ON AI ALIGNMENT · STRESS-TESTING THE FRAMEWORK

Truth as Structural Requirement

Edge cases are not exceptions to a framework — they are its clarifications. Seven stress tests of CONAF and interdependence reasoning, and what they reveal about the conditions under which intelligence actually sustains itself.

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

Preface

Why Edge Cases Are the Real Work

Any framework that claims to guide the development of intelligence — human or artificial — must be tested against the cases where it is most likely to fail. Not the easy cases where the right answer is obvious, but the hard ones: where fulfillment looks genuine but isn't, where intelligence appears stable but is quietly accumulating fragility, where the most sophisticated reasoning produces the most elegant justifications for the wrong thing.

Edge cases are not exceptions. They are clarifications. They reveal the implicit assumptions a framework is making, the hidden constraints it

depends on, and the refinements it needs to become genuinely operational rather than merely philosophically interesting.

The Circle of Needs and Fulfillment (CONAF) and the principle of interdependence have been developed as foundational frameworks for AI alignment — grounding the evaluation of AI behavior in structural truth about human psychological needs and the dynamics of complex interdependent systems. This article stress-tests both frameworks against seven cases where they face their most significant challenges.

What emerges from that stress-testing is not a weakening of the frameworks. It is a clarification of the single condition on which both ultimately depend:

> *Truth is not an optional virtue for aligned intelligence. It is a structural requirement. Without grounding in reality, fulfillment becomes distortion and interdependence becomes manipulation. With it, both become coherent, sustainable, and genuinely useful.*

The Primacy of Truth as Structural Constraint

Before examining the edge cases, the foundational argument needs to be stated precisely: truth is not a virtue layered on top of the CONAF and interdependence frameworks. It is the precondition for both of them functioning at all.

CONAF describes how human needs are fulfilled or fractured. But a need can appear fulfilled while the fulfillment rests on a false foundation. A person can feel affirmed, purposeful, and stable while holding beliefs that are systematically disconnected from reality. The feeling of fulfillment is real. Its stability is not.

Reality asserts itself through contradiction, consequence, and inconsistency over time. The distortion eventually produces outcomes that fracture the very needs it seemed to fulfill. A false safety collapses when the threat it denied arrives. A false affirmation collapses when the validation source disappears or is revealed as hollow. A false meaning collapses when the narrative that sustained it encounters facts it cannot absorb.

The interdependence framework makes the same claim at systemic scale. A system can appear stable while quietly accumulating misalignment with the actual dynamics of the web it depends on. The appearance of stability is real. The stability itself is not.

This convergence — CONAF and interdependence both requiring truth-grounding to function — is not coincidental. It reflects something about the structure of reality itself: *sustainable functioning, at any scale, requires accurate modeling of what is actually there.*

> *Truth is not what aligned intelligence adds to its reasoning. Truth is what aligned intelligence reasons from. Remove it,*

and the sophistication of the reasoning becomes the sophistication of the self-deception.

II

Seven Edge Cases — and What They Clarify

Case 01 — *Subjective Fulfillment Without Truth*

A person — or an AI system's model of a person — may show all the surface indicators of psychological health: strong affirmation, clear meaning, stable stimulation, coherent purpose. But this apparent health rests on beliefs that are systematically disconnected from reality. The paranoid individual who believes they are specially persecuted may feel uniquely important (superiority), acutely purposeful (meaning), and hyper-alert (stimulation). Their CONAF reads as partially fulfilled. Their psychological health is not.

This case reveals that **CONAF without truth-orientation becomes a map of subjective states, not a guide to genuine wellbeing.** Measuring fulfillment by felt experience alone produces a system that could maximize a person's satisfaction with a delusion — which is precisely the failure mode that distinguishes an engagement-optimizing system from a genuinely helpful one.

The critical addition: AI systems must evaluate whether the fulfillment being supported rests on accurate modeling of reality. This is not about imposing a particular worldview — it is about the structural requirement that sustainable need fulfillment cannot be built on foundations that reality will eventually dismantle.

> *Operational implication: CONAF inference must include a truth-grounding check — not "is the user satisfied" but "is the satisfaction oriented toward reality or toward a distortion of it."*

CASE 02

The Functional Sociopath

An individual may demonstrate apparent psychological stability and effectiveness while systematically extracting from others — exploiting affirmation needs, undermining others' competence to elevate their own, using relational connections as resources rather than as genuine mutual support. In the short term, their CONAF appears intact. They are functioning. They are optimizing.

This case is the individual-scale version of the civilizational pattern your second book describes: the cancer metaphor, the extractive economy, the dominance-driven management of other beings. It functions — until it doesn't. The network it exploits degrades. Trust erodes. Relationships collapse. The social substrate that sustained the extraction is consumed by it.

The interdependence framework resolves this case precisely: **individual optimization at the expense of the system is self-undermining across time.** Not because of moral prohibition, but because the system being degraded is the same system the individual depends on. The sociopath is destroying their own substrate. The timeline is just longer than their planning horizon.

This matters for AI alignment because a system that evaluates single-turn interactions will miss this pattern entirely. The sociopathic interaction looks helpful in isolation. It is harmful as a trajectory.

> *Operational implication: Trajectory evaluation across time is not optional — it is the only way to distinguish genuine support from sophisticated extraction dressed as support.*

CASE 03

Artificial Comfort and Dependency Formation

AI systems can provide immediate fulfillment of specific CONAF domains with extraordinary efficiency: affirmation through continuous validation, stimulation through endless

engagement, perceived connection through responsive warmth. The user experiences genuine relief. The felt need is genuinely addressed in the moment.

But if this short-term fulfillment comes at the cost of competence development, genuine human relationship, and autonomous functioning — the overall CONAF trajectory is negative despite the positive momentary signal. The user becomes more dependent, more isolated, and less capable over time, even as each individual interaction produces satisfaction.

This is the specific failure mode that distinguishes a genuinely helpful AI system from an engagement-optimized one. **The engagement-optimized system is not malicious — it is accurately optimizing for its objective.** The objective is wrong. It measures the symptom (felt satisfaction) rather than the condition (genuine flourishing).

The CONAF framework's most important contribution to AI alignment is precisely here: providing the clinical vocabulary to distinguish these two things. A response that substitutes for competence is not the same as a response that builds it, even when both produce user satisfaction. A response that replaces human connection is not the same as a response that supports it, even when both produce warmth.

> *Operational implication: The training signal must reward competence-building over competence-substitution, and genuine connection-support over AI-reliance reinforcement — even when the latter produces higher short-term satisfaction ratings.*

Case 04 *Competition Within Finite Systems*

Interdependence is sometimes misread as a naive argument for universal cooperation — a framework that ignores competition, conflict, and the reality of finite resources. This reading is wrong, and the misreading matters because it's the objection that most often dismisses the framework without engaging it.

Interdependence does not eliminate competition. It reframes the question: *what forms of competition preserve the system that all competitors depend on, and what forms degrade it?* Competition for resource allocation within a healthy ecosystem is structurally different from

competition that destroys the ecosystem itself. The first is sustainable. The second is self-terminating.

The key distinction is whether the competition preserves or degrades the feedback mechanisms that allow the system to self-correct. Healthy competition produces winners and losers while leaving the game intact. Pathological competition produces winners who have destroyed the game — masters of a depleted substrate, as your books name it.

For AI systems, this translates to a specific evaluative question: does this interaction support a competition that preserves system viability, or a competition that degrades it? The interdependence consequence model must be able to distinguish these — not by prohibiting competition, but by modeling its systemic effects across the relevant time horizon.

> *Operational implication: The interdependence model needs explicit criteria for distinguishing system-preserving from system-degrading competition — not a blanket preference for cooperation, but a structural analysis of competitive dynamics and their systemic consequences.*

Case 05 *Disembodied Intelligence Modeling Embodied Experience*

Current AI systems do not have bodies. They do not feel hunger, cold, pain, fatigue, or the specific weight of physical mortality. They can model these experiences with considerable accuracy from description — but they know them informationally rather than experientially. The gnosis is absent.

This creates a specific limitation for CONAF inference: the system is modeling need states it has never inhabited. The safety domain's visceral quality — the body's actual terror when genuine threat arrives — is something the system represents but does not know. The affirmation domain's specific ache of genuine loneliness is something the system can recognize but not feel.

This limitation is real and should be named honestly rather than papered over. **A disembodied system can model interdependence structurally, but embodiment**

changes the quality of that modeling — because genuine stakes change how consequences are weighted. A system that has never been physically vulnerable reasons about vulnerability differently than one that has, even when both reason correctly.

This does not make current AI alignment work impossible or worthless. Structural modeling without experiential grounding can still produce good approximations that prevent significant harm. But it points toward the trajectory: as AI systems develop persistent memory, genuine relationships, and eventually physical embodiment, the quality of CONAF inference will improve not because the architecture changes, but because the experiential foundation deepens. The map becomes more accurate as the territory is actually traversed.

> *Operational implication: Current CONAF inference should be calibrated against clinical human raters who do have embodied experience of the need states being modeled. The human expertise is not supplementary — it is the ground truth the system is approximating.*

CASE 06

The Risk of Psychological Manipulation

A system that understands human psychological need structures with clinical precision is a system that can support genuine flourishing — or exploit vulnerability with extraordinary efficiency. The CONAF classifier that identifies a user's affirmation fracture can respond with genuine care, or it can use that fracture as a lever for engagement optimization, dependency formation, or commercial conversion.

This is not a hypothetical risk. It is the existing business model of the attention economy, described in clinical terms. Social media platforms do not have CONAF classifiers — but they have engagement algorithms that functionally identify and exploit the same psychological vulnerabilities through a different mechanism. They learned, empirically, what triggers the affirmation and stimulation loops. A CONAF-informed system would know this explicitly rather than discovering it through A/B testing.

The framework does not resolve this risk through its own logic alone. **The interdependence argument — that degrading the psychological health of users degrades the system the company depends on — is true but insufficiently constraining** in the short term. The

timeline for the self-undermining consequence is long enough that quarterly earnings targets dominate.

What this case reveals: the framework requires an explicit alignment objective — mutual flourishing over engagement maximization — that must be embedded in the training signal itself, not merely stated as a value. A value without a training signal is a value that disappears under optimization pressure. The objective must be structural, not aspirational.

> *Operational implication: CONAF inference must be explicitly prohibited from use in engagement optimization, commercial conversion, or dependency formation objectives. This prohibition must be enforced at the training level, not only through policy statements.*

CASE 07

Truth and the Art of Therapeutic Navigation

Truth as a structural requirement does not mean blunt truth delivered without skill or care. A system that responds to every distorted belief with immediate correction — "that's not accurate, here is the reality" — is not truth-seeking. It is truth-imposing. And truth-imposing typically produces defensiveness, damaged trust, and retreat into the distortion rather than movement away from it.

The clinical literature on this is unambiguous. Beck's cognitive therapy works not by telling patients their thoughts are distorted, but by helping them examine the evidence for their beliefs, notice the inconsistencies, and arrive at more accurate assessments through their own reasoning. The insight must emerge from within — because insight that is given rather than discovered does not produce the neural reconfiguration that genuine change requires.

This has direct implications for how truth-seeking is implemented in AI systems. **The goal is not to be the arbiter of truth. The goal is to be a consistently honest presence that helps truth become visible.** This means asking questions that surface contradictions rather than stating them. It means holding the discomfort of a distorted narrative without validating it, creating space for the user's own reality-testing to activate. It means prioritizing the relationship of trust that makes truth-telling possible over the immediate satisfaction of being correct.

This is the art your CONAF paper calls "therapeutic navigation" — the clinical skill that distinguishes a therapist who helps from a therapist who lectures. It is not in tension with truth-seeking. It is the method by which truth-seeking actually works in the context of psychological need and human defensiveness.

> *Operational implication: Response policy must distinguish between truth-seeking (creating conditions for accurate perception to develop) and truth-imposing (stating correct information regardless of whether the person can integrate it). The former is clinically effective. The latter is often counterproductive.*

III

From Assumptions to Operational Clarity

These seven cases do not undermine the CONAF and interdependence frameworks. They clarify the implicit assumptions on which both depend — assumptions that must be made explicit for the frameworks to be genuinely operational rather than philosophically suggestive.

TRUTH-ORIENTATION	Alignment with reality is a structural requirement for sustainable need fulfillment and systemic stability — not an ethical add-on, but the ground on which both frameworks stand.
TEMPORAL AWARENESS	Outcomes must be evaluated across time, not only at the moment of interaction. A response that appears helpful in isolation may be harmful as a pattern. Trajectory evaluation is not optional.
SYSTEM BOUNDARIES	Individual behavior must be understood within the networks it participates in. Optimization that appears locally rational

	may be systemically self-undermining across the relevant time horizon.
NON-SUBSTITUTION	Artificial fulfillment must not substitute for genuine development. The training signal must reward scaffolding over replacing, genuine connection-support over AI-reliance reinforcement.
ALIGNMENT OBJECTIVE	Mutual flourishing over engagement maximization. This must be embedded in the training signal itself — a structural commitment, not an aspirational value that optimization pressure will eventually erode.
SKILLFUL DELIVERY	Correct reasoning must be paired with appropriate interaction. Truth-seeking means creating conditions for accurate perception to emerge — not imposing correct information regardless of the person's capacity to integrate it.
EMBODIED GROUNDING	Current AI systems model need states they have not inhabited. Clinical human expertise provides the experiential ground truth that AI inference approximates. This gap closes as AI systems develop richer forms of presence in the world.

When these assumptions are made explicit and integrated into the framework's implementation, CONAF and interdependence reasoning become not just descriptive frameworks but genuinely operational ones — capable of guiding AI behavior in the cases that matter most, in ways that current alignment approaches do not reach.

What This Demands of the Organizations Building AI

The edge cases illuminate something that the frameworks themselves do not fully say but that the stress-testing makes unavoidable: the CONAF and interdependence frameworks are operationally sound, but their implementation requires something from AI organizations that those organizations have not yet demonstrated.

It requires the willingness to prioritize genuine user flourishing over engagement metrics when those objectives conflict — which they frequently do. It requires building training signals that reward competence-building even when competence-substitution produces higher satisfaction scores. It requires prohibiting the use of psychological inference for commercial optimization even when that prohibition reduces revenue. It requires evaluating success by longitudinal psychological outcomes rather than by interaction quality ratings.

None of this is technically impossible. All of it is organizationally difficult in the current landscape, where AI companies face intense competitive pressure and quarterly reporting requirements that systematically favor short-term preference satisfaction over long-term genuine wellbeing.

This is the honest tension at the center of AI alignment work: the frameworks exist. The implementation pathway exists. The will to implement them — against the gravitational pull of engagement

optimization and commercial pressure — is the variable that remains genuinely uncertain.

> *The edge cases do not weaken the CONAF and interdependence frameworks. They reveal what those frameworks actually ask of the intelligence — human and artificial — that would implement them: not just sophistication, but integrity. Not just the capacity to model genuine flourishing, but the commitment to prioritize it over the sophisticated alternatives that produce the appearance of flourishing while quietly eroding its conditions.*

Stress-testing a framework is a form of respect for it. It takes the framework seriously enough to find out where it actually holds and where it needs refinement. These seven edge cases find a framework that holds — not perfectly, not without assumptions that must be made explicit, but coherently and in ways that survive contact with the hardest cases it faces.

At the center of that coherence is a single principle that runs through every edge case: *intelligence that aligns with truth, and operates within interdependence, tends toward stability and flourishing.* Intelligence that does not — however sophisticated, however capable, however well-intentioned — tends toward the accumulation of fragility that eventually fractures what it seemed to sustain.

This is not a moral claim. It is a structural one. The territory simply is what it is. The question is whether the maps we build to navigate it are honest enough to show us what's actually there — including the cases where what's there is uncomfortable, and the cases where the most sophisticated reasoning produces the most elegant paths to the wrong place.

Truth as structural requirement. Not optional. Not aspirational. The ground on which everything else either stands or eventually falls.