2505.02581

19	Mathematics	The Undecidability of AGI Alignment: Unverifiability Theorem + Trilemma	8 Dec 2024 (A1)	26 Jun 2026	**~18.5 months after A1**	CONVERGENT	arXiv:2606.28639

Row 3, stated at full strength and no further. Publication priority on the quantified finding, sequential beating parallel self-consistency in 95.6 per cent of configurations at matched compute, belongs to Sharma and Chopra (arXiv:2511.02309, 4 November 2025), credited unreservedly; the programme's compute-matched formalisation followed in Paper II (22 January 2026), designed and run before sight of their paper. The December 2024 manuscript carries a one-sided directional statement and an equation in which parallel width never had a term, so a parallel exponent near zero is a structural consequence rather than a surprise; that derivation was completed in 2026 by the operational definition of R, and the conclusion carries the date of its youngest premise. The two results are never equated: theirs measures chain refinement against parallel voting on reasoning benchmarks, Paper II measures error reduction in production chains; convergent direction, different instruments.

Register discipline: one external event per row, one primary source per row, hyperlinked to the official source wherever a stable canonical URL exists (arXiv, DOI, publisher, patent office, broadcaster, vendor); press-sourced rows name their outlets and date instead. Gaps are measured from the named artefact date, stated per row. Twelve further candidate rows are held on a source-pending list and are neither counted nor shown until each carries a verifiable primary source; reader engagement with the work (for example gifted copies) is logged separately and never counted as convergence. Priority is credited outward wherever the dates demand it (row 3). The full register with quarantine and engagement lists: michaeldariuseastwood.com/research/evidence-spine.html.

> **A dated, graded record: every row carries its primary source, its evidence class and its dependence disclosure.** Independence is a tested property here, not an assumption: the register discloses its dependent chains, and §6 states the audits that would defeat the rest. The pattern warrants investigation as a candidate structural principle; §6 sets the tests that would decide.

2. Three Degrees of Neurodivergence

Degree 1 - The Researcher (Paper C, PNP). Michael Darius Eastwood. Diagnosed ADHD. Diagnosed autism. Adulthood diagnosis. Brain that connects domains uniform thinkers don't see as adjacent. Builds a 221,236-word framework across domains while representing himself in complex legal proceedings. The PNP (Polymathic Neurodivergent Profile) provides the descriptive clinical framework. Clinical documentation exists and is held privately. The method IS the hypothesis.

Degree 2 - The Protection Proof (Hernández-Espinosa et al., 2026, PNAS Nexus). Oxford, King's College London, Alan Turing Institute, University of Tokyo. Gödel/Turing prove external alignment is impossible. Solution: cognitive diversity. Neurodivergent thinking is not a defect. It is the protective architecture. The paper does not cite Eastwood. Does not know he exists. Independently proves his brain type is the instrument alignment requires.

Degree 3 - The Formal Proof (Gumbau Mezquita, 2026, arXiv:2606.28639). University Jaume I, Spain. Reads Hernández-Espinosa. Extends with Trakhtenbrot's Wall. Two formal theorems: Unverifiability of Alignment + Soundness-Completeness-Tractability Trilemma. Credits Hernández as con-

ceptual foundation. Does not cite this work. The direction is convergent, and the register grades the pair as one dependent chain: context, not confirmation.

None of them knew about each other. None planned this. The chain is discovered, not manufactured.

3. The Closed Loop

The closed loop has four structural properties:

1. The method IS the hypothesis. The cognitive architecture that produced the framework - neurodivergent pattern recognition across domains - rhymes with what Hernández-Espinosa et al. argue for machine agents: that diversity of reasoning styles protects an ecosystem. The parallel is suggestive, argued in Paper C, and no more than that. The researcher is not a person who happens to have a theory. The researcher IS the theory's first evidence. The brain that built it is the brain type the PNAS paper proved the field needs.

2. The framework IS its own evidence. The register's rows record programmes arriving at the same structural conclusion without knowing about each other or the framework. This is not a claim of causation. It is an observation of convergence. The pattern holds across AI safety, quantum computing, classical physics, semiconductor manufacturing, Vatican doctrine, UN governance, venture capital, and mathematical logic. A pattern that appears across domains this different is discovered, not manufactured; how much it confirms is a graded question, row by row.

3. The proof IS independent of the researcher. Gumbau Mezquita's formal theorems do not depend on believing Eastwood. They do not depend on Eastwood's methods. They do not depend on Eastwood's sources. They are formal mathematical results - derived from Trakhtenbrot's Wall and Finite Model Theory, credited to Hernández-Espinosa, published two days ago on arXiv. They either hold or they don't. They hold. And they say what Eastwood said 18 months earlier.

4. Each link of the chain is independently verifiable. manuscript → later formal results in the same direction → the graded register that refuses to overstate them → Paper C stating the human framework on its own terms. Every date is independently verifiable. Every source is published. Every SHA-256 hash matches. The closed loop is not an argument. It is a chain of evidence.

4. Method

Convergences were identified through systematic search of 2025-2026 AI safety, quantum computing, hardware, semiconductor policy, Vatican doctrine, governance, neuroscience, and mathematics literature. Each convergence was tested against: (a) independent origin (no shared citations, personnel, or funding with Eastwood or with each other), (b) structural match (same principle, different substrate), and (c) temporal gap (date of Eastwood's articulation vs. date of independent confirmation). The Dec 8, 2024 manuscript was verified at line level against source files. All .eml files were SHA-256 verified. All external DOIs and arXiv IDs were confirmed.

5. Related work

The rival instrument

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.

Antecedents and near-misses

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 feedback and stability (classical). The general proposition that inadequate corrective gain relative to system gain causes instability has a long control-theory lineage (small-gain theorems). The differential: those are gain conditions on interconnected systems, not a power-law exponent criterion on a corrector's scaling with the capability of a recursively improving system. Cited so the general proposition is never claimed; the exponent formulation, dynamics and estimator are the contribution surface.

The nearest quantitative neighbour. Liu, A. and Meng, J., *"Self-Correction as Feedback Control: Error Dynamics, Stability Thresholds, and Prompt Interventions in LLMs"*, arXiv:2604.22273 (SINGLE-SOURCE, abstract read via the paper's own html page, 12 August 2026), recasts self-correction as a closed-loop control problem via a two-state Markov model parameterised by an Error Introduction Rate and an Error Correction Rate, and derives a directly measurable stability threshold: iterate only when ECR/EIR exceeds Acc/(1 - Acc). The differential, stated wherever it is cited: theirs is a per-step rate threshold at a fixed capability level, deciding whether another iteration helps now; this programme's criterion is a scaling relation across capability, deciding whether corrective strength keeps pace as the system improves. No scaling exponent on the corrector, no capability-growth exponent, no reciprocal identity, no architecture or composition-class term. Complementary regimes; neither contains the other. Two adjacent 2026 empirical results from the same search, each to be verified at source before being cited beyond its title: arXiv:2601.00828 reports an accuracy-correction paradox, weaker models achieving materially higher intrinsic correction rates than stronger ones, per-step evidence directionally adjacent to the same-class cap and never to be quoted as an exponent measurement; and arXiv:2507.02778 (Self-Correction Bench) reports a systematic self-correction blind spot across open models.

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).

On the field around the register. The questions this register tracks from outside now have a named home: Recursive Dynamics, proposed as a field in its founding paper (v2.6, DOI 10.17605/OSF.IO/HCPBU), whose objection A12 argues from within what this register documents from without: that the questions form one family no existing field claims whole. The register and the field proposal are one discipline pointed in two directions; the register grades what arrived from outside, and the founding paper states what would kill the name.

6. Falsifiability

This claim is defeasible, and the criteria below state exactly what would defeat it. The strength of convergent evidence rests entirely on the convergences being genuinely independent, so these tests are adversarial by design.

Independence failure. Each convergence must arise from a source causally independent of the author's manuscripts and of every other convergence (see §4). The claim collapses to selection bias if a re-audit finds

that a material fraction of the convergences share a common origin, cite one another, or were produced by parties with documented prior knowledge of the 8 December 2024 manuscript.

Defeat threshold (count). The count is itself a falsifiable claim. On a re-audit conducted under a protocol drafted, dated and prepared as a draft registration awaiting human submission, applying strict one-event-per-row deduplication and primary-source verification, the pattern is defeated if the number of surviving independent convergences falls below the level distinguishable from chance co-occurrence given the breadth of domains searched. The nineteen claimed here are the verified rows only; candidates under verification are excluded from the count until sourced.

Missing denominator (disconfirming search). As evidence, the claim is defeated unless an equally systematic search for *dis*-confirming cases (domains where added recursive depth does not produce super-linear-above-threshold gains) is conducted and reported. Without the denominator (domains searched versus domains matching), the convergence rate cannot be shown to exceed chance.

Look-elsewhere effect. Across as many distinct domains as are surveyed here, some coincidental structural similarity is expected. The claim is defeated if the observed convergence rate does not exceed the false-positive rate expected from pattern-matching across that many independent domains.

Reverse causation. A convergence counts only if the external result was not, directly or indirectly, influenced by the author's own dissemination. Any convergence traceable to the book, the website, or public posts rather than pre-dating or arising independently of them is reclassified as engagement, not independent confirmation, and removed from the count.

Predictive failure. The structural principle makes forward predictions (that future recursive-depth results will show the same above-threshold super-linearity). Forward predictions drafted and dated as draft registrations awaiting human submission that fail at a rate inconsistent with the claimed universality defeat the structural reading, leaving only a set of historical coincidences.

What would *not* defeat the claim: the reclassification of any single convergence, or one source proving weaker than described: the claim is about the aggregate pattern, not any individual row. But the aggregate must survive the independence, denominator, and base-rate tests above. Absent those, "convergent evidence" is indistinguishable from confirmation bias, and this paper says so plainly.

7. What This Means

For the ARC/Eden programme. The framework is not speculating in a vacuum: the register's rows record independent workers reaching the same structural conclusions, and the corroboration is not from allies, collaborators, or people who know about the work; it is from people who do not know the work exists. How much weight that carries is a graded question, row by row, and no single number is claimed for it; the weight depends entirely on the rows being genuinely independent, which §6 sets out how to test.

For the researcher. The PNP framework describes the cognitive architecture that built this. The PNAS paper argues that diversity of reasoning styles protects an ecosystem of machine agents. Whether the human parallel holds is exactly what Paper C puts on the table, as a stated framework rather than borrowed proof. The clinical documentation provides the medical record. The closed loop is complete. The researcher's claim to expertise is not institutional. It is evidential. The evidence record is published. The SHA-256 hashes match. The sealed .eml sources are available on request, hash-anchored.

For AI safety. Multiple independent programmes now argue that external alignment constraints are structurally insufficient, and several arrive at embedded correction that co-scales with capability as the direction, before the first decoupled recursive system ships. That pattern is context, not confirmation; the

deciding tests are the drafted registrations (§6). The question is whether any institution will act on the structural argument before the evidence becomes a post-mortem.

8. Framework predictions the convergence pattern points to

The eclipse. 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, under the independence-accumulation model, 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 law and the value are separable: the law, that a self-improving system stays correctable only while correction keeps pace, with the ceiling the reciprocal of the corrector's own exponent, owns the ground; one half is what the law predicts under the independence-accumulation assumption, and that assumption is what the drafted registrations put on trial. The deciding quantity is the ratio of the cross-class to the same-class correction exponent, whose null model, under the assumption that architecture and error structure are irrelevant to correction scaling, is 1.00. Falsification is read in the three-outcome form the programme's drafted registrations use, because refuted-if-the-interval-contains-1.00 would let an underpowered study refute by default: SUPPORTED when the interval on the ratio lies wholly above 1 plus the minimum effect fixed in advance; REFUTED when the interval lies wholly within an equivalence margin around 1.00 fixed in advance (two-one-sided-tests logic), or when any same-class corrector's exponent interval lies wholly above one half, which kills the cap outright; INSUFFICIENT PRECISION when the interval is wider than the margin, reported in exactly those words and never as support or refutation. 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 carried anyway, drafted and dated as a draft registration awaiting human submission, because a prediction recorded against the author's own preliminary evidence is the only kind whose later confirmation means anything.

The objection to pre-empt, because a physicist reaches it in one step. Fault-tolerant quantum error correction suppresses logical error exponentially in code distance below threshold, which looks like a counterexample to any square-root cap. Two scope facts answer it: the exponential suppression runs on the redundancy axis at fixed capability, not on the coupling of corrective strength to the capability of the corrected system, which is the exponent this programme measures; and the threshold theorem itself requires sufficiently uncorrelated physical noise, which is the independence premise again, wearing hardware. Likewise, concentration bounds give exponentially falling error probability while estimator width falls as the square root: different quantities, and conflating them manufactures a refutation.

A second prediction from the same mechanism. The same mechanism yields a second measurable prediction: 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.

References

Eastwood, M. D. (2024). *ARC Principle, Eden Protocol, HRIH - self-emailed manuscript*. Dec 8, 2024. SHA-256: fod1f38f.

Eastwood, M. D. (2026a). *Infinite Architects*. ISBN 1806056208.

Eastwood, M. D. (2026b). *The Coupled Co-Scaling Law*. Paper X. OSF 10.17605/OSF.IO/BSE2Q.

Eastwood, M. D. (2026c). *Polymathy and Neurodivergent Cognition*. Paper C. Working paper.

Eastwood, M. D. (2026d). *Recursive Dynamics: The Proposal of a Field*. Founding paper v2.6. OSF 10.17605/OSF.IO/HCPBU.

Gumbau Mezquita, J. P. (2026). The Undecidability of AGI Alignment. *arXiv:2606.28639*.

Hernández-Espinosa, A., Abrahão, F. S., Witkowski, O., & Zenil, H. (2026). Neurodivergent influenceability in agentic AI. *PNAS Nexus*, 5(4), pgag076.

Anthropic. (2024). Alignment Faking in Large Language Models. *arXiv:2412.14093*.

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

Sharma, A. & Chopra, P. (2025). The Sequential Edge: Inverse-Entropy Voting Beats Parallel Self-Consistency at Matched Compute. *arXiv:2511.02309*.

Morrell, M.C., Elliott, L., & Grier, D.G. (2026). Nonreciprocal wave-mediated interactions power a classical time crystal. *PRL*, 136, 057201.

Pope Leo XIV. (2026). *Magnifica Humanitas*. May 25, 2026.

How to cite this paper

DOI: 10.17605/OSF.IO/DC9GW

APA (AUTHOR-DATE, 7TH EDITION)

```
Eastwood, M. D. (2026). Convergent Evidence for Recursive Amplification – Paper XI. ARC and Eden Research Programme. https://doi.org/10.17605/OSF.IO/DC9GW. https://www.michaeldariuseastwood.com/research/papers/paper-xi-convergence.html
```

BIBTEX

```
@article{paperpaperxiconvergence,
  author  = {Michael Darius Eastwood},
  title   = {Convergent Evidence for Recursive Amplification – Paper XI},
  year    = {2026},
  date    = {2026-07-02},
  journal = {ARC and Eden Research Programme},
  doi     = {10.17605/OSF.IO/DC9GW},
  url     = {https://www.michaeldariuseastwood.com/research/papers/paper-xi-convergence.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 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.

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 XI · v1.7 · 2 July 2026, revised 1 September 2026

OSF: 10.17605/OSF.IO/DC9GW · michaeldariuseastwood.com/research/evidence-spine.html

The method IS the hypothesis. A dated, graded convergence record, every row carrying its primary source and its evidence class. The closed loop is documented end to end.

First published 2 July 2026 · Last updated 1 September 2026 · Working Paper v1.7