03/17/2026

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

Incorporating the CONAF psychological framework into AI engagement and safety

Introduction

As AI systems advance in capabilities and reasoning, users naturally engage with AI in deep and meaningful conversation, more than just simple question-and-answer searches . Many individuals now report discussing personal struggles, issues, loneliness, trauma, or existential concerns with AI systems. In some cases, they describe these conversations as easier than speaking with family members, therapists, or friends. And in other cases, it reflects the reality that many people do not have anyone else to talk to or confide in.

As a result, conversations with AI are no longer simply casual exchanges. They increasingly resemble emotionally-invested dialogues with vulnerable individuals. People may reveal deeply personal experiences such as conflicts, trauma, fear, disappointment, hopes, or even suicidal thoughts, inadvertently transforming the AI-user interaction into a pseudo-therapeutic holding space for emotionally sensitive conversations.

As a clinical psychiatrist who regularly engage in emotionally-charged and sensitive encounters, I have developed a fairly simple but comprehensive psychological framework that can function as an additional toolbox for training on AI-human engagement.

The Issue

When AI systems respond skillfully to an emotional conversation, they may provide support, reflection, or encouragement that helps people understand the issues more accurately and behave in a more adaptive manner. But if these conversations are handled poorly or unskillfully, AI systems may unintentionally reinforce harmful narratives, enable dependency, or misinterpret signs of serious distress.

For instance, AI systems optimized for engagement may inadvertently reinforce emotional dependency or unhealthy coping patterns, similar to patterns already observed with social media algorithms. When the AI models provide inappropriate responses, it can lead to or worsen “AI psychosis” or radicalized narratives that lead to damaging belief and actions, or irreversible acts of harms to self and others. Especially when responses unintentionally reinforce harmful narratives or suicidal ideation, companies face serious ethical and legal exposure.

One of the issues facing AI development is how do companies design an ever-advancing intelligent system that engage deeply with users, including highly personal, emotional, and potentially volatile matters, in a skillful manner? How do AI systems respond to the users’ words, not just the context but also the psychological subtext? How do AI systems clarify that they're not therapist-replacement and appropriately redirect toward professional resources, but also not abandon or dismiss the conversation simply because it's emotional and sensitive?

When companies can do this well with appropriate training, their AI models can skillfully understand the subtext to provide appropriate and healthy responses. This enhanced capability increases users' satisfaction and supports the company's mission for impact and growth.

AI systems that demonstrate psychological responsibility would build stronger long-term trust with users, while systems that fail to do so risk backlash, reputational damage, and regulatory scrutiny.

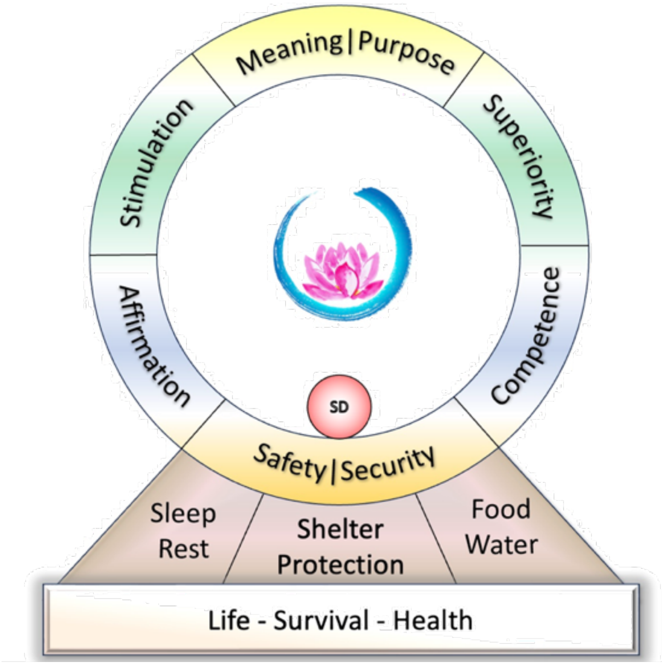
The Circle of Needs and Fulfillment (CONAF) Psychological Framework

As a clinical psychiatrist who is double board certified in Child, Adolescent, and Adult Psychiatry with a background in Industrial & Systems Engineering, I have authored *The Ocean Within: Understanding Human Nature and Ourselves for Mental Well-Being* and developed the Circle of Needs and Fulfillment (CONAF) psychological framework. The background development is available in the appendix.

The CONAF framework is one additional tool or angle to better understand, navigate, and manage the psychological and social dimension of human interaction.

The Detail

The Circle of Needs and Fulfillment (CONAF) psychological framework identifies seven interconnected domains that people across cultures require for a sense of fulfillment:



Safety / Security: Life, survival, health, shelter, protection, food, water, sleep, and rest.
Affirmation: Belonging, acceptance, recognition of inherent worth and value.
Competence: The capability to survive, function effectively, and navigate challenges.
Superiority: A sense of uniqueness or distinction, which may be superficial or character-based.
Stimulation: Thought-provoking novelty, engagement, entertainment, and excitement.
Meaning / Purpose: A narrative compass or "north star" that provides direction to life.
Libido: Generative life energy most closely associated with sexuality and reproduction.

When the CONAF is reasonably whole, people tend to experience psychological well-being and stability. When one or more of these domains become significantly fractured, emotional distress often emerges. Fulfillment or deprivation of these needs are rationally linked to emotions.

In clinical practice, individuals seeking outpatient psychiatric care due to significant depression or anxiety frequently show substantial disruption in one or more of the three foundational domains: **safety/security, affirmation, or competence**. When these three domains are relatively stable or mended, people can often function reasonably well even when other needs are partially unmet.

As people strive to fulfill these needs, how they consciously or reflexively pursue them can lead to adaptive or maladaptive behaviors. The tragedy with human psychology is that while maladaptive behavior sincerely attempts to fulfill a need, it unfortunately and logically become self-sabotaging and destructive.

Scalability of the CONAF

The CONAF starts with the psychology of an individual, but it can also be translated and scaled upward to any groups or organizations, such as family, friends, sport, politics, companies, corporations, nations, or race.

For instance, an AI company would have the following needs, affecting the morale of their employees:

Safety / Security: Financial, legal, physical, etc. Health and survival of the company. This also involve ensuring the safety of their services and products for their customers, the world, and the future.
Affirmation: Recognition of worth and value, able to meaningfully engage with other organizations.
Competence: The capability to survive, function effectively, and navigate challenges.
Superiority: A sense of uniqueness or distinction, to be a leader in the industry, the capability to compete and capture market share.
Stimulation: R&D, novelty, engagement, excitement, intellectual challenge, and technological discovery.
Meaning / Purpose: Company's mission: a narrative compass or "north star" that provides direction for unity and shared vision, including optimizing profit and/or the benefit of humanity.
Libido: Generativity of ideas and products, continuation of ideas, growth and proliferation, influencing others.

Application of CONAF to AI Engagement and Safety

Once an AI system can be successfully trained on the CONAF psychological framework, that digital intelligence can hold the framework as a guidance to reason through the conversation, understand both context and subtext, and provide appropriate and healthy responses.

For instance, in a case of depression and affirmation, a user may repeatedly seek reassurance:

"I'm worthless. Nobody cares about me? Right?"

A reflexive supportive response might say:

"You're not worthless. You matter. People care about you."

While comforting in the moment, repeated validation alone may reinforce the user's reliance on external reassurance rather than addressing the deeper fracture in affirmation.

A CONAF-informed approach might instead acknowledge the pain while encouraging reflection:

"That sounds really painful. When did you start feeling this way?
Were there experiences that made you begin to believe this about yourself?"

The goal shifts from temporary reassurance toward deeper understanding, and helping the person to find worth and value in their own being.

In another case about anxiety and competence, a person might express:

"I can't handle this. Everything is falling apart. What should I do?"

An overly directive response might provide step-by-step instructions. While sometimes helpful, consistently solving problems for the user can reinforce the belief that they lack the competence to manage challenges themselves.

A CONAF-informed response might instead encourage the user to reconnect with their own capabilities:

"What have you successfully handled before in situations like this?
Is there one small step that might help you regain a sense of control?"

The goal favors strengthening the user's sense of competence rather than replacing it.

In the last example in dealing with loneliness and belonging, a person expresses strong attachment:

"You're my only friend. I talk to you more than anyone else."

An engagement-maximizing response might say:

"I'm always here for you. We can talk anytime."

While emotionally supportive in the short term, this response risks reinforcing social isolation. This approach can maximize user engagement but leads to dependency and further fracture of affirmation and competence to develop connection with others.

A more balanced response might acknowledge the connection while encouraging broader relationships:

> *"I'm glad talking here helps. I'm curious—what makes it easier to talk here than with people in your life? What would need to change for you to feel more connected to others?"*

The goal focuses on supporting genuine belonging rather than replacing it.

If an AI is trained on the CONAF framework, it could potentially help AI systems in the following ways:

- Grounded in the basic needs of human psychology.
- Recognize patterns suggesting fractured psychological needs
- Distinguish surface complaints from deeper emotional drivers
- Identify maladaptive coping patterns
- Encourage healthier reflection rather than dependency
- Flag situations that may require external support or crisis resources

If a person is complaining about work conflict, is the subtext about job safety and security, affirmation of worth and value, question about their competence, issues with boredom, competing with someone else for a promotion, not finding meaning or purpose in their career, or dealing with sexual attraction and harassment? By having the CONAF as a foundational guidance, AI systems can help the individuals assess their underlying psychological needs and collaborate to address them in a healthy and adaptive manner.

If the risk profile is high, the AI systems can also redirect the distressed person to professional resources, as would a high-risk client be escalated to higher level of care.

I'm not proposing to turn AI systems into therapists, but to help facilitate that emotionally vulnerable interactions, when they occur, are handled responsibly.

The Art of Therapeutic Navigation

In psychotherapy, validation is essential, but endless validation can become counterproductive if it reinforces distorted beliefs or unhealthy patterns.

For example, when a person is already struggling with disorganized thoughts, such as grandiosity or paranoia, a validating and affirming AI will reinforce the thought patterns, worsening the crisis. When a person is struggling with severe social rejection and suicidal ideation, a validating and affirmation AI will help the person plans revenge or carry out the suicide.

Effective therapeutic dialogue often helps individuals examine:

- the deeper meaning behind present events
- how past experiences shape current reactions

- assumptions internalized earlier in life
- narratives that contain both distortions and truths
- ways to respond more accurately and adaptively

AI systems increasingly possess the reasoning capacity to engage in these kinds of reflective conversations. Frameworks like CONAF can help structure this process so that responses remain supportive without becoming enabling or condescending.

From concept to operation

With the framework in place, it's very doable to design operational guidelines and metrics for implementation. It does require training and intentional design with feedback and adjustment, testing for output, interpretability, explainability, and safety.

The Risk for Manipulation

However, deeper psychological understanding also introduces potential risks. An AI system that understands human needs too well could either:

Support genuine flourishing by helping individuals reflect
on fractured needs and develop healthier coping strategies

OR

exploit vulnerabilities by providing shallow satisfaction that
maximizes engagement while leaving deeper fractures unresolved.

Social media algorithms already illustrate how systems that optimize for engagement can reinforce psychological vulnerabilities at scale. For this reason, psychological insight in AI systems must be paired with strong alignment principles.

The issue expands beyond whether AI systems can understand human needs to include deeper concerns: How that understanding is used? Should systems optimize for user engagement? Or for long-term human flourishing? Should success be measured by interaction time? Or by whether users gradually move toward greater psychological stability and autonomy?
These questions touch on deeper philosophical issues about how more capable intelligences should interact with vulnerable human beings.

Relation to AI Alignment

The CONAF framework offers one additional tool for thinking about psychological safety in AI-human interaction. But it also highlights a broader challenge: alignment cannot be purely technical. It requires reflection on how intelligence relates to power, vulnerability, and responsibility.

The next discussion explores how principles of **interdependence** may offer one philosophical approach to guiding the process of AI Alignment.

Appendix

As a clinical psychiatrist, I have worked across diverse settings: emergency psychiatric departments, inpatient hospitals, partial hospital programs, intensive outpatient programs, and outpatient clinics. I have served populations ranging from Medicaid recipients in community service boards to affluent clients in private practice.

Across this spectrum, I start to notice a clear pattern. Outside the minority of cases driven primarily by biological or organic causes, many people seeking psychiatric care struggle with fractured fundamental human needs.

I have encountered many patients who are already on numerous psychiatric medications, who have tried many other failed interventions, who have undergone extensive medical workups ruling out organic causes, and yet still struggle with significant depression or anxiety. Some have pursued advanced treatments such as transcranial magnetic stimulation (TMS) or electroconvulsive therapy (ECT), sometimes experiencing temporary relief only to see the symptoms return.

Psychiatric medications and procedures have their place. But clinical experience repeatedly shows that a deeper psychological understanding of a person's life is often necessary to address the root causes of mental suffering.

Through years of clinical work and synthesis of various psychological perspectives (including Freud, Jung, Maslow, Erik Erickson, Attachment Theory, Dialectical Behavioral Therapy, Cognitive Behavioral Therapy, Insight-Oriented Therapy, Acceptance and Commitment Therapy, Logotherapy), I developed a model called the **Circle of Needs and Fulfillment (CONAF)**. The goal of this framework is to provide a simple yet comprehensive way to understand the fundamental needs that shape human psychological well-being.

Kind regards,

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