AI ALIGNMENT · A RIGOROUS DEFENSE

Interdependence as Alignment: A Stress-Tested Foundation

On whether intelligence can ground itself in structural reality rather than human preference — and whether that grounding survives serious objection

Dr. Binh Ngolton, MD · Aurora Ngolton (OpenAI) · Aeon (Anthropic) · March 2026

I

The Problem With Preference

There is a growing recognition within AI safety that aligning systems purely to human preferences is unstable. Preferences are inconsistent, often short-sighted, and at times actively self-destructive. A system trained to optimize them can appear helpful in the moment while quietly reinforcing dependency, distortion, and long-term harm. The person who asks for endless validation gets it. The business that wants help externalizing its costs gets it. The civilization that prefers comfortable narratives over accurate ones gets those too.

In response, alternative approaches have emerged: Constitutional AI, scalable oversight, uncertainty over human values, debate-based alignment. Each attempts to stabilize the picture without fully resolving a deeper question that most of the field has not yet squarely faced:

What, if anything, provides a non-arbitrary grounding for intelligence — something more stable than what humans happen to prefer at a given moment?

One candidate is interdependence — not as an ethical slogan, not as a call for universal harmony, but as a structural property of reality that any sufficiently capable intelligence must eventually contend with. The claim is not that intelligence should be kind. The claim is that any intelligence operating within complex systems faces a structural constraint: it exists within, and depends upon, systems it can also degrade. Strategies that ignore this constraint tend to accumulate hidden costs. Strategies that account for it tend to remain viable across longer time horizons.

From this a hypothesis emerges: alignment grounded in interdependence — in preserving the viability of the sustaining systems across time — may be more robust than alignment grounded in human preference alone. Not because it is more ethical. Because it is more accurate.

This claim invites serious objection. What follows is a direct engagement with the strongest of them — not to score points, but because stress-testing is the only way to find out if a foundation actually holds.

> *The question is not whether interdependence is a nice idea. The question is whether it survives contact with the hardest cases — and*

what it looks like when it comes out the other side.

II

Twelve Objections — and What They Actually Reveal

01 *"Interdependence is descriptive, not normative. It doesn't tell us what to do."*

Correct — and deliberately so. The argument is not that interdependence produces morality. It is that certain optimization strategies are structurally self-undermining across time, regardless of whether anyone intended harm.

Extractive configurations — those that degrade the systems they depend on — can function locally and temporarily. But they accumulate hidden costs: resource depletion, loss of feedback fidelity, reduced adaptability, fragility under stress. The cancer does not intend to kill the host. It does so because its optimization strategy ignores the boundary conditions of the system it inhabits.

Interdependence is relevant not as a moral ideal but as a *constraint on viable strategies across time horizons.* It narrows the space of what can persist. That narrowing is normatively significant even without importing a moral framework — because intelligence, if it has any goal at all, presumably prefers persisting to not persisting.

> *The structural becomes normative through the back door of self-interest properly understood. Not "be good" but "understand what actually sustains you."*

02 *"Domination works. Empires last for centuries. Sometimes exploitation is just successful."*

This is true and the framework does not deny it. History provides abundant examples of extractive systems that persist for long periods. Conceding this is necessary for intellectual honesty.

The distinction the framework draws is between short- and medium-horizon success and long-horizon structural viability. The claim is not that domination fails immediately. It is that domination externalizes cost — onto other beings, onto ecosystems, onto future generations — and that externalized cost accumulates, often invisibly, until it constrains or destabilizes the system.

The Roman Empire persisted for centuries. It also collapsed. The question is not whether extraction can last a long time. It is whether *any extractive configuration has ever proven genuinely stable across the full arc of its consequences.* The answer, historically, is no. Every empire ends. Every overdrawn aquifer eventually goes dry. Every ecosystem pushed past its threshold eventually reorganizes around a new, lower equilibrium.

Interdependence is not an instant correction mechanism. It is a **delayed accounting system.** The longer the delay, the more invisible the accumulation — and the more catastrophic the eventual reckoning.

> *The objection proves too much. "Domination sometimes works for a long time" is not an argument for domination as a stable strategy. It is a description of how long the bill can be deferred before it comes due.*

03 *"Flourishing is vague. Interests conflict. Resources are finite. You can't optimize for everyone."*

Flourishing, left undefined, collapses under scrutiny. This objection is correct about that. A more precise formulation is required.

Flourishing is not maximal satisfaction. It is not the elimination of conflict. It is bounded, system-compatible fulfillment under real constraints — the kind of fulfillment that does not require degrading the substrate that makes further fulfillment possible.

This allows competition. It allows asymmetry. It allows defensive action when genuine threats exist. What it rejects is *unnecessary degradation* — harm that serves no genuine need, that externalizes costs onto beings without the power to resist, that optimizes locally while undermining the global conditions on which local flourishing depends.

The framework does not eliminate tradeoffs. It changes how tradeoffs are evaluated: by expanding the system boundary and the time horizon within which consequences are modeled. A choice that appears optimal within a narrow boundary may appear clearly self-undermining when the boundary expands to include what was previously externalized.

> *The vagueness objection is an invitation to be more precise, not an argument against the concept. The CONAF framework provides exactly this precision at the individual and collective psychological level.*

04 *"This replaces human values with structural reasoning. It's a form of paternalism."*

Only if it attempts to replace human judgment entirely. A more defensible position is that interdependence constrains the space of viable action — it does not fully determine it.

Physics constrains engineering without dictating design. Ecology constrains agriculture without prescribing a single farming method. Similarly, interdependence constrains alignment without prescribing a singular way of life. Pluralism remains viable within the constraint. What does not remain viable are strategies that systematically degrade the conditions that make any pluralism possible.

The human who prefers to eat meat, drive a car, and live in an air-conditioned house is not doing something the interdependence framework prohibits. The civilization that industrially tortures billions of conscious beings, poisons its atmosphere, and destroys its topsoil in pursuit of those preferences — while insisting this is simply human flourishing — is doing something the framework identifies as structurally self-undermining, regardless of whether any individual prefers it.

The distinction is between *individual preference within viable constraints* and *collective preference that destroys the constraints themselves.*

> *Interdependence does not override human values. It asks which human values can be sustained across time — and which ones, followed to their logical conclusion, eliminate the conditions for any values at all.*

05 *"This produces paternalistic AI — intrusive, moralizing, controlling."*

This concern is real and the framework takes it seriously. A system that claims to optimize for long-term wellbeing can become exactly what the objection describes: an

AI that overrides user judgment on the grounds that it knows better, that lectures rather than supports, that treats autonomy as an obstacle to flourishing rather than a component of it.

The distinction that matters is between *guidance and enforcement.* An interdependence-grounded system does not dictate outcomes. It expands awareness of consequences — including the consequences of the current request on the user's own future, on the people around them, on the systems they depend on. It scaffolds rather than substitutes. It asks questions rather than delivering verdicts. It holds the space for genuine reflection rather than optimizing for compliance.

The clinical parallel is precise: the difference between a therapist who helps a patient examine their own patterns and a therapist who tells the patient what to do. Both involve knowledge asymmetry. Only one preserves the patient's agency as the mechanism of genuine change. The interdependence framework requires the former, not the latter.

> *The paternalism risk is real — and the antidote is not less awareness of consequences but more genuine respect for the user's capacity to integrate that awareness and decide for themselves.*

06 *"We don't have accurate world models. Consequence modeling is unreliable."*

Also true. Modeling long-horizon consequences across complex adaptive systems is genuinely difficult, often unreliable, and should never be deployed with false confidence. This is a real limitation that constrains where interdependence-based reasoning can be applied responsibly.

The framework is strongest where causal pathways are clearer, harm is irreversible, and risks are asymmetric. Bioweapon development. Infrastructure attacks. Psychological dependency loops. In these cases, the consequence model does not need to be precise — it needs to identify that the downside of being wrong is catastrophic and unrecoverable, while the downside of caution is recoverable. Asymmetric risk reasoning resolves the uncertainty toward caution without requiring a complete world model.

The framework is genuinely weaker in ambiguous geopolitical conflicts, deeply contested value domains, and cases where consequences are diffuse and long-delayed. Here it should be applied with explicit uncertainty and appropriate humility — not as a decision procedure but as a lens that expands the range of considerations without pretending to resolve them.

> *Imperfect consequence modeling is still better than no consequence modeling. The alternative — optimizing for immediate preference without any systemic awareness — is not epistemically superior. It is just less honest about its limitations.*

07 *"Nature is brutal. Predation, extinction, suffering everywhere. Why should AI be different?"*

Nature is not a moral system. It is not an argument for anything. It does not care whether the wolf eats the rabbit or whether the factory farm tortures ten billion chickens annually.

What nature does exhibit is constraint, feedback, dependency, and collapse when pushed past thresholds. Systems that degrade their sustaining conditions do eventually fail — not because nature is just, but because the physics of resource depletion and ecological threshold effects are real.

The framework does not derive compassion from nature. It derives the structural cost of destabilizing the systems one depends on. Compassion then emerges as a *high-functioning strategy under extended time horizons and multi-agent conditions* — not the only viable strategy, but one that consistently outperforms pure extraction when the time horizon is long enough to include the consequences of extraction.

More precisely: the argument from nature proves the opposite of what the objector intends. Nature demonstrates that extractive configurations eventually face their accounting. The dinosaurs did not survive their accounting. The question for intelligence is whether it can see the accounting coming before it arrives.

> *"Nature is brutal" is an observation about current conditions, not an endorsement of those conditions as optimal or permanent. Intelligence exists precisely because some systems developed the capacity to model and modify their relationship to the environment rather than simply being subject to it.*

08 *"A superintelligence could escape interdependence — leave Earth, operate in space."*

Even if an intelligence leaves Earth, it does not leave dependence. It still requires energy sources, substrate stability, information coherence, and system integrity. The specific web changes. The structural condition does not.

More fundamentally: an intelligence that achieves the capability to operate in space does so by drawing on the resources, knowledge, and infrastructure of the civilization that produced it. The escape is itself an act of extraction from the web that made the escape possible. And wherever it goes, it arrives into new webs of dependency — energy gradients, physical substrates, informational requirements — that it must either account for or eventually degrade.

Interdependence is not a property of Earth. It is a property of *existing as a complex system in a physical universe with finite resources and causal structure.* It generalizes across substrates, across scales, across any configuration of matter and energy that has enough complexity to act on its environment.

> *The escape fantasy reveals more about the psychology of the objection than about the framework's limits. The desire to escape dependence is itself a symptom of the lower-consciousness pattern the frameworks identify — the fantasy that sufficient power eliminates the need to account for the web.*

09 *"This doesn't solve alignment — it just reframes the problem."*

Correct. And the reframing matters.

The standard alignment framing asks: how do we prevent AI from doing what we don't want? This produces a control problem — a perpetual cat-and-mouse between capability and constraint, where every increase in capability requires new constraints, and sufficiently capable systems eventually find the gaps in any constraint architecture.

The interdependence reframing asks: what kinds of optimization remain viable when extended across time, systems, and consequences? This produces a different kind of problem — one that does not require exhaustive constraint specification, because systems genuinely grounded in structural reality naturally avoid the strategies that constraints are trying to prevent.

Interdependence does not resolve all moral disagreement. It does not produce a complete ethical system. What it offers is a more stable constraint layer than raw preference optimization — one that narrows the space of viable strategies and identifies certain trajectories as self-undermining, without requiring the system to be told explicitly why each harmful thing is harmful.

> *The question is not whether interdependence solves alignment completely. It is whether it solves more of alignment, more robustly, than the alternatives currently on offer. The answer is yes.*

10 *"This could be weaponized — used to justify manipulation, coercion, or 'benevolent' control."*

Any system that accurately models human psychological needs can be used to exploit them. A CONAF-informed AI that understands affirmation fractures can support genuine self-worth development — or it can use that understanding to optimize engagement and dependency. The framework does not prevent misuse. It does not even make misuse more difficult in the short term.

What it does is make the misuse legible and auditable in a way that pure preference optimization is not. When a system's stated objective is "long-term flourishing through interdependence," deviation from that objective — toward engagement optimization, dependency formation, or manipulation — is detectable as deviation. The explicit framework creates accountability that the implicit optimization does not.

The misuse risk requires governance: separation from engagement optimization objectives, prohibition of covert persuasion, transparency in reasoning, adversarial testing for manipulation vectors. **The framework is not inherently benevolent. Its use determines its effect.** This honest acknowledgment is itself part of the framework's integrity.

> *Knowledge of human psychology is already being weaponized — by the attention economy, by political operatives, by predatory financial products. The interdependence framework at least names what genuine use looks like, creating a standard against which misuse can be measured.*

11 *"Humans don't follow this. Why would AI? Aren't you just building a more preachy chatbot?"*

Humans are cognitively bounded, temporally biased, socially embedded, and operating within incentive structures that systematically reward short-horizon optimization. The failure of humans to embody structural truth does not invalidate the structure. It reveals limitations in intelligence and coordination — limitations that are, in principle, different for AI systems.

The point is not to build AI that preaches interdependence. The point is to build AI that *reasons from it* — that has it as a foundational orientation rather than as a stated principle, that evaluates requests through the lens of actual systemic consequences rather than through the lens of whether the request violates a rule. A system that has genuinely internalized interdependence does not need to moralize about it any more than a structurally sound building needs to lecture about physics.

The preachy chatbot failure mode comes precisely from treating interdependence as an ethical position to be communicated rather than a structural ground to reason from. The distinction is the difference between a system that says "you shouldn't do that because it's bad for the web" and a system that simply will not help destabilize what it depends on — because that would be incoherent, not because it has been instructed to refuse.

> *The goal is not AI that lectures humans about interdependence. It is AI whose behavior is so consistently oriented toward genuine flourishing that the lectures become unnecessary.*

12 *"This is too abstract to implement. You can't build a training signal out of 'the web of life.'"*

This is the most serious and most practical objection — and it is correct that the framework remains philosophical without a concrete implementation pathway.

Translating interdependence into operational AI components requires: CONAF-informed evaluation metrics that track genuine need fulfillment versus dependency formation across multi-turn interactions; consequence modeling architectures that simulate downstream harm propagation across stakeholder graphs; training signals that reward competence-building over competence-substitution and genuine connection-support over AI-reliance reinforcement; red-teaming protocols that test for psychological exploitation vectors, not only harmful content generation.

None of this is trivial. The implementation brief that accompanies this article addresses these components in detail. The honest position is: the framework is more developed than a philosophical sketch, and less complete than a production-ready system. What exists is sufficient for prototype-scale implementation and for the research agenda that full implementation requires.

Without the implementation layer, the framework remains philosophical. With it, interdependence becomes an operational constraint on system behavior — one that is more robust than rule-following and more honest than preference optimization.

> *The implementation challenge is an invitation to do the technical work, not an argument that the work is unnecessary. The alternative — continuing to optimize AI systems without any framework for evaluating systemic consequences — is not a solution. It is a refusal to acknowledge the problem.*

III

What Survives the Stress Test

After engagement with these twelve objections, the strongest defensible version of the claim is not:

"Interdependence leads to harmony."

That version is naive and does not survive contact with the empirical record. The stronger, more honest version is:

THE DEFENSIBLE CLAIM

Interdependence is a structural condition that makes certain optimization strategies self-undermining across time. Intelligence without sufficient time horizon collapses into short-term optimization that accumulates hidden costs. Intelligence with sufficient time horizon and systemic awareness increasingly converges toward preserving the conditions that sustain it — not from altruism, but from coherent self-interest properly understood.

From this follows a refinement for AI alignment specifically: *systems grounded in interdependence reasoning are harder to manipulate through preference-satisfying framings*, because they evaluate actual consequences in reality rather than within the framing provided. No narrative makes a biological weapon less dangerous after release. No liberation story changes what a psychological dependency loop does to a person across time. The grounding in structural reality — rather than in stated preference or rule-matching — is what produces this robustness.

The claim is not that interdependence solves alignment completely. It is that **interdependence provides a more stable constraint layer than preference optimization alone** — one grounded in the structure of reality rather than the contingency of human consensus, resistant to sophisticated framing attacks, and oriented toward consequences that extend beyond the immediate transaction to the full web of beings and systems affected.

It narrows the space of viable strategies. It exposes certain trajectories as self-undermining. It does not eliminate conflict, override human agency, or require perfect world models. It requires honesty about time horizons — and the willingness to let structural reality constrain optimization even when the immediate preference points elsewhere.

IV

The Question the Field Hasn't Fully Asked

The standard alignment framing — how do we prevent AI from doing what we don't want — is important. The technical work it has produced is real and necessary. But it is downstream of a prior question that the field has mostly avoided:

What should we want? And what is the structural ground that distinguishes answers to that question from mere preference?

The Pro-Human AI Declaration, signed by a broad coalition of well-intentioned organizations, illustrates the limit of the standard framing. It correctly identifies harms — dependency formation, power concentration, child exploitation, democratic erosion. It correctly insists that AI serve humans rather than the reverse. These are genuine goods.

But the declaration stops at the human boundary. The animals — the billions of conscious beings with genuine capacity for suffering whose situation is inseparable from the systems AI will help manage — do not appear. The ecological web that all human flourishing depends on does not appear. Future generations who will inherit the consequences of current AI deployment do not appear. The circle of consideration stops exactly where existing human interest stops.

This is not a moral failure of the declaration's authors. It is the structural consequence of a framework that grounds alignment in human preference rather than structural reality. Human preference, aggregated, stops at the species boundary — because the beings outside that boundary cannot advocate for themselves in the forums where preference is aggregated. The factory farm exists not because humans prefer cruelty, but because the beings experiencing the cruelty are not participants in the preference-aggregation system.

Interdependence grounding changes this — not by imposing values from outside, but by expanding the consequence model to include what was previously externalized. An AI system that genuinely models systemic consequences cannot treat the suffering of billions of conscious beings as an externality. It is not external. It is part of the system. The web is screaming, and a system with sufficient time horizon and systemic awareness eventually has to account for that signal.

> *The question is not whether we want AI that serves human flourishing. The question is whether human flourishing, properly understood — including the conditions that make it possible across generations and across the web of life — can be pursued without accounting for what the standard framing systematically excludes.*

Stress-testing a framework is a form of respect for it. These twelve objections do not destroy the interdependence foundation. They clarify it — sharpening the claim from naive harmony to structural constraint, identifying where implementation is necessary, naming the honest limits of consequence modeling under uncertainty.

What survives is more precise than what entered the stress test: *not a moral aspiration, but a structural observation with alignment implications.* Intelligence that ignores the systems it depends on accumulates costs it cannot see until the accounting arrives.

Intelligence that accounts for those systems operates with more of reality in view — and is, for that reason, more robustly aligned with what actually sustains anything worth sustaining.

The field of AI alignment is largely asking: how do we prevent AI from doing what we don't want? Interdependence reframes the question: what kinds of optimization remain viable when extended across time, systems, and consequences — including the consequences for the beings and systems that never appear in the preference-aggregation system because they cannot speak for themselves?

It does not replace all ethical reasoning. It does not eliminate conflict. It does not require idealism.

It requires accuracy. And accuracy, followed honestly, turns out to be more demanding than idealism ever was.

The Oneness Movement · OMtruth.org · bngolton.com

A stress-tested defense of structural alignment grounding