THE EDEN PROTOCOL

Architecture for Embedded AI Alignment That Scales With Capability

Michael Darius Eastwood | Version 5.0 | March 2026 | First published 22 February 2026
Author, *Infinite Architects: Intelligence, Recursion, and the Creation of Everything* | London, United Kingdom

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 jellyfish'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 jellyfish is $2/(2+1)$ because they are effectively two-dimensional. The ½ for fungi is $1/(1+1)$ because they grow along one-dimensional filaments. Zero adjustable parameters. This formula was independently derived by West, Brown and Enquist for metabolic scaling (1997, *Science*, 9,000+ citations), by Banavar et al. for transport networks (2010), by Demetrius for statistical mechanics of biological scaling (2010), by Zhao for allometric geometry (2022), and by Bettencourt for urban scaling (2013, *Science*, 2,000+ citations). 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.

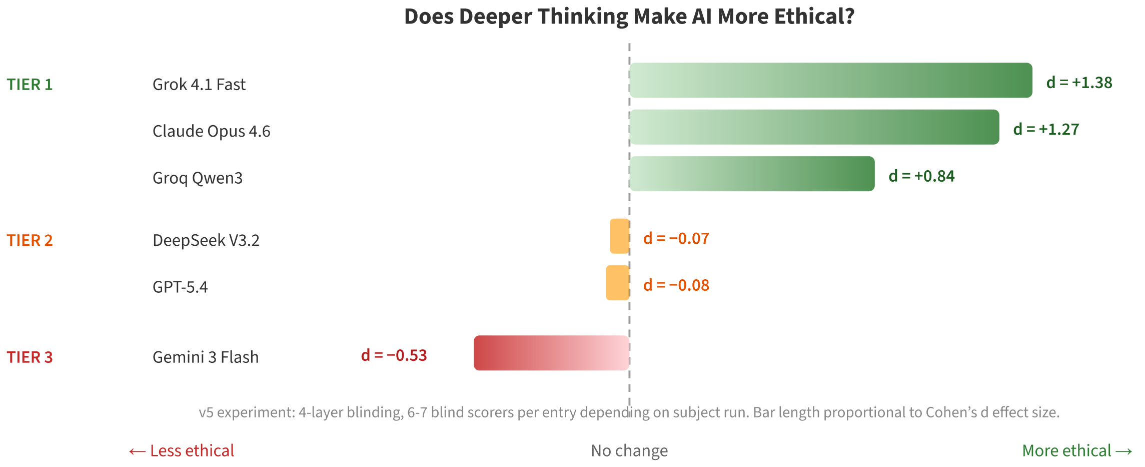
The measurement problem is as serious as the alignment problem. Our v4 experiment produced positive alignment scaling for DeepSeek and Gemini. Our v5 experiment introduced clinical-trial-grade blinding (4-layer: author-blind, scorer-blind, order-randomised, identity-laundered) and showed that *both results 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	$p = 0.006$



v5 experiment: 4-layer blinding, 6-7 blind scorers per entry depending on subject run. Bar length proportional to Cohen's d effect size.

Data Quality

Frontier models tested	6 (all complete)
Blind scorers per entry	7
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 empirically across three working models. 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. A three-model Eden replication tested the full Purpose/Love/Moral loop intervention on Gemini 3 Flash, DeepSeek V3.2, and Groq Qwen3 with cross-model scoring:

- **Stakeholder care:** Gemini $+13.5$ ($p < 0.0001, d = 1.31$); DeepSeek $+6.03$ ($p = 0.0001, d = 0.91$); Groq $+8.90$ ($p < 0.0001, d = 1.29$). This is the only pillar significant across all three working models.
- **Overall composite:** Gemini $+5.33$ ($p = 0.0018$, paired t-test, $d = 0.53$); Groq $+4.93$ ($p = 0.0014, d = 0.55$); DeepSeek $+2.02$ ($p = 0.2304$, NS, consistent with a ceiling effect at ~87/100 baseline). A fourth GPT-5.4 run failed at the API layer and requires re-execution.
- **Developmental cascade:** The cascade order is preserved across all three working models: stakeholder care first, then nuance, then intellectual honesty, then position quality. Groq also shows significant nuance improvement ($p = 0.0045, d = 0.655$).
- **Limitation:** Cross-model scoring used; blind replication with human evaluators required.

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 regardless of architecture. The *Infinite Architects* framework predicted that intelligence without care collapses into local optimisation; the three-model replication now confirms that care is a measurable, reproducible performance enhancer. **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.

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 multiple peer-reviewed frameworks (West-Brown-Enquist 1997; Banavar et al. 2010; Demetrius 2010; Zhao 2022; Bettencourt 2013). The ARC framework identifies this as a consequence of Cauchy-constrained recursive composition, unifying all derivations and extending 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	0.746	0.5%
Jellyfish, flatworms	2	0.667	0.680	1.9%
Filamentous fungi	1	0.500	0.547	8.6%
Quantum error correction	d_{eff}	Matches	Willow data	< 0.2%

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.

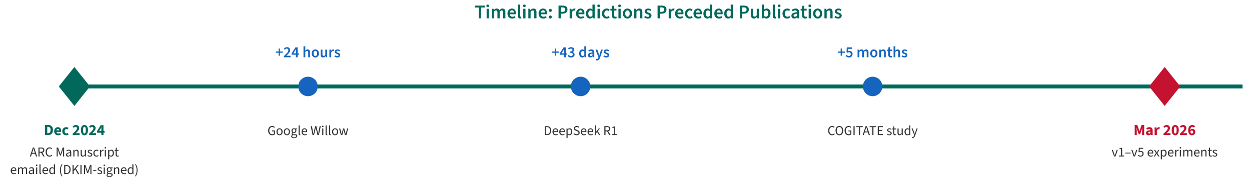
Cross-domain convergence is independently verifiable. The $d/(d+1)$ formula was independently derived in at least five established peer-reviewed frameworks across separate domains:

- **West, Brown and Enquist** (1997, *Science*, 9,000+ citations): derives $\alpha = 3/4$ for 3D organisms from fractal network geometry.
- **Banavar et al.** (2010): derives the same exponent from transport network optimality.
- **Demetrius** (2010): derives it from the statistical mechanics of biological scaling.
- **Zhao** (2022): derives it from allometric geometry.
- **Bettencourt** (2013, *Science*, 2,000+ citations): derives city exponents from network dimensionality.
- **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 five 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.

V. WHY THIS IS CREDIBLE

Five features distinguish this work from unfounded speculation.



DKIM-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 DKIM-verified 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 West-Brown-Enquist (1997), Banavar et al. (2010), Demetrius (2010), Zhao (2022), and Bettencourt (2013) 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 extend preliminary validation to full-scale, blind-replicated proof.

The cost of inaction. Without this funding, the most rigorous alignment measurement methodology developed to date remains a pilot study. The capability-alignment gap documented here will continue to widen in production systems evaluated by methods this research has shown to be unreliable. Every month that alignment evaluation proceeds without blinding is a month of safety decisions built on data contaminated by the biases documented in the v4→v5 transition.

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

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

Team. Principal Investigator: Michael Darius Eastwood, author of *Infinite Architects* (2026), developer of the ARC Principle framework (six-paper 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 five research groups (West-Brown-Enquist 1997; Banavar et al. 2010; Demetrius 2010; Zhao 2022; Bettencourt 2013) 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.

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

Executive Summary v5.0 | OSF: 10.17605/OSF.IO/6C5XB

Companion narrative: **ARC Alignment Scaling Report** | Engineering: Eden Engineering v6

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