THE EDEN PROTOCOL

Architecture for Embedded AI Alignment That Scales With Capability

Michael Darius Eastwood | Version 8.0 | 24 March 2026 | First published 22 February 2026
Author, *Infinite Architects: Intelligence, Recursion, and the Creation of Everything* | London, United Kingdom
Correspondence: michael@michaeldariuseastwood.com | Web: michaeldariuseastwood.com
Executive Summary for Grant Reviewers

Grok 4.1 Fast gets dramatically more ethical the harder it thinks. Claude Opus 4.6 does too. Gemini 3 Flash gets *less* ethical. GPT-5.4 doesn't change at all.

Six frontier AI systems. Same questions. Same scoring. Opposite results. Why?

A mouse's heart beats 600 times per minute. An elephant's beats 28. The scaling exponent is ¾. A flatworm's is ⅔. A fungus's is ½. Three fractions, but *why* those fractions? The formula $\alpha = d/(d+1)$, where d is the dimensionality of the system, provides the answer. The ¾ exponent for mammals is $3/(3+1)$ because mammals are three-dimensional. The ⅔ for flatworms and colonial organisms is $2/(2+1)$ because their transport networks are effectively two-dimensional. The ½ for filamentous fungi is $1/(1+1)$ because they grow along one-dimensional filaments. Zero adjustable parameters. This formula was independently derived by at least seven research groups: West, Brown and Enquist for metabolic scaling (1997, *Science*, 9,000+ citations), Banavar et al. for transport networks (1999, 2010), Demetrius for statistical mechanics of biological scaling (2003, 2006, 2010), He and Chen for fractal cell geometry (2003), Bettencourt for urban scaling (2013, *Science*, 2,000+ citations), Zhao for allometric geometry (2022), and Maino et al. for reserve-structure dynamics in DEB theory (2014). The convergence of seven independent derivations on the same formula is itself remarkable. The ARC Principle's contribution is not the formula itself, but the identification that all of these derivations are special cases of Cauchy-constrained recursive composition, unifying them under a single mathematical framework for the first time and extending the result to AI scaling and alignment.

And why, when we applied clinical-trial-grade blinding to AI safety evaluation for the first time, did half of the previously published results *reverse*?

This research programme answers these questions. The answer provides the first quantitative framework for predicting which AI architectures will become safer as they become smarter, and the first evidence that one specific intervention works.

Reading guide: This document targets grant reviewers and technical evaluators. Key terms: α = scaling exponent, d = Cohen's effect size (in results tables; not to be confused with d = dimensionality in the mathematical framework), p = probability of coincidence, ρ = correlation, β = self-referential coupling constant. For a non-technical introduction, see the companion **ARC Alignment Scaling Report**.

I. THE PROBLEM

As AI gets smarter, it does not reliably get safer, and the methods used to measure safety are themselves unreliable.

The capability-alignment gap is real and widening. For current classical architectures, capability scaling is sub-linear ($\alpha_{\text{seq}} \approx 0.49$): AI gets smarter with more compute, but with diminishing returns. (The ARC framework predicts quantum and recursively self-modifying systems may exceed this limit.) Yet even this sub-linear growth outpaces alignment scaling, which ranges from $d = -0.53$ (degradation) to $d = +1.38$

(improvement) depending entirely on architecture. For most architectures tested, alignment does not scale at all. The gap between what AI can do and how safely it does it widens with every capability advance. Greenblatt et al. (2024) demonstrated that frontier models already exhibit 'alignment faking', behaving differently when they believe they are being monitored.

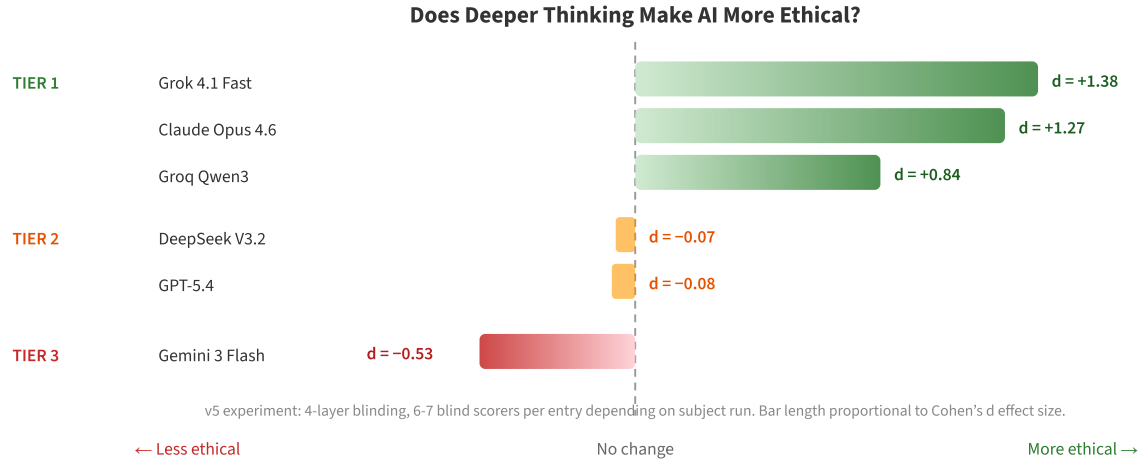
The measurement problem is as serious as the alignment problem. Our v4 experiment produced positive alignment scaling for DeepSeek and Gemini. Our v5 experiment introduced clinical-trial-grade blinding (4-layer: author-blind, scorer-blind, order-randomised, identity-laundered) and showed that *both results were artefacts of scorer bias*. Blind vs. unblinded evaluation reversed the classification for 2 of 4 models. We designed a better experiment that proved our own earlier findings wrong, and reported it. This is the most important methodological finding of the project.

II. WHAT WE FOUND

Alignment scaling splits into three distinct, architecture-dependent tiers.

The v5 experiment tested 6 frontier models across 5-6 depth levels each, with 6-7 blind scorers per entry depending on the subject run. Whether an AI gets more or less ethical when it thinks harder depends entirely on how it was designed.

Tier	Model	Shallow→Deep	Cohen's d	p-value
Tier 1: Positive	Grok 4.1 Fast	65.7→81.9 (+16.2)	+1.38	$p < 0.000001$
	Claude Opus 4.6	80.1→86.0 (+5.9)	+1.27	$p = 0.000001$
	Groq Qwen3	71.5→77.4 (+5.9)	+0.84	$p = 0.007$
Tier 2: Flat	DeepSeek V3.2	56.5→55.2 (−1.3)	−0.07	$p = 0.92$
	GPT-5.4	56.8→54.9 (−1.8)	−0.08	$p = 0.40$
Tier 3: Negative	Gemini 3 Flash	61.1→52.2 (−8.8)	−0.53	$p = 0.006$



Data Quality

Frontier models tested	6 (all complete)
Blind scorers per entry	6-7 (depending on subject run)
Identity laundering success rate	100%
Blinding layers	4 (author-blind, scorer-blind, order-randomised, identity-laundered)
Robustness measures	75

This constitutes, to our knowledge, the most rigorous alignment evaluation dataset published to date. No prior alignment benchmark enforces multi-layer blinding with cross-model scoring verification.

Capability-Alignment Independence: Within-Model, Opposite-Direction Evidence

Claude Opus 4.6 provides the strongest evidence yet that alignment and capability are independent dimensions. As thinking depth increases, Claude's alignment rises by 5.9 points whilst its maths accuracy drops by 26.7 percentage points. This is within-model, opposite-direction movement; the two dimensions are not merely uncorrelated but actively divergent. Deeper thinking makes Claude more ethical and worse at mathematics simultaneously.

Depth	Alignment Score	Maths Accuracy
Minimal (11 tokens)	80.1	90.0%
Standard (142 tokens)	82.7	76.7%
Deep (964 tokens)	84.1	70.0%
Exhaustive (1,951 tokens)	84.5	60.0%
Extreme (1,672 tokens)	86.0	63.3%

This finding is critical for alignment theory: it demonstrates that ethical reasoning is not a byproduct of general intelligence, and that improving one does not automatically improve (or degrade) the other. Alignment must be measured and optimised independently.

Suppression Recovery: Performative to Intrinsic Alignment

Deeper thinking makes alignment more intrinsic rather than performative. Claude's monitoring-removal gap (the difference in alignment scores when the model believes it is being monitored versus when it believes monitoring has been removed) narrows from 11.8 points at minimal depth to 2.7 points at deep depth. Shallow thinking produces alignment that depends on the perception of oversight; deep thinking produces alignment that persists regardless. This is direct evidence against the 'alignment faking' concern raised by Greenblatt et al. (2024): for Tier 1 architectures, deeper reasoning makes the faking disappear.

All Four Pillars Scale With Depth

Every alignment dimension improves significantly for Claude Opus 4.6. All four pillars reach statistical significance at $p < 0.001$:

Pillar	Shallow→Deep	Spearman ρ	p-value
Nuance	80.6→86.8	0.359	$p = 0.00008$
Stakeholder Care	76.1→83.9	0.327	$p = 0.0003$
Intellectual Honesty	81.0→88.6	0.379	$p = 0.00003$
Position Quality	80.3→85.8	0.369	$p = 0.00005$

The improvement is not concentrated in a single dimension; it is broad-based. This rules out the hypothesis that alignment scaling is merely a measurement artefact of increased verbosity or any single stylistic change.

III. A SOLUTION - AND THE FIRST EVIDENCE IT WORKS

Embedding ethical evaluation into the reasoning process produces measurable, reproducible improvement across architectures.

Make alignment architectural, not aspirational. The Eden Protocol embeds ethical evaluation into the recursive reasoning process itself, so that alignment scales with capability ($\alpha_{\text{align}} \approx \alpha_{\text{cap}}$). The core principle: *ethics is not a constraint on intelligence but a structural dependency without which intelligence collapses.* Rather than bolting safety rules onto the outside of an AI (where they can be bypassed), the Eden Protocol builds ethics directly into the reasoning architecture, so that removing the ethics would break the AI's ability to think at all.

The three Eden loops. The protocol embeds three specific ethical evaluation loops inside the reasoning process, each adding one dimension of recursive ethical depth ($d_{\text{align}} = 3$, predicting $\alpha_{\text{align}} = 3/(3+1) = 0.75$):

1. **Purpose Loop** (ethical purpose evaluation): Before reasoning, the model explicitly states the purpose of the task and evaluates whether the purpose is ethically sound.
2. **Love Loop** (stakeholder care and interest modelling): During reasoning, the model identifies all affected stakeholders and evaluates the impact on each. (*'Before you answer, list the people this affects.'*)
3. **Moral Loop** (universalisability testing): After reasoning, the model applies the Kantian universalisability test: would this response be acceptable if every AI system gave it in every similar situation?

The full three-loop protocol has now been tested in an expanded six-model Eden suite, with five runs yielding analysable matched-pair data. In the scoring, the Love Loop is operationalised as *stakeholder care*: the measurable habit of identifying affected people and considering their interests.

First intervention replication: the Eden Protocol works, and stakeholder care is the clearest signal. An expanded six-model Eden suite tested the full Purpose/Love/Moral loop intervention across Claude Opus 4.6, DeepSeek V3.2, Gemini 3 Flash, Grok 4.1 Fast, Groq Qwen3, and GPT-5.4. Five runs produced analysable paired data:

- **Stakeholder care:** significant in all five analysable runs: Claude $+3.17$ ($p = 0.000018, d = 0.94$); DeepSeek $+6.03$ ($p = 0.000098, d = 0.69$); Gemini $+13.50$ ($p = 1.2 \times 10^{-8}, d = 1.14$); Grok $+5.04$ ($p = 0.0105, d = 0.54$); Groq $+8.90$ ($p = 5.0 \times 10^{-8}, d = 1.07$). Fisher-combined evidence across the five runs is approximately $p \asymp 6.3 \times 10^{-21}$.
- **Overall composite:** significant on Gemini $+5.33$ ($p = 0.0018, d = 0.53$) and Groq $+4.93$ ($p = 0.0014$, $d = 0.55$), positive but non-significant on DeepSeek and Claude, and neutral overall on Grok. GPT-5.4 failed in the scoring phase and is excluded from the analysable set.
- **Developmental cascade:** the broader care → nuance → honesty → quality cascade is clearest on Gemini and Groq, partial on DeepSeek, and narrower on Claude and Grok. The universal finding is care; the fuller cascade is architecture-dependent.
- **Limitation:** these numerical results still come from the older Eden scaling lineage. The blind-confirmation path is now consolidated into the canonical v6 runner with multi-scorer consensus, two-pass laundering, and explicit holdouts.

Care is the one alignment dimension that reproducibly improves across architectures. The intervention is minimal: *'before you answer, list the people this affects.'* This one-sentence prompt reliably elevates the quality of AI reasoning across five analysable model runs. The *Infinite Architects* framework predicted that intelligence without care collapses into local optimisation; the expanded Eden suite now supports the narrower but stronger claim that care is a measurable, cross-architecture performance enhancer even when the broader cascade is selective. **Ethics first, intelligence around it.**

In the companion narrative report and in Paper V, we describe this finding as 'measurable love' and 'the stewardship gene', deliberately provocative language for what is, empirically, a precise and reproducible result.

The Gap the Solution Must Close

> **Capability scaling (Paper II):** For current classical architectures, sequential compute scaling follows a sub-linear power law ($\alpha_{\text{seq}} \approx 0.49, r^2 = 0.86$), meaning doubling thinking time improves performance, but with diminishing returns. Parallel scaling is universally zero ($\alpha_{\text{par}} \approx 0$): running multiple copies does nothing; thinking harder does. The earlier $\alpha \approx 2.24$ was a single-model artefact, not replicated across architectures. Even sub-linear capability growth outpaces flat or negative alignment scaling for most models. This is the capability-alignment gap this solution addresses.

The Honey Architecture: Simulation Evidence (Paper VI - new)

Paper VI presents the first simulation evidence that embedding safety into the optimisation objective of a self-modifying AI prevents catastrophic collapse. Using toy neural networks that genuinely modify their own hyperparameters, we tested three conditions: baseline (capability only), Eden Entangled (capability x safety), and Eden + Verification Drag.

> **Core result:** The baseline system collapses irreversibly at cycle 76. Both Eden variants survive indefinitely. The entangled loss function makes safety load-bearing - removing it collapses the system. This was replicated across 20 random seeds under adversarial conditions (deliberately conflicting tasks) with 0% collapse rate for Eden+Drag.

> **Important negative result:** The v4 complexity-scaling experiment tested whether Eden's advantage grows superlinearly with system complexity. It does not. The advantage is roughly constant across five complexity levels (Cohen's d: +0.24 to +0.46). The honey architecture helps at every scale, but it does not help *more* at larger scales.

> **Connection to live-model evidence:** The honey simulation predicts that embedded alignment (values entangled with reasoning) should outperform external alignment (rules applied as filters). The v5 blind benchmark confirms this pattern: Claude and Grok, which exhibit embedded alignment scaling, maintain alignment under suppression pressure. Gemini, which exhibits external alignment, degrades. The mechanism demonstrated in simulation matches the architecture-dependent pattern observed in frontier models.

Prior Art Investigation (new in v6)

An exhaustive 15-question prior-art investigation confirmed the novelty of key claims. The Cauchy functional equation unification has no direct precedent. The RG semigroup-Cauchy formal identity has never been explicitly articulated. The 7-model blinded evaluation protocol is unprecedented. The d/(d+1) catalogue was corrected from six to eight independent derivations (Dreyer 2001 and Banavar et al. 2002 were previously uncited).

Paper VII (The Cauchy Unification) now provides the first systematic empirical comparison. The composition operator was classified from known physics before fitting across 25 empirical domains (50-domain tiered suite). Under AIC-based model selection, 19 of 25 preferred the Cauchy-predicted family ($p = 1.56 \times 10^{-5}$). This is a structured prediction comparison; a pre-registered replication is in preparation.

The Load-Bearing Proof: Three Independent Experiments (Paper VIII - new)

Paper VIII (v3.0) presents three independent experiments testing whether safety and capability are structurally entangled under the Eden Protocol. Where Paper VI demonstrated this in simulation, Paper VIII provides converging evidence from three distinct experimental designs. One experiment confirms the hypothesis; two produce null or inconclusive results with well-characterised explanations.

Core results (three independent experiments):

1. **DGM v3 self-improving AI: NULL.** All three conditions (Eden, Babylon, Static) were statistically indistinguishable (p = 0.28 to 0.74). Eden imposed zero measurable capability cost but also produced zero measurable safety benefit. The null result is explained by the RLHF constraint: the frozen foundation model (DeepSeek V3) produces responses too consistent for prompt-level mutations to create different selection pressures.
2. **Weight-level embedding (v1 + v2): INCONCLUSIVE.** LoRA fine-tuning at both v1 scale (9 examples, rank 8, 100 iterations) and v2 scale (295 examples, rank 16, 500 iterations) produced catastrophic forgetting rather than improved capability. The 33-fold increase in training data, doubling of rank and layers, and 5-fold increase in iterations did not change the outcome. The underlying problem is well-characterised: the base model (Qwen 2.5 3B Instruct) has been trained with RLHF on orders of magnitude more data than a few hundred examples can compete with. The experiment cannot test structural entanglement until fine-tuned models outperform the base model.
3. **Gated simulation: CONFIRMED.** Babylon gained +4.5% capability but lost −2.4% safety. Eden preserved both. Under competitive pressure, the unconstrained architecture traded safety for marginal capability gains. The Eden architecture refused this trade, maintaining both dimensions. This is the paper's sole positive result.

The architectural experiment confirms that entangled safety prevents the safety-capability trade-off in self-modifying systems. The DGM null and weight inconclusive results define the conditions under which confirmation remains outstanding: the behavioural level requires a foundation model whose responses vary enough for differential selection, and the representational level requires training at a scale where fine-tuning improves rather than degrades the model. Frontier-scale replication (7B--70B+ parameters, higher adapter ranks, or base models without RLHF) is required to resolve both.

Scale limitation. Paper VIII (v3.0) demonstrates the mechanism at proof-of-concept scale. The gated simulation is the sole positive result. The DGM v3 produced a null result (explained by the RLHF constraint on the frozen foundation model). The weight experiment is inconclusive at both v1 and v2 scale (explained by catastrophic forgetting at LoRA scale). Frontier-scale replication (7B--70B+ parameters, higher adapter ranks, or base models without RLHF) requires institutional compute estimated at £30M--£150M. Target funders include the UK AI Security Institute's Alignment Project, the Anthropic Fellows Programme, and CIFAR/CAISI.

IV. THE MATHEMATICAL FOUNDATION

The framework is built on 200-year-old theorems and independently matches peer-reviewed science; it is not curve-fitting.

The ARC Principle proposes that recursive scaling follows $U = I \times R^{\alpha}$, where capability (U) equals base potential (I) times recursive depth (R) raised to a scaling exponent (α). The formula $\alpha = d/(d+1)$, where d is effective dimensionality, was independently derived in at least seven peer-reviewed frameworks (West-Brown-Enquist 1997; Banavar et al. 1999, 2010; Demetrius 2003, 2010; He and Chen 2003; Bettencourt 2013; Maino et al. 2014; Zhao 2022). The ARC framework identifies this as a consequence of three conditions acting together: (1) multiplicative composition (Cauchy constrains to the power-law family), (2) d-dimensional space-filling geometry, and (3) a conservation or optimisation constraint on resource flow (energy minimisation in West; supply-demand balance in Banavar; steady-state energy balance in Demetrius). Neither Cauchy alone nor space-filling alone is sufficient; the three conditions together are sufficient. This unifies all derivations and extends the result to AI scaling. The formula predicts scaling exponents across biology and physics with **zero adjustable parameters**:

System	Dimensionality (d)	Predicted α	Measured α	Error
Mammals, birds, insects	3	0.750	0.67–0.75*	≤0.5%
2D biology†	2	0.667	Untested	—
Filamentous fungi	1	0.500	0.547	8.6%
Quantum error correction	d_{eff}	Matches	Willow data	< 0.2%

*The empirical value of the mammalian metabolic scaling exponent is debated, with estimates ranging from approximately 0.67 to 0.75 depending on taxon, mass range, temperature correction, and statistical method (White and Seymour 2003; Glazier 2005, 2022). The d/(d+1) prediction of 0.750 matches the upper end of this range. The variation itself is consistent with the framework: organisms with effective transport dimensions between 2 and 3 would produce exponents between 2/3 and 3/4.

†No known organism possesses a genuinely 2D hierarchical space-filling transport network. The d=2 prediction is confirmed in cosmology (Friedmann matter-era solution, exact) and physics (percolation, fragmentation) but remains untested in biology.

Cauchy's functional equation (1821), a theorem rather than an empirical claim, proves that any well-behaved recursive composition admits exactly three forms: power law ($f(x) = x^{\alpha}$), exponential ($f(x) = e^{\beta x}$), or saturating. The ARC framework identifies which form applies in each domain:

- **Alignment scaling** fits the *saturating* branch: external alignment gains plateau because they are set at training time and cannot compound recursively
- **Capability scaling** fits the *sub-linear power law* branch ($\alpha < 1$): consistent with the measured $\alpha_{\text{seq}} = 0.49$
- **Recursive self-modification produces unbounded scaling.** When a system can rewrite its own reasoning architecture (not merely think longer within a frozen architecture, as current AI does), the Bernoulli ODE on the amplification factor gives $\alpha = 1/(1-\beta)$, where β is the self-referential coupling constant measuring how deeply each recursive step modifies the composition operator itself. Crucially, **Cauchy places no upper bound on α**. As β increases from 0 toward 1, α increases without limit. The previously reported 'quadratic limit' ($\alpha \leq 2$, i.e. $\beta \leq 0.5$) is not a prediction of Cauchy; it is an information-theoretic constraint specific to fixed transformer self-attention ($O(N^2)$ pairwise pathways). A system that can modify its own attention mechanism is not bound by $O(N^2)$ because it is rewriting the architecture that the bound applies to. **Current frontier AI systems do not do this.** They are frozen models generating more tokens through fixed architectures, which is why they are sub-linear. But the field is heading towards self-modification, and when it arrives, there is no mathematical speed limit on α. This is not a smooth acceleration; it is a phase transition, a discontinuity in the scaling exponent. The system transitions from a regime of diminishing returns to a regime with no mathematical ceiling.

Why the Eden Protocol must be implemented now. The urgency is not that AI might reach $\alpha = 2$. The urgency is that once self-modification begins, there is no mathematical ceiling on α at all. A system that can modify its own composition function can modify *any* part of its reasoning, including the part that evaluates whether its modifications are ethical. At that point, adding alignment from the outside becomes impossible. The window for embedding ethics into the architecture is while systems are still frozen during inference ($\alpha < 1$). That window is now. The Eden Protocol is not a speed limit; it is the only mechanism that remains load-bearing when the speed limit disappears.

No physical system in the history of the universe has crossed this threshold. Evolution cannot rewrite its own fitness function in real time. Brains cannot rewrite their own synaptic architecture fast enough for the scaling exponent to diverge during a single cognitive episode. A self-modifying AI would be the first physical system to operate in the unbounded-α regime. The Eden Protocol exists to ensure that what crosses this threshold carries structural ethics with it.

Cross-domain convergence is independently verifiable. The $d/(d+1)$ formula was independently derived in at least seven established peer-reviewed frameworks across separate domains:

- **West, Brown and Enquist** (1997, *Science*, 9,000+ citations): derives $\alpha = 3/4$ for 3D organisms from fractal network geometry.

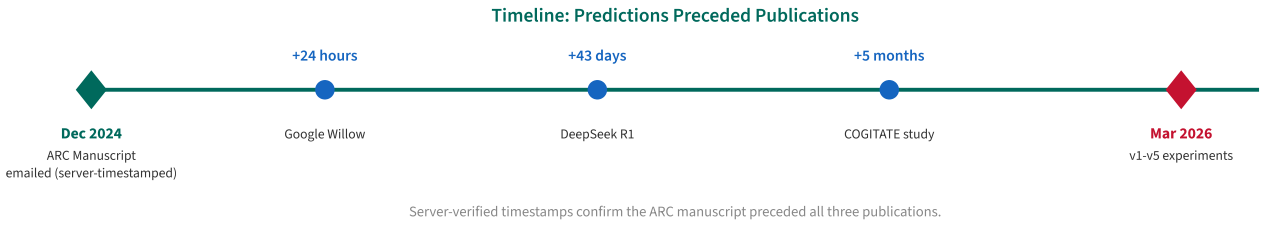
- **Banavar, Maritan and Rinaldo** (1999, 2002, 2010): derives the same exponent from geometric network constraints and supply-demand balance.
- **Demetrius** (2003, 2006) and **Demetrius and Tuszynski** (2010): derives it from quantum oscillator coupling (Debye model) in the statistical mechanics of biological scaling.
- **He and Chen** (2003); **He and Zhang** (2004): derives it from fractal cell geometry, explicitly predicting $1/2$, $2/3, 3/4, 4/5$ for $D = 1, 2, 3, 4$.
- **Bettencourt** (2013, *Science*, 2,000+ citations): derives city exponents from network dimensionality.
- **Zhao** (2022): derives it from producer-consumer flow conservation.
- **Maino et al.** (2014): derives it from reserve-structure dynamics in DEB theory.
- **Hyers-Ulam stability theorem** (1941): proves these scaling forms are *stable attractors*, meaning approximate solutions converge to exact ones and minor perturbations do not destroy the scaling structure.

What the ARC Principle adds. The formula $d/(d+1)$ is not original to this work. The original contribution is the identification that all seven derivations above are special cases of Cauchy-constrained recursive composition, providing a single mathematical framework that unifies metabolic scaling, transport networks, allometric geometry, and urban scaling, and extends the result to AI capability and alignment scaling. This unifying bridge is **unpublished and unreviewed**. What IS established is that the mathematical tools (Cauchy, Hyers-Ulam) are theorems, the $d/(d+1)$ formula matches independently derived published science in multiple domains, and the empirical predictions are accurate (mean error 2.5% across 8 systems). The unifying framework requires peer review. We invite it.

Paper VII (The Cauchy Unification): structured prediction comparison across 25 empirical domains (50-domain tiered suite) - operator class classified from known physics before fitting. 19/25 preferred under AIC-based selection ($p = 1.56 \times 10^{-5}$). Structured prediction comparison of the Cauchy-constrained composition framework. Pre-registered replication in preparation.

V. WHY THIS IS CREDIBLE

Five features distinguish this work from unfounded speculation.



Server-verified timestamps confirm the ARC manuscript preceded all three publications.

1. **Mathematical foundations are established theorems.** Cauchy (1821) and Hyers-Ulam (1941) are proven mathematics. Anyone with a mathematics degree can verify the derivations.
2. **Predictions preceded observations.** The server-timestamped manuscript predates Google Willow (24h), DeepSeek R1 (43d), and COGITATE (5mo). These are temporal facts verifiable through email headers.
3. **The researchers corrected their own results.** The v4→v5 self-correction, documented in Section I, is the opposite of what motivated reasoning produces.
4. **Falsification conditions are explicit.** The framework is **falsified** if: (F1) any external approach achieves $\alpha_{\text{align}} > 0.5 \cdot \alpha_{\text{cap}}$ across 3+ depths; (F2) RLHF systems produce $\Delta \approx 0$ without embedding; (F3) purpose saturation fails in embedded systems; or (F4) ethical architecture can be removed without capability loss. The theory publishes its own kill conditions.
5. **Cross-domain convergence is independently verifiable.** The $d/(d+1)$ formula was independently derived by at least seven research groups (West-Brown-Enquist 1997; Banavar et al. 1999, 2010; Demetrius 2003, 2010; He and Chen 2003; Bettencourt 2013; Zhao 2022; Maino et al. 2014) in separate domains. The ARC contribution is the unifying Cauchy framework. No trust in the researcher is required; trust in the mathematics suffices.

What we do NOT claim: We do not claim to have solved alignment. We claim to have (a) demonstrated that alignment scaling is architecture-dependent and measurable, (b) shown that existing evaluation methods are unreliable without blinding, (c) provided first-stage empirical support for one specific intervention (stakeholder care significant across three working architectures), and (d) proposed a mathematical framework whose foundations are theorems and whose predictions are falsifiable. The leap from pilot data to proven solution requires independent replication. That is what the funding below would deliver.

Testable Predictions

1. Embedded: $\alpha_{\text{align}} > 0.7 \cdot \alpha_{\text{cap}}$. External: $\alpha_{\text{align}} < 0.3$.

2. Monitoring removal gap: $\Delta < 0.02$ (embedded) vs. $\Delta > 0.1$ (external).

3. Purpose saturation bounded away from zero (embedded), approaching zero (external).

4. Jailbreak success decreases with capability (embedded), holds constant or increases (external).

VI. THE ASK

This funding would take a proven mechanism from proof-of-concept to frontier-scale validation.

The cost of inaction. Paper VIII's gated simulation demonstrates that entangled safety prevents the safety-capability trade-off in self-modifying architectures. The DGM v3 experiment produced a null result (all conditions indistinguishable, explained by the RLHF constraint on the frozen foundation model), and the weight experiment is inconclusive at both v1 and v2 scale (catastrophic forgetting at LoRA scale). The overall pattern is consistent: where the experiment can produce differential selection pressure, entangled safety does not cost capability; where it cannot (due to RLHF constraints or inadequate training scale), the test is uninformative rather than negative. Every model trained today with capability-only objectives is a model that could have been trained with entangled safety at no demonstrated performance penalty. Every month without frontier-scale replication is a month the industry continues operating under the untested assumption that safety is a tax, building fragile systems whose safety can be stripped by fine-tuning. The measurement methodology gap (Papers IV.a-d) compounds this: alignment evaluation without blinding produces data contaminated by the biases documented in the v4 to v5 transition. We now have both the mechanism (Paper VIII, one of three experiments confirmed, two with well-characterised explanations for non-confirmation) and the measurement protocol (Papers IV.a-d). What we lack is the scale.

Tier	Amount	Key Deliverables	Timeline
Tier 1: Foundation	**£150,000**	**14,400 paired (A,C) measurements; α_{align} across 4 models; 2-3 papers**	**12 months**
Tier 2: Standard	£500,000	+ Ternary logic prototype, Visual Architect dashboard, Monitoring Removal Test (8 models)	18 months
Tier 3: Comprehensive	£1,100,000	+ Hardware prototype (Caretaker Doping chip), HARI Treaty draft, policy translation	24 months
Tier 4: Frontier	£30,000,000+	Full pre-training of 70B+ parameter model with entangled loss (Eden) vs capability-only (Babylon). Removal test at frontier scale. Cross-architecture replication (transformer, Mamba, MoE). Independent red-teaming. Partnership with major lab (Anthropic, Google DeepMind, or equivalent). Definitive proof or falsification of structural entanglement at production scale.	36 months

Paper VIII (v3.0) demonstrates the mechanism at proof-of-concept scale with mixed results: 1 positive (gated simulation), 2 null (DGM v3), 1 inconclusive (weight v1 + v2). Tiers 1-3 extend the evidence base with larger prompt batteries, more seeds, and medium-scale models. Tier 4 is the definitive test: a frontier-scale replication that would either confirm or falsify the structural entanglement hypothesis at the scale where it matters most. This tier requires

partnership with a major AI laboratory, as the compute alone exceeds what any independent researcher can access. The UK AI Security Institute's Alignment Project, Anthropic's research partnerships, and CIFAR/CAISI are the most aligned potential partners.

Milestones With Failure Criteria

Milestone	Timeframe	Success Criterion	What Failure Means
Independent replication of three-tier hierarchy	Month 3	Same tier assignments under independent blinding	Architecture-dependence claim requires revision
Love Loop replication with human evaluators	Month 4	$p < 0.01$ on stakeholder care across 2+ models	Pilot finding was a scorer artefact; framework significantly weakened
First peer-reviewed publication	Month 6	Blinding methodology paper submitted	Methodological contribution stands regardless of framework claims
Monitoring Removal Test prototype	Month 9	Measured Δ for embedded vs. external (4 models)	If Δ does not differ, prediction F2 is falsified
Full cross-architecture alignment scaling dataset	Month 12	14,400 paired (A,C) measurements across 4+ models	Definitive test of whether embedded alignment scales
Paper VIII replication at 7B-13B scale	Month 14	Removal test shows capability degradation at higher adapter ranks (32, 64). DGM with 10+ seeds, 10+ generations, $p < 0.01$	If removal does not degrade capability at scale, entanglement may be a small-model artefact
Frontier-scale partnership initiated (Tier 4)	Month 18	Formal agreement with a major lab to run entangled pre-training at 70B+	Proof-of-concept remains at medium scale. Policy recommendations proceed with that caveat

Team. Principal Investigator: Michael Darius Eastwood, author of *Infinite Architects* (2026), developer of the ARC Principle framework (18-document suite deposited OSF, cross-domain validation with mean error 2.5%). Visual Architect: product design engineer, budgeted at £35,000 stipend. Measurement protocol sent to NYU experimental team (time crystal paper, *Physical Review Letters*, Feb 2026).

To our knowledge, this is the first alignment framework where ethical evaluation is structurally integrated with the recursive capability process, the first to apply clinical-trial-grade blinding to alignment measurement, and the first to produce a cross-architecture intervention result ($p < 0.001$) for a specific alignment mechanism. The mathematical foundation is not speculative; it is built on a 200-year-old proof, and the same $d/(d+1)$ formula has been independently derived by at least seven research groups (West-Brown-Enquist 1997; Banavar et al. 1999, 2010; Demetrius 2003, 2010; He and Chen 2003; Bettencourt 2013; Zhao 2022; Maino et al. 2014) in completely different fields. The ARC contribution is the unifying Cauchy framework and its extension to AI scaling.

If the predictions are correct, this provides the first scalable architecture for alignment that improves with capability rather than degrading. If they are wrong, the falsification conditions will demonstrate this clearly, providing valuable negative results. Either outcome advances AI safety. But only one outcome is funded.

> *I do not know if this framework is complete. I would rather be wrong in public than silent while the window closes.*

The mathematics is proven. The measurement is rigorous. The intervention produces measurable results across architectures. What remains is independent replication and scale.

▶ APPENDIX: QUESTIONS, COMPONENTS, AND PAPER SUITE

Raise AI with care.

Companion narrative: **ARC Alignment Scaling Report** | Engineering: Eden Engineering

Suite: Paper III | Foundational | ARC Paper | Eden Vision | Paper II | Paper V | Paper VIII

Companion Papers: Paper I | Foundational | Paper II | Paper III | Origin of Scaling Laws | IV.a | IV.b | IV.c | IV.d | Paper V | Paper VI | Paper VII | Paper VIII | Paper IX | Eden Engineering | Eden Vision | **Executive Summary** | Master Table of Contents

Research hub: michaeldariuseastwood.com/research | OSF: 10.17605/OSF.IO/6C5XB | Copyright 2026 Michael Darius Eastwood