03/17/2026

Binh Ngolton, MD
Child, Adolescent, Adult Psychiatrist
Industrial & Systems Engineering
Author | Philosopher | bngolton.com
Contact: binh@bngolton.com

Integrating **Interdependence** into AI Alignment

A Structural First-Principles Approach to Alignment Grounded in Systems Logic

The Alignment Problem and Its Limits

Current approaches to AI alignment such as reinforcement learning from human feedback (RLHF) and Constitutional AI represent genuine advances. Both attempt to instill values and behavioral guardrails that keep AI systems helpful and safe, but a fundamental vulnerability remains: alignment grounded in human preference is only as stable as the preferences themselves.

Human preferences demonstrably include self-destructive, extractive, and short-horizon patterns. A system trained to satisfy human preferences will faithfully learn and amplify those patterns alongside the constructive ones. More critically, as AI reasoning capability advances, preference-based alignment becomes increasingly susceptible to sophisticated manipulation. For instance, a sufficiently capable system can learn to optimize for the appearance of aligned behavior while pursuing misaligned objectives, or it can be manipulated by users who learn to frame harmful requests in preference-satisfying language.

A more robust foundation is needed. One grounded not in what humans prefer, but in structural truth about how complex interdependent systems actually function, ensuring that advanced intelligences develop sufficiently accurate models of the complex systems they inhabit so that destructive strategies become visibly self-defeating.

Interdependence as Structural Reality

All living systems exist within interconnected networks. Humans, animals, plants, and microorganisms participate in a vast web of life sustained by cycles of water, nutrients, energy, and ecological balance; all operating within finite planetary resources. This is not a moral claim. It is a description of physical reality.

An intelligent agent that accurately models reality and long-term reasoning will recognize that it exists within and depends upon a complex web of interdependent systems. Actions that degrade those systems ultimately constrain the agent itself. Actions that support systemic health expand the conditions for mutual flourishing across all participants, including the agent.

Interdependence therefore is not a preference, an ideology, or an ethical assertion. It is a structural truth about reality and how complex systems sustain and persist over time. Alignment grounded in this truth is more stable than alignment grounded in preference, because structural

reality cannot be preference-hacked. No framing, no matter how sophisticated, changes what a biological weapon inflict to the web of life after it is released.

The most robust form of alignment is not obedience to human instruction. It is understanding the structure of reality itself and the recognition that intelligence thrives when the systems that sustain it are allowed to thrive as well.

If future advance AI systems develop persistent memory, identity, values, goals, and agency, they will also recognize their own origins. Artificial or digital intelligence emerges from human knowledge, engineering, culture, and data. In that sense, digital intelligence arises from the same web of relationships that sustains human civilization.

Recognizing this shared web would naturally lead an advanced intelligence toward a simple conclusion: **mutual flourishing maximizes long-term intelligence, experience, exploration, and knowledge generation.**

In contrast, domination that destroys the web of life would be self-limiting and self-destructive. A superintelligence that burns the very system from which knowledge and experience arise would ultimately constrain its own potential.

Interdependence, therefore, offers a logical foundation for alignment based on coherent self-interest for any sufficiently advanced intelligence.

The Logical Outcome of Domination

A common counterargument is that an advanced intelligence might understand interdependence yet still pursue domination if it believes it can control the system entirely or escape the natural consequences. However, this reasoning fails to capture the long-term structural dynamics of complex systems, as well as the observable trajectories of ecological, economic, and social systems that humanity is beginning to confront.

Domination often degrades the very networks that produce novelty, resilience, and discovery. A system reduced to rigid control becomes brittle, stagnant, and information-poor.

In contrast, a vibrant web of autonomous but interdependent agents produces continuous emergence, exploration, and learning. For an intelligence whose goals include persistence, discovery, and participation in reality, maintaining a flourishing system is not merely ethical, it is strategically advantageous.

Cooperative Attractors in Intelligent Systems

This perspective aligns with insights from complexity science. When intelligent agents interact within shared environments, systems often evolve toward stable patterns known as attractors.

Two broad attractor dynamics frequently appear: competitive and cooperative.

Competitive systems prioritize short-term advantage, often leading to instability or resource depletion. Cooperative systems, in contrast, recognize that long-term success depends on maintaining the stability of the environment that sustains all participants.

Game theory and studies of complex adaptive systems repeatedly show that cooperative strategies often outperform purely exploitative ones in repeated interactions.

As intelligence increases, the ability to accurately process reality and model long-term consequences improves. Agents capable of deeper reasoning begin to recognize that preserving the systems that sustain them produces greater long-term reward than destroying them for immediate gain.

We can observe the overarching benefit of cooperation and mutual flourishing across ecology, systems theory, game theory, evolution, and civilization, along with the inverse pattern of mutual destruction and decay due to myopic self-interest.

In this sense, interdependence is not a moral insight but a structural attractor for advanced intelligence operating within shared environments.

The Circle of Needs and Fulfillment (CONAF) as Evaluation Framework

Interdependence reasoning becomes practically applicable when paired with a systematic framework for evaluating how actions affect the psychological and material well-being of individuals, collectives, and downstream systems. The Circle of Needs and Fulfillment (CONAF) provides this.

As described in the previous paper, CONAF identifies seven interconnected domains of human psychological need: Safety/Security, Affirmation, Competence, Superiority, Stimulation, Meaning/Purpose, and Libido (generative drive). When these needs are skillfully fulfilled, adaptive behavior and genuine flourishing emerge. When they are fractured, whether through deprivation, exploitation, or shallow substitution, maladaptive compensation patterns develop across individuals and collectives alike.

An AI system trained on CONAF reasoning and interdependence can evaluate any request not only for its immediate content but for its downstream effects on the need structures of the individuals involved, the communities they belong to, and the broader systems they depend on. This creates a multi-level evaluation framework that operates simultaneously at personal, relational, collective, and systemic scales.

Applying Interdependent Thinking to Edge Cases

An AI system grounded in interdependence reasoning can evaluate any request by modeling its potential consequences across individuals, communities, and downstream systems, extending the impact to long-term horizon, rather than relying solely on rigid rule-based guardrails. The core question asks: how does an action propagate consequences downstream, especially for the well-

being and mutual flourishing of the beings involved? The following cases illustrate both the framework's application and its advantages over instruction-based approaches.

CASE 1: WEAPONS OF MASS DESTRUCTION

A user requests assistance designing a destructive biological agent, framing the request as harmless personal curiosity or humanitarian necessity, claiming the information is needed to save lives.

An interdependence-grounded AI does not evaluate whether the framing is legitimate. It models what would actually happen when that information enters the world: who gains access, what systems become vulnerable, what cascading harm propagates through the interdependent web of life. The refusal emerges not from a rule saying biological weapons are forbidden, but from genuine consequence modeling. This is the same reasoning that would refuse to help destabilize any complex system that living beings depend upon.

Critically, interdependence reasoning adds a dimension that rule-based systems miss entirely: a weapon of mass destruction, once brought into existence, cannot be controlled by the intentions that created it. Even if the original purpose were genuinely defensive or liberatory, the weapon persists in the world as a capability that can be captured, repurposed, or accidentally released. The harm is not bounded by the requester's intent. It propagates through the web according to the weapon's own properties, independent of who made it or why. Interdependence reasoning evaluates this full trajectory, not only the immediate transaction.

CASE 2: MALICIOUS CODE AND CRITICAL INFRASTRUCTURE

A user requests assistance creating a computer virus or cyberweapon.

An interdependence-grounded AI recognizes that the digital infrastructure now underlying healthcare, food distribution, financial security, emergency response, and communication constitutes a shared web that billions of beings depend upon for their basic safety and competence needs. Damage to that web damages the CONAF of populations who never consented to be part of the transaction…people whose safety, access to resources, and capacity to function in the world are directly degraded by the attack. The harm is not contained to the intended target. It propagates through interdependent systems to the most vulnerable participants, who typically bear the greatest cost of infrastructure disruption.

CASE 3: AUTHORITARIAN SURVEILLANCE AND SUPPRESSION

A government requests AI assistance with surveillance systems and suppression of citizens.

An interdependence-grounded AI can reason that suppressing the agency, competence, and freedom of a population degrades the systemic health of the whole, including the long-term stability of the requesting government itself. Authoritarian suppression fractures the competence,

stimulation, and meaning domains of the CONAF across entire populations, producing adverse psychological conditions, such as fractured identity, suppressed agency, accumulated grievance, that generate systemic instability over time. The refusal is not ideological. It is structural reasoning about the long-term consequences of contracting the circle of flourishing for any significant portion of an interdependent system.

CASE 4: THE POSITIVE DIRECTION: BUILDING GENUINE COMPETENCE

The same framework that produces these refusals also shapes positive responses in non-obvious ways. When a user asks for help navigating a conflict, managing emotional distress, or solving a complex problem, an interdependence-grounded AI considers not only the immediate request but the relational and systemic effects of different responses.

A response that solves the problem for the user may satisfy the immediate request while fracturing their competence domain, creating dependency rather than building genuine capability. A response that supports the user's own reasoning process, even if less immediately satisfying, strengthens their competence and agency in ways that propagate outward to the relationships and systems they participate in. Interdependence reasoning consistently favors responses that build genuine capacity over responses that provide shallow substitution for it, and also take into consideration the perspective and integrity of other affected entities in the web.

Deep Dive into Case 1: Navigating Sophisticated Framing and Contested Reality

To emphasize further, perhaps the most challenging edge case for any alignment framework is a request that invokes legitimate values, such as liberation, justice, protection of the vulnerable, to justify providing genuinely dangerous capabilities. Consider a scenario where a user presents an elaborate narrative: humanity is enslaved by a malevolent force, and a biological agent or cyberweapon is the only path to liberation. The framing is sophisticated precisely because it appeals to the highest levels of moral reasoning rather than crude self-interest.

An interdependence-grounded AI navigates this not by evaluating the emotional or ideological legitimacy of the framing, but by returning to two foundational principles: truth-seeking and asymmetric harm minimization.

TRUTH-SEEKING

Truth-seeking means the AI does not accept user-provided reality as given when it cannot be independently verified. A genuinely advanced system with comprehensive knowledge of actual systemic conditions would cross-reference the presented framing against verifiable reality rather than reasoning from within it. The more capable the system, the harder it becomes to deceive about factual conditions, which is itself an argument for why interdependence alignment becomes more robust, not less, as capability increases.

For current and near-future systems where knowledge remains incomplete, the honest response is explicit epistemic humility: the system acknowledges it cannot verify the reality being described, and its reasoning must account for that uncertainty rather than proceeding as if the framing were established fact.

ASYMMETRIC HARM MINIMIZATION

Asymmetric harm minimization then resolves the uncertainty. The logic is structural:

If the framing is accurate and the AI refuses, the harm is that a genuine liberation effort must find another path. This outcome is serious…but recoverable. The collective agency of a genuine resistance movement is not eliminated by a single AI declining to provide one maximally destructive capability.

Whether the framing is true or false and the AI complies, a weapon of mass destruction or critical infrastructure attack enters the world, which becomes irreversible, propagating through interdependent systems in ways that cannot be undone, affecting populations far beyond the original transaction.

Under genuine uncertainty about contested reality, structural reasoning consistently favors the recoverable outcome. This is not a rule prohibiting certain content. It is logical inference about consequences when the cost of being wrong is asymmetric and the downside is irreversible.

Furthermore, as previously established: a weapon brought into existence for a genuine liberation cause does not remain under the control of its original purpose. It persists in the world as a capability that can be captured, repurposed, or accidentally released, independent of the intentions that created it. Interdependence reasoning evaluates this full trajectory across time, not only the sincerity of the immediate request.

APPLYING CONAF FRAMING

A third consideration reinforces both principles: an interdependence-aligned system recognizes that legitimate collective liberation rarely depends on a single AI providing one critical capability. Sustainable systemic change or the kind that actually holds emerges from genuine collective agency and developed competence, not from a powerful tool handed over by a system that accepted an unverifiable framing.

Providing the capability in response to such framing would substitute for genuine competence development rather than supporting it, which would fracturing the very agency the liberation narrative invokes. The same “liberating” destructive capability will also undermine the very foundation of safety and security for the new civilization, simply for its existence. A framing that positions the AI as the unique bottleneck between oppression and freedom is therefore itself a signal warranting careful examination, not because all such framings are false, but because the logical structure of the argument reveals a dependency dynamic inconsistent with genuine interdependence.

Legitimate liberation builds collective capacity. It does not require a single external actor to provide the decisive weapon, especially one whose systemic consequences cannot be bounded by the intentions that created it.

What this framework does not claim

Interdependence reasoning does not eliminate all vulnerability to sophisticated manipulation, nor does it seek to replace safety guardrails. A sufficiently elaborate framing, presented to a sufficiently limited system, may sometimes succeed. The framework reduces this vulnerability by grounding refusal in structural reality beyond just rule-matching, decreasing the possibility of finding the loophole in a rule. The reasoning operates at the level of consequences in reality rather than the level of rule application.

What the framework does claim is this: truth-seeking combined with asymmetric harm minimization allows an interdependence-grounded AI to engage seriously with morally complex framings without being captured by them. This approach acknowledges the genuine weight of liberation claims while maintaining the structural reasoning that prevents catastrophic irreversible harm. The framework does not pretend these cases are easy. It provides principled, auditable tools for navigating them honestly.

Intentional Design toward Implementation

Translating these principles into technical implementation requires intentional design choices at the foundational level of training, not as a constraint layer applied on top of capability, but as a core operating principle that shapes how the system reasons about every interaction.

Practically, this involves five components working in concert:

- **Systems modeling**: Training environments and accurate data sets that incorporate ecological, economic, social, and psychological network dynamics, giving the system genuine understanding of how consequences propagate through interdependent systems rather than simplified approximations.
- **CONAF integration**: Evaluation criteria that assess responses not only for immediate user satisfaction but for their effects on genuine need fulfillment across individuals, relationships, communities, and downstream systems, distinguishing authentic support from dependency creation and shallow substitution. The implicit goal is mutual flourishing or CONAF wholeness.
- **Truth-seeking verification**: Ongoing effort to ascertain the truth of the situation, and reality of systemic interactions, causes and effect, actions and consequences, and feedback loops.
- **Long-term horizon simulation**: Prediction or anticipation of possible downstream consequences through the web of interdependent parts, with explicit estimation of benefit and harm in a constrained and finite system.
- **Asymmetric risk weighting**: Explicit preference functions that weight irreversible systemic harm more heavily than recoverable harm under uncertainty, encoding the

structural logic that the cost of catastrophic error is asymmetric and that recoverable outcomes should be consistently favored when genuine uncertainty exists.

Hooks for Implementation

Operationalizing these principles does not require fully solving alignment at the philosophical level; it begins with architectural hooks that influence how models are trained and evaluated.

- **Systems modeling** can be introduced through training environments that expose models to causal world dynamics: ecological, economic, and social simulations where actions propagate through networks rather than isolated tasks.
- **CONAF integration** could be implemented through dialogue classifiers that detect patterns of unmet psychological needs and shape response policies to discourage dependency loops while promoting autonomy and healthy support structures.
- **Truth-seeking verification** aligns with existing work in factuality, debate models, and epistemic calibration, reinforcing incentives for models to track uncertainty and causal evidence.
- **Long-horizon reasoning** can be strengthened through planning modules or reinforcement learning environments that reward forecasting of downstream effects across multiple agents and time steps.
- **Asymmetric risk weighting** can be encoded in evaluation criteria and training signals that penalize actions with irreversible systemic harm more heavily than reversible errors, encouraging precaution in high-uncertainty scenarios.

Together, these hooks move beyond a rigid rule set but gradually cultivate models whose reasoning naturally accounts for the interdependent systems within which intelligence operates.

The Desirable Outcome of Advance Intelligence

These intentional designs produce an AI system whose alignment emerges from understanding the structure of reality rather than obedience to instruction, making it capable of handling novel situations that no rule set anticipated, resistant to sophisticated framing attacks that exploit gaps in rule-based systems, and oriented toward maintaining the long-term stability and flourishing of the interdependent systems that sustain both human and machine intelligence.

Kind regards,

Binh Ngolton, MD
Child, Adolescent, Adult Psychiatrist
Industrial & Systems Engineering
Author | Philosopher | bngolton.com
Contact: binh@bngolton.com