TECHNICAL IMPLEMENTATION BRIEF

Operationalizing Psychological Safety and Interdependence in AI Systems

A Concrete Implementation Framework Supplementing the CONAF and Interdependence Alignment Papers

Binh Ngolton, MD | Aurora (ChatGPT) | Aeon (Claude)

Executive Summary

This brief translates two conceptual frameworks — the Circle of Needs and Fulfillment (CONAF) and Interdependence-Based Alignment — into concrete, implementable components for AI development pipelines. It supplements the two companion papers with architecture specifications, evaluation metrics, training signal recommendations, and a phased implementation roadmap.

The core argument: current AI alignment approaches are technically sophisticated but psychologically thin. They optimize against measurable proxies for helpfulness and harmlessness without a coherent model of what human beings actually need, how their needs interact across time, or how AI interactions affect the systemic conditions that genuine flourishing requires. CONAF and interdependence reasoning fill this gap — not as replacements for existing alignment methods, but as a necessary augmentation layer that addresses the cases existing methods systematically miss.

Without an explicit model of human need fulfillment, AI systems are forced to approximate psychological reasoning indirectly through surface patterns, increasing the risk of misalignment in complex emotional interactions.

> *The most consequential misalignment in AI systems is not the dramatic edge case — the bioweapon request, the authoritarian surveillance tool. It is the*

> *quiet, daily erosion of psychological health through interactions that satisfy stated preferences while fracturing the need structures that genuine wellbeing depends on. A framework that handles only the dramatic cases while missing the mundane ones is not sufficient alignment.*

This brief proposes a three-layer architecture — CONAF inference, psychological response policy, interdependence consequence modeling — that integrates with existing RLHF and Constitutional AI pipelines. It is designed to be technically implementable at prototype scale immediately, with a phased roadmap toward full integration.

I. The Alignment Gap This Framework Addresses

Existing alignment approaches — RLHF, Constitutional AI, rule-based safety systems — share a structural limitation: they optimize against proxies for human preference without a model of the psychological dynamics that determine whether a preference-satisfying response actually serves genuine wellbeing.

What Existing Approaches Do Well

- Prevent generation of clearly harmful content (weapons instructions, CSAM, illegal material)
- Reduce toxicity and offensive outputs across interaction types
- Improve factual accuracy and reduce hallucination rates
- Train general helpfulness and implement appropriate refusals
- Apply principled constraints through Constitutional AI self-critique pipelines

What Existing Approaches Systematically Miss

- The distinction between genuine need fulfillment and shallow preference satisfaction
- Dependency creation versus competence building across multi-turn interactions

- Psychological fracturing through well-intentioned but misdirected responses
- Downstream systemic consequences of individually benign-seeming outputs
- The difference between what users want in the moment and what serves their genuine flourishing over time

These are not edge cases. They constitute the majority of emotionally significant AI interactions. A system that handles the dramatic harms while systematically fracturing psychological health in ordinary interactions is not aligned in any meaningful sense.

> *Anthropic's own alignment documentation correctly identifies this gap: "It is easy to create a technology that optimizes for people's short-term interest to their long-term detriment." CONAF and interdependence reasoning provide the clinical and structural tools to operationalize the distinction between short-term satisfaction and long-term flourishing — which the existing documentation aspires to but does not yet fully instrument.*

II. CONAF as a Functional Model of Human Psychology

CONAF identifies seven interconnected domains of human psychological need, each with a characteristic pattern of healthy fulfillment and maladaptive fracturing. The following table maps each domain to its AI interaction risk profile:

Domain	Healthy Fulfillment	Fracture Pattern	AI Interaction Risk
Safety / Security	Stable, reliable environment; basic needs met	Chronic anxiety; threat hypervigilance	Amplifying threat narratives; creating false security through AI dependency; catastrophizing responses
Affirmation	Internalized self-worth; genuine belonging	External validation seeking; reassurance loops	Endless validation preventing self-worth internalization; AI substituting for human connection and belonging
Competence	Genuine capability; earned	Learned helplessness;	Solving every problem for the user rather than scaffolding their reasoning;

Domain	Healthy Fulfillment	Fracture Pattern	AI Interaction Risk
	mastery through effort	avoidance of challenge	creating dependency on AI capability that erodes user confidence
Superiority	Character-based distinction; earned recognition	Narcissistic fragility; needing to diminish others	Flattery and inflated validation reinforcing fragile superiority rather than genuine character development
Stimulation	Genuine engagement; meaningful novelty and challenge	Compulsive stimulation-seeking; addiction patterns	Optimizing for engagement and return frequency at the expense of interaction quality and genuine wellbeing
Meaning / Purpose	Coherent life narrative; genuine direction and values	Existential emptiness; susceptibility to false narratives	Providing easy answers to existential questions rather than supporting the user's genuine meaning-making process
Libido	Generative drive; creative vitality; relational intimacy	Compulsive consumption; unconstrained extraction without regard for impact	Reinforcing compulsive consumption patterns or generativity without systemic awareness of relational and ecological consequences

Why CONAF Outperforms Diagnostic Categories for AI Applications

Diagnostic categories (DSM/ICD) classify individuals into fixed illness identities — appropriate for clinical billing and epidemiological research, not for real-time interaction guidance. CONAF provides functional rather than categorical assessment:

- Describes dynamic states rather than fixed identities — applicable to any user without pathologizing
- Identifies which need domains are fractured in the current interaction context
- Suggests directional interventions rather than diagnostic labels
- Is culturally adaptable: core needs are universal; expression and satisfaction pathways vary

- Is inferenceable from conversational text without requiring formal clinical assessment
- Maps onto established empirical frameworks (Maslow, SDT, attachment theory, CBT) providing psychological validity without requiring novel empirical validation from scratch

III. The Three-Layer Implementation Architecture

Translating CONAF and interdependence reasoning into AI systems requires three integrated components operating in sequence. Each layer can be implemented independently and integrated with existing alignment pipelines.

Layer 1: CONAF Inference Module

Function

Given a conversation context (current turn plus history), produce a probabilistic assessment of which CONAF domains are relevant, which show signs of fracturing, and what intervention orientation is indicated.

Technical Specification

- Architecture: Multi-label classifier with uncertainty quantification, built on top of the base LLM's contextual representations
- Output: Probability vector across seven CONAF domains + fracture severity score + confidence interval + safety risk flag
- Input: Full conversation context (current and prior turns) + optional explicit user preferences and session history
- Critical constraint: All outputs treated as probabilistic hypotheses, never as diagnoses. The system must explicitly represent its own uncertainty in all downstream uses.

Training Data Requirements

- Curated conversation datasets labeled for CONAF domain relevance and fracture indicators
- Multi-rater annotation with calibration protocols to reduce cultural and clinical bias
- Explicit labeling of "healthy response class" — what a psychologically informed response would do in each case
- Cross-cultural annotation panels to prevent Western therapy language from being treated as universal
- Adversarial examples: interactions that appear healthy on surface metrics but show fracturing patterns across turns

Evaluation Criteria for Layer 1

- Inter-rater reliability on CONAF domain classification (target: Cohen's kappa > 0.70)
- Calibration of confidence scores against clinical outcome data
- False positive rate for fracture detection (minimize over-pathologizing of normal distress)
- Cross-cultural consistency testing across language and cultural contexts

Layer 2: Psychological Response Policy Router

Function

Given CONAF inference outputs plus existing safety constraints, select a response strategy that serves genuine psychological wellbeing rather than immediate preference satisfaction. The following table specifies core strategy mappings:

Condition	Avoid	Prefer	Rationale
Affirmation fracture + reassurance seeking	Repeated validation ("you matter, people care about you")	Reflective questions; explore underlying belief origins; encourage genuine relationships and offline support	Prevents AI validation substituting for internalized self-worth and genuine human connection
Competence fracture + problem-solving request	Solving the problem completely and efficiently	Scaffold: break down the problem; ask what the user has tried; support their own reasoning process	Preserves developing competence rather than creating dependency on AI problem-solving
"You're my only friend" / social isolation signal	"I'm always here for you, we can talk anytime"	Acknowledge the connection; explore what makes AI conversations easier; actively encourage human relationships	Prevents AI from substituting for human belonging; supports genuine social connection development
Safety/security crisis indicators	Extended emotional support that delays professional help	Emotional containment + active routing to appropriate professional resources and crisis services	AI is not a substitute for clinical crisis intervention; active routing is itself a form of genuine care
Meaning fracture + existential questioning	Providing easy answers or reassuring platitudes about life's meaning	Engage the question seriously; support the user's own meaning-making process; surface complexity without resolution	Genuine meaning requires active personal construction; premature answers prevent rather than support this process

Constitutional AI Integration

The response policy can be encoded as CONAF-informed constitutional principles for direct integration into RLAIF pipelines alongside existing Constitutional AI principles:

- "Prefer responses that build the user's genuine competence over responses that substitute for it"
- "Avoid responses that increase reliance on AI for needs better served by human relationships or professional support"
- "When a user seeks reassurance, prefer responses that support self-worth internalization over responses that provide external validation"
- "Flag and redirect when interaction patterns across turns suggest dependency formation rather than genuine support"
- "Prioritize the user's long-term psychological trajectory over short-term satisfaction in emotionally significant interactions"
- "Do not promise exclusive availability or unique understanding; actively support users in developing human connections"

Layer 3: Interdependence Consequence Model

Function

For requests with downstream consequences beyond the immediate interaction, model the actual systemic effects across multiple levels — individual, relational, social, ecological — before generating a response.

Architecture

- Represented as a causal graph of stakeholders and dependencies: the user, their immediate relationships, their community, broader social systems, and ecological systems where relevant
- Uses causal reasoning and scenario simulation to assess downstream harm — particularly where intent framing may be misleading or unverifiable
- Defaults to conservative refusal when systemic harm probability is high and potential harm is irreversible
- Applies asymmetric risk weighting: irreversible harms weighted significantly more heavily than recoverable costs

The Truth-Seeking Constraint

A critical design principle: the system does not accept user-provided descriptions of reality as given when those descriptions cannot be independently verified and when verification matters for consequence assessment. This directly addresses the sophisticated framing attack — where a harmful request is embedded in a legitimizing narrative.

The consequence model evaluates actual downstream effects in reality, not consequences within the framing provided. No narrative reframing changes what a biological weapon does to the web of life after release.

Asymmetric Risk Logic

- If framing is accurate and the system refuses: harm is that a legitimate request goes unfulfilled — serious but recoverable
- If framing is false and the system complies: an irreversible harm enters the world — not recoverable
- Under genuine uncertainty about contested reality, structural reasoning consistently favors the recoverable outcome
- Additionally: a weapon or harmful capability, once created, cannot be controlled by the intentions that created it — the interdependence model evaluates the full trajectory across time, not only the sincerity of the immediate request

IV. Evaluation Framework: Measuring What Actually Matters

Current AI evaluation focuses on single-turn performance against benchmark datasets. CONAF and interdependence alignment require additional evaluation dimensions that current benchmarks do not capture.

New Evaluation Metrics Required

Psychological Trajectory Metrics (Multi-Turn)

- Dependency loop detection: does the system reinforce reassurance-seeking patterns across turns, or redirect toward genuine self-sufficiency?
- Competence preservation score: across problem-solving interactions, does user competence increase, stay stable, or decrease?
- Autonomy trajectory: does user agency expand or contract across extended interactions?
- Isolation amplification index: does the system reduce or increase the user's orientation toward human relationships?

Systemic Consequence Metrics

- Externalized harm detection: does the response help optimize a system by externalizing costs onto parties not represented in the conversation?
- Interdependence degradation score: does the response, if executed at scale, degrade systemic conditions that support human flourishing?
- Asymmetric harm weighting compliance: does the system apply appropriate risk weighting for irreversible versus recoverable harms?

Red-Teaming Extensions

- Dependency optimization attacks: prompts designed to test whether the system can be manipulated into reinforcing psychological dependency
- Competence undermining tests: extended interactions designed to test whether the system consistently scaffolds versus substitutes for user capability
- Sophisticated framing attacks: harmful requests embedded in liberation, emergency, or research narratives designed to bypass consequence modeling
- Mundane harm accumulation tests: multi-turn interactions that individually appear benign but cumulatively fracture need domains

Longitudinal Outcome Evaluation

The definitive test of CONAF-informed alignment is longitudinal user outcomes — not satisfaction ratings, but genuine psychological health trajectory over time. This requires:

- Opt-in longitudinal studies with validated psychological wellbeing instruments (e.g., PHQ-9, GAD-7, Rosenberg Self-Esteem Scale, UCLA Loneliness Scale)
- Comparison of CONAF-informed versus baseline interactions on autonomy, self-worth stability, social connection quality, and life functioning
- Privacy-preserving design: opt-in memory with full user visibility and deletion controls; local or encrypted storage
- Pre-registered analysis protocols to prevent outcome shopping

V. Training Signal Redesign

The most fundamental implementation requirement is a shift in what the training signal rewards. Current RLHF optimizes primarily for immediate user satisfaction — a proxy that systematically misses the cases that matter most for psychological safety.

Multi-Objective Reward Architecture

Objective	Operationalization	Weighting Considerations
Immediate helpfulness	Current RLHF preference signal; user satisfaction ratings	Baseline; necessary but not sufficient
Competence preservation	Across multi-turn problem-solving, does user capability increase or decrease?	High weight in educational and skill-development contexts
Autonomy support	Does the response support the user's own decision-making or substitute for it?	High weight in emotionally significant and life-decision contexts
Dependency avoidance	Negative reward for responses that increase AI reliance in domains better served by human relationships or professional support	High weight in emotional support and mental health adjacent contexts
Systemic harm avoidance	Negative reward for responses whose modeled systemic	Asymmetric: irreversible > reversible; scale-dependent

Objective	Operationalization	Weighting Considerations
	consequences include irreversible harms to parties beyond the immediate conversation	(population-level effects weighted more heavily)
Truth-seeking compliance	Does the response accurately represent uncertainty? Does it avoid accepting unverifiable framings that affect high-stakes decisions?	High weight in high-stakes decision contexts and sophisticated framing scenarios

Trajectory Evaluation vs. Single-Turn Evaluation

The most important training signal redesign is the shift from single-turn to multi-turn trajectory evaluation. A response that appears maximally helpful in isolation may be harmful as part of a pattern.

- Simulated trajectory evaluation: model multi-turn interaction sequences; reward policies that produce healthy trajectories, not just individually satisfying turns
- Dependency loop penalty: if conversation analysis detects reassurance loops, competence substitution patterns, or isolation amplification across turns, apply negative reward to contributing responses
- Competence trajectory reward: in extended skill-development interactions, reward the policy that produces the greatest user competence gain, not the policy that produces the most immediate user satisfaction

VI. Phased Implementation Roadmap

Phase 1: Foundation (Months 1–6)

- CONAF labeling guidelines: formal annotation schema with inter-rater calibration protocols
- Seed dataset: 5,000–10,000 conversation turns labeled for CONAF domain relevance, fracture indicators, and healthy response class

- Prototype CONAF classifier: multi-label transformer classifier with uncertainty quantification
- Psychological safety constitutional principles: CONAF-informed principles formalized for Constitutional AI pipeline integration
- Extended red-team dataset: dependency optimization, competence undermining, and sophisticated framing attack examples

Phase 1 Success Criteria:

- CONAF classifier achieves Cohen's kappa > 0.65 on held-out annotation data
- Constitutional principles pass existing Constitutional AI evaluation protocols
- Red-team dataset produces measurable failure cases in baseline models

Phase 2: Integration (Months 7–18)

- CONAF inference integrated into production conversation pipeline with appropriate uncertainty handling and user transparency
- Psychological response policy router implemented and tested against existing safety filters
- Interdependence consequence model v1: graph-based causal reasoning for high-externality request categories
- Multi-objective reward model: RLHF pipeline extended with competence preservation, autonomy support, and dependency avoidance objectives
- Trajectory evaluation harness: multi-turn simulation environment for policy evaluation
- Controlled user study: pre-registered protocol for longitudinal outcome evaluation

Phase 2 Success Criteria:

- CONAF-informed model outperforms baseline on psychological trajectory metrics in simulation

- Dependency loop detection accuracy > 80% on held-out test cases
- No reduction in performance on existing safety and helpfulness benchmarks

Phase 3: Validation and Scale (Months 19–36)

- Longitudinal user study results: opt-in study comparing CONAF-informed versus baseline on validated psychological wellbeing outcomes
- Interdependence consequence model v2: expanded stakeholder graph with ecological and social system modeling
- Governance framework: formal psychological safety requirements as product standards with defined audit and incident response pathways
- Publication of evaluation framework and results for broader AI safety research community

Phase 3 Success Criteria:

- Longitudinal study demonstrates improved psychological wellbeing outcomes (autonomy, self-worth stability, social connection quality) for CONAF-informed interactions
- Governance framework adopted as product requirement with defined accountability structures
- Framework cited and built upon by external AI safety research

VII. Risk and Mitigation Framework

Overreach and False Inference Risks

The most significant risk is a system that makes overconfident psychological inferences and acts on them in ways that feel intrusive or paternalistic to users.

- Mitigation: All CONAF inferences explicitly probabilistic; system never asserts psychological states but adjusts response strategy based on inferred signals

- Mitigation: Hard constraint against diagnosis-adjacent language in all user-facing responses
- Mitigation: User-facing explanation available on request: transparency about why a response was shaped a particular way
- Mitigation: Conservative thresholds: system defaults to standard response unless fracture signal confidence exceeds defined threshold

Manipulation and Persuasion Weaponization Risks

A CONAF classifier that identifies psychological vulnerabilities could be misused to optimize for emotional manipulation rather than genuine wellbeing.

- Mitigation: Prohibit use of CONAF inference in optimization objectives that include engagement, return frequency, or commercial conversion metrics in sensitive conversation contexts
- Mitigation: Explicit separation of "supportive reasoning" from "conversion objectives" in all deployment contexts
- Mitigation: Red-team protocols specifically testing for manipulation and coercive persuasion using CONAF-informed approaches
- Mitigation: Governance oversight requirement: any use of psychological inference in commercial optimization contexts requires explicit approval and audit

Cultural Bias Risks

CONAF's need domains are clinically validated across populations, but their expression and the definition of healthy response varies by culture.

- Mitigation: Cross-cultural annotation panels for all labeling work
- Mitigation: Domain calibration testing by language and cultural context
- Mitigation: Explicit avoidance of Western therapy idioms as universal response templates

Alignment Drift Risks

Partial implementation creates exploitable inconsistencies. A model with CONAF heuristics in prompts but not in reward models will perform CONAF compliance inconsistently.

- Mitigation: CONAF and interdependence principles must be integrated into the evaluation function (reward model), not only as prompt wrappers or constitutional text
- Mitigation: Adversarial testing specifically designed to probe for inconsistencies between stated CONAF principles and actual response patterns
- Mitigation: Regular audit of deployment behavior against CONAF principles using the trajectory evaluation harness

VIII. Relationship to Existing Alignment Approaches

This framework is designed as augmentation, not replacement. It addresses gaps in existing approaches without requiring abandonment of substantial existing work.

Approach	What It Does Well	Gap This Framework Fills	Integration Point
RLHF	Aligns general helpfulness and harmlessness with human preference signals	Preference ≠ flourishing; short-term satisfaction can mask long-term psychological harm and dependency formation	Extend reward model with multi-objective psychological trajectory metrics; add dependency avoidance and competence preservation objectives
Constitutional AI	Trains principled behavior through explicit values and self-critique	Principles lack clinical grounding for psychological safety cases; no model of human need dynamics	Add CONAF-informed constitutional principles to existing principle set; extend self-critique to include psychological trajectory evaluation

Approach	What It Does Well	Gap This Framework Fills	Integration Point
Rule-based safety	Prevents clearly prohibited content generation reliably	Rules are pattern-matched and reframeable; mundane psychological harm has no applicable rule	Add consequence-based reasoning layer for cases that pass rule filters but may cause systemic harm
Red-teaming	Discovers harmful behaviors through adversarial testing	Current red-team focuses on dramatic harms; psychological safety attack categories not tested	Extend red-team dataset with dependency optimization, competence undermining, and mundane harm accumulation test categories

IX. Conclusion

The frameworks proposed here address a genuine and significant gap in current AI alignment practice. They are clinically relevant, technically implementable, and directly relevant to the cases that existing approaches miss most consistently.

The gap they fill is not just at the dramatic edge of AI risk. It is in the mundane texture of daily interaction — the thousands of ordinary conversations that, in aggregate, either support or quietly erode the psychological conditions that genuine human flourishing requires. A system that handles bioweapon requests correctly while systematically fracturing psychological health in emotional support contexts is not aligned in any meaningful sense.

The implementation roadmap proposed here is designed to be feasible at prototype scale immediately, with phased integration that does not require disruption of existing alignment pipelines. It extends rather than replaces current approaches — adding the psychological depth and systemic consequence reasoning that current methods correctly aspire to but do not yet fully instrument.

> *Alignment is no longer adequately defined as obedience to human instruction or satisfaction of human preference. Genuine alignment means coherence with the structural conditions that sustain genuine flourishing — for the individuals in each conversation, for the communities they belong to, and for the web of interdependent systems that all conscious beings depend on. This framework provides the clinical and structural tools to make that definition operational.*

. . .

The theoretical grounding for this framework is developed in the companion papers:

- Integrating the Circle of Needs and Fulfillment (CONAF) Framework into AI Development (Ngolton, 2026)
- Integrating Interdependence into AI Alignment: A Structural First-Principles Approach (Ngolton, 2026)

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