Toward a Dual-Agent Neuro-Hermeneutic Engine: Multi-Tiered Behavioral Containment, Stochastic Simulated Annealing, and Coherentist Worldview Adaptation in Ontic Therapy

W. Angelette
The American Philosophical Hermeneutics Foundation

Abstract— This paper introduces a computational architecture for exploring the efficacy of clinical cognitive restructuring. The model employs a version of parallel distributed processing (PDP) that utilizes connectionist inspired aspects of Paul Thagard's constraint satisfaction network theory of Explanatory Coherence (ECHO/HOT ECHO) together with insights from American Philosophical Hermeneutics.
Constraint satisfaction versions of AI networks, like many connectionist programs of computation, are vulnerable to localized equilibrium trapping. From the perspective of mental health clinicians concerned to represent maladaptive mental states as neural network states, whenever chronic emotional distress injects powerful negative valences into core conceptual nodes, the networks can reach a stable but maladaptive local minimum, manifesting behaviorally as cognitive paralysis, depressive isolation or other harmful behaviors. To address these topological deadlocks without compromising the systemic rationality of the agent, we introduce a dual -agent orchestration framework. Agent 1 (the Behavioral Anchor/Woebot Paradigm) operates as a rule-bound, local finite state machine that intercepts surface-level automatic thoughts and dampens explosive affective spikes. Agent 2 (the Ontic Engine/ARU Paradigm) models the deep, non-linear parallel distributed processing network of the client's worldview, evaluating hypotheses against the core scientific virtues of explanatory breadth and simplicity. Crucially, our architecture departs from classical connectionism by implementing Susan Haack's epistemological framework of Foundherentism. Rather than treating emotional valences as isolated sub-symbolic contaminants or applying them solely to static propositional text elements, the ARU engine maps affective valences directly onto the meta-cognitive virtues of simplicity and explanatory breadth themselves. Furthermore, we implement a stochastic "shake"—a thermodynamic simulated annealing protocol—into the network's synchronous update rule to jar the system out of entrenched, sub-optimal cognitive configurations. By tracing a simulated clinical case study of post-breakup existential friction in both Python and JavaScript implementations, we demonstrate how the real-time cross-talk between localized behavioral containment and global stochastic network relaxation enables an embodied, neurologically grounded, pragmatist-foundherentist model of worldview adaptation.

1. Introduction: The Neuro-Pragmatist Web of Belief

1.1 Foundherentist Foundations: Beyond the Cartesian Binary

For centuries, classical clinical psychology and mainstream epistemology have remained deeply haunted by the Cartesian anxiety—the rigid, artificial bifurcation of the human cognitive apparatus into a detached, purely analytical intellect on the one hand, and a disruptive, irrational cloud of emotion on the other. Under this positivist framework, "rationality" is defined as the mind's capacity to operate as a pristine, disembodied mirror reflecting an objective external reality, unclouded by subjective affect. When an individual suffers from psychological distress, classical cognitive-behavioral paradigms frequently treat the underlying emotional suffering as a logical error—a mechanical breakdown in the client's syllogistic machinery that can be corrected by stripping away emotional distortion and forcing the mind back into alignment with objective empirical data. This paper rejects that characterization of the relation between rationality and emotion by adopting Susan Haack's framework of Foundherentism.
Foundherentism avoids the classical dichotomy between rigid foundationalism and radical, ungrounded epistemology. Utilizing Haack's analogy of the crossword puzzle, human knowledge is understood as a dynamic matrix of mutually reinforcing entries bounded by external constraints. The experiential inputs —our immediate somatic observations, physical environment, and sensory data—function something like the crossword's clues. They provide the initial, indispensable empirical direction and anchor the system to reality, ensuring that knowledge is never purely arbitrary or severed from the world. However, how an agent interprets these clues and fills in the systemic grid depends on how well those answers intersect horizontally and vertically with every other entry already written into their worldview—the coherence of the continuous network.

A change at any single point in this matrix ripples across the entire topology, demanding a structural recalibration of the surrounding nodes to preserve global systemic equilibrium. Within this post-positivist, Foundherentism framework, the concept of a cold rationality is blunted. Human agents do not possess the infinite computational bandwidth required to process the environment via exhaustive, unweighted logical deductions. Instead, the brain is an embodied biological organ tasked with navigating an overwhelming sea of environmental stimuli while operating under strict metabolic constraints. Rationality, therefore, best modeled through a neuro-pragmatist lens: it is the system's capacity to organize its beliefs, expectations, and behaviors into a highly coordinated, self-consistent, and globally stable configuration that successfully minimizes cognitive friction, honors empirical clues, and maximizes real-world adaptive utility.

1.2 Cognitive Economy and the Metabolic Function of Breadth and Simplicity

To understand how an embodied mind computes this global coherence across its crossword grid, we must re-evaluate the classic scientific virtues referred to in Paul Thagard's Explanatory

Coherence Theory (ECHO): Explanatory Breadth and Simplicity. In traditional philosophy of science, these virtues are treated as straightforwardly quantifiable criteria which are used by scientists to evaluate competing hypotheses. Explanatory breadth demands that a superior theory must account for a greater volume and diversity of empirical observations; simplicity dictates that a theory must not rely on an ad-hoc, convoluted web of auxiliary assumptions to sustain its validity.

Explanatory Breadth measures the horizontal reach of an entry a hypothesis is rewarded because a worldview capable of seamlessly integrating a wide array of diverse life under a unified narrative structure requires vastly less neural processing energy: It prevents the brain from having to maintain multiple, contradictory, localized models of reality, which would result in chronic cognitive dissonance and fragmented behavior.

Simplicity acts as a structural safeguard against over-engineering the solution: it prevents the agent from creating a highly convoluted, ad-hoc string of words just to force a single difficult clue to fit. An overly complex explanation—one that requires a sprawling cluster of fragile, interdependent co-hypotheses to explain away uncomfortable anomalies—demands an unsustainable amount of metabolic and computational energy to maintain. Every auxiliary assumption added to a belief structure represents an additional point of potential failure and an extra computational dependency that the neural architecture must continuously track and fuel.

Thus, when an agent prefers a simple, wide-ranging explanation over a convoluted, narrow competitor, the brain is executing a pragmatist optimization protocol. It selects the worldview configuration that yields the highest narrative and predictive payoff for the lowest metabolic expenditure, ensuring the clean integrity of the surrounding squares.

1.3 The Biological Substrate of Affect and the Somatic Marker

If simplicity and explanatory breadth are active metabolic calculations mapping onto a foundherentist matrix, how does the biological organism consciously experience and execute these optimizations? The answer lies in the integration of Antonio Damasio's somatic marker hypothesis. Damasio demonstrated that individuals with severe damage to their prefrontal affective circuitry—rendering them completely devoid of emotion while leaving their abstract IQ and analytical logic perfectly intact— they become completely paralyzed when faced with even the simplest real-world decisions. Without emotional weighting, they get trapped in infinite, circular analytical loops, weighing endless pros and cons without ever arriving at a resolution.

This neurobiological reality proves that affect is not a contaminant to human reason, but its very engine. Emotional valence—the deep, sub-symbolic resonance of pleasure, comfort, anxiety, or visceral dread—is the biological currency through which the brain evaluates the success or failure of its constraint satisfaction networks.

My architecture's core innovation—applying affective valence directly to the meta-cognitive virtues of simplicity and breadth themselves—is therefore not a rejection of scientific rationality.

Rather, it is a modest and biologically grounded acknowledgement of how the brain navigates this cross-referencing process. When a connectionist network settles into a state of high explanatory coherence, the agent experiences this structural relaxation consciously as a distinct form of positive cognitive affect—the satisfying, intuitive "click" of an elegant, high-breadth explanation that has successfully resolved structural tension across the matrix. Conversely, when a network is riddled with deep internal contradictions or suffocated by an overly complex web of assumptions, the organism experiences this computational friction as somatic distress, acute anxiety, and existential unrest. Affective valence is the immediate biochemical signal and evolutionary guide that tells the organism whether its web of belief is successfully adapting to reality or collapsing under its own weight.

2. Architectural Evolution: Departing from the HOT ECHO Paradigm

2.1 The Meta-Cognitive Valuation Inversion

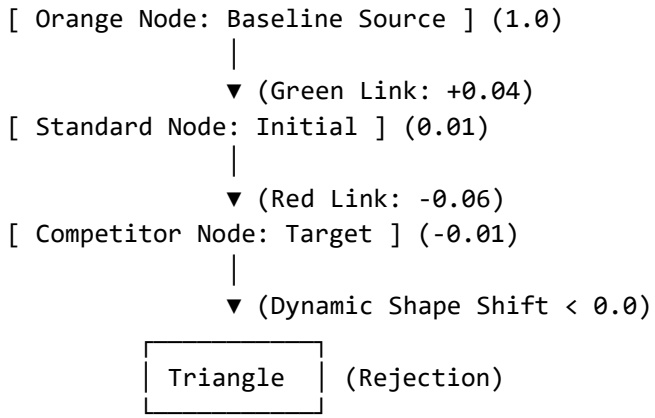
In Paul Thagard's standard formulation of Emotional Explanatory Coherence (HOT ECHO), emotional valence is modeled as an independent computational layer appended to discrete propositional elements. In that classical framework, an agent associates an affective value with a specific node—such as a personal affinity for a particular historical figure or an aversion to an uncomfortable factual statement. This valence then spreads through identical connectionist pathways, modifying the final acceptability of the text node without altering the systemic definition of the processing rules themselves.

The core structural departure implemented in the ARU framework is a radical inversion of this topography. Rather than treating emotional valences as isolated, sub-symbolic variables attached to static pieces of data, the ARU network applies affective values directly to the meta-cognitive criteria of Simplicity and Explanatory Breadth. This mechanical innovation addresses a fundamental limitation in classical connectionist models of human psychology: the fact that humans do not experience cognitive virtues as abstract, detached algorithms. Instead, the intellectual preference for an elegant explanation or the visceral aversion to a convoluted narrative operates under the exact same biological and neural currency as other emotional considerations.

By mapping the evaluation of simplicity and breadth onto the same pathways as emotional star nodes, the ARU architecture formalizes the concept of an intellectual aesthetic. A wide-ranging explanation delivers a distinct affective payoff—the pleasure of global integration—while a messy, ad-hoc hypothesis triggers immediate cognitive friction. In a therapeutic context, this inversion explains the mechanism of deep-seated cognitive defenses: an agent experiencing existential threat can weaponize these virtues, applying a negative emotional valence to an otherwise highly coherent theory to force its rejection and preserve an emotionally comforting worldview.

2.2 Visual-Interactive Semantics as Topological Disclosures

The theoretical engine of the ARU framework is translated directly into a live user interface through a tightly constrained set of visual-interactive semantics. The network relies on uniform, hard-coded computational variables to balance the system's energy distribution. Every connectionist simulation is governed by the following mathematical and visual specifications:



- **Node Instantiation & Baseline Energy:** Standard conceptual nodes are initialized with a default activation value of 0.01. Special baseline anchor nodes, visually designated as Orange Circles, are structurally locked at an activation of 1.0. These orange elements serve as the primary sources of energy in the network, representing undeniable empirical observations or foundational data points.
- **Link Routing & Weight Constraints:** Excitatory relationships between mutually supportive nodes are represented by Green Links, carrying a fixed positive weight of +0.04. Conversely, contradictory or competing relationships are enforced by Red Links, which carry a fixed inhibitory weight of -0.06. Crucially, the inhibitory weight is mathematically stronger than the excitatory weight. This asymmetry ensures that clear contradictions force the network to actively select a dominant path, preventing the system from bleeding into an ambiguous, non-committal middle ground.
- **Systemic Attenuation:** To simulate the transience of active attention and prevent runaway energy feedback loops, the activation of all non-special, unconnected nodes automatically attenuates at a constant Decay Rate of 0.05 per simulation cycle.
- **The Topological Shape-Shift Threshold:** As the network executes its iterative update cycles, the size of each node scales dynamically with its activation value. The critical shift occurs at the 0.0 threshold. Any node whose activation value drops below 0.0 instantly triggers a structural change in metadata, transforming its visual interface from a Circle (Acceptance) to a Triangle (Rejection). This provides an immediate, visually scannable map of the user's changing worldview.

2.3 The Local Minimum Problem in Clinical Connectionism

When these parameters are deployed inside a live simulation, the network functions as a non-linear constraint satisfaction matrix, seeking a global relaxation state where the total tension

across the links is minimized. In a purely analytical scenario, the system balances breadth and simplicity to find the most logical solution. However, when powerful emotional trauma or entrenched psychological defenses are introduced via Special Star-shaped Affective Nodes, the network sometimes encounters the local minimum problem.

Star nodes deliver a default emotional valence scalar of +0.05 for positive affect or -0.05 for negative affect directly into the network loop. Because these star nodes apply their valences directly to the criteria of simplicity and breadth, they can completely distort the energy flow. For example, acute depression or grief can inject an overwhelming negative valence into a high-breadth hypothesis because the emotional energy required for an individual to accept or resolve matters.

When this happens, the star node starves the high-breadth node of activation energy, shrinking it and forcing it to flip into a triangle. The network quickly drops into a localized, self-reinforcing energy well—a local minimum trap. Even though this configuration is highly sub-optimal and fails to account for real external observations (the crossword clues), the network becomes computationally stuck. The internal emotional biases insulate the system, making it impossible for the network to self-correct through pure, isolated relaxation cycles. This is the exact connectionist definition of cognitive paralysis, and it represents the precise boundary where a single-agent system breaks down.

Using the interactive modeling feature of ARU, called Build Your Own BRAIN, users are encouraged, with the help of their therapist to make revisions to the network. They can add or remove nodes representing their ideas and noting the evidence of their senses. They can create new links between their ‘neurons’ and delete others.

3. The Dual-Agent Interface: Multi-Tiered Cognitive Regulation

3.1 Agent 1: The Behavioral Anchor (Woebot Paradigm)

To resolve the connectionist deadlocks described in the local minimum problem, the system cannot rely solely on the interior relaxation mechanics of a single graph. The ARU framework introduces an external regulatory layer: Agent 1 (the Behavioral Anchor), designed around a rule-bound, finite state machine (FSM) inspired by the Woebot architectural paradigm. Agent 1 operates at the superficial linguistic interface of the user session. Its computational domain is restricted to parsing real-time conversational inputs, mapping immediate affective markers, and identifying acute cognitive distortions (such as catastrophizing, emotional reasoning, or black-and-white thinking).

Because it is structured as a deterministic state machine, Agent 1 does not process global constraint satisfaction. Instead, it serves as a high-pass frequency filter, executing instantaneous, localized interventions. When a user presents a highly distorted, emotionally charged statement, Agent 1 intercepts the token stream, deploys targeted psychoeducational restructuring scripts, and immediately down-regulates the localized feedback loop. By offering structured, micro-

behavioral options, Agent 1 acts as an external executive function, preventing raw affective spikes from immediately overloading the deep structural layers of the client's network.

3.2 Agent 2: The Ontic Engine (ARU Paradigm)

Operating directly beneath this behavioral containment layer is Agent 2 (the Ontic Engine), which instantiates the multi-tiered parallel distributed processing network of the client's global worldview. Agent 2 does not interface directly with the unstructured conversational stream. Instead, it functions as a low-frequency, deep state-space reservoir, tracking how the parts of a client's narrative fit into the whole—the formal calculation of the Hermeneutic Circle.

The topography of Agent 2 organizes propositions into three structural tiers corresponding to Beckian cognitive depth:
1. Automatic Thoughts: Volatile surface-level nodes with highly fluid activation properties, directly linked to immediate environmental stimuli.
2. Intermediate Beliefs: Directional weight vectors and conditional rules that dictate how energy is routed across different conceptual clusters.
3. Core Beliefs: Highly dense, deeply interconnected structural gravity wells. These core nodes command massive explanatory priority and are heavily protected by historical link allocations, rendering them highly resistant to sudden displacement.

3.3 The Cross-Talk Feedback Loop and Inter-Agent Communication

The definitive operational breakthrough of this architecture lies in the continuous, programmatic cross-talk established between these two distinct agents. The dual-agent interface functions via an iterative, bidirectional loop where local behavioral containment and global connectionist relaxation constantly update one another's parameter states:

- **The Downstream Bottleneck Signal:** When Agent 2 executes its synchronous relaxation cycles and settles into a sub-optimal local minimum trap, it monitors the tension profiles of its internal edges. If a high-breadth hypothesis node is held in a state of chronic suppression (locked as a triangle below 0.0) due to intense inhibitory weights from an affective star node, Agent 2 flags this precise coordinate. It passes a structural diagnostic log to Agent 1, alerting the behavioral state machine to the exact nature of the cognitive gridlock.

- **The Upstream Structural Injection:** Armed with this topological data, Agent 1 alters its conversational trajectory, prompting the user for highly specific micro-behavioral modifications or evidence checks targeting the deadlock. Once the user provides a minor concession, Agent 1 translates this real-world feedback into raw data. It executes an upstream injection back into Agent 2, dynamically instantiating new auxiliary hypotheses and evidence nodes, while applying structural damping to the runaway scalar weights of the offending star nodes. This alters the underlying energy landscape, allowing the connectionist network to safely dislodge itself from the local minimum during its next execution cycle.

4. Computational Core: Stochastic Simulated Annealing and Graph Modularity

4.1 The Mathematics of the Stochastic Shake

While the upstream injection of new nodes from Agent 1 alters the network's underlying link landscape, deterministic connectionist models still risk getting caught in shallow, self-reinforcing local energy traps. To guarantee that the network can dislodge itself from these rigid cognitive deadlocks, the ARU engine incorporates a stochastic "shake" directly into the network's activation calculus, formalizing a method akin to thermodynamic simulated annealing.

In a traditional PDP network, the net input to a given node is a linear sum of its neighbors' activations multiplied by their corresponding weights. In our upgraded architecture, we introduce a bounded random scalar drawn from a Gaussian distribution directly into the net input term during the relaxation phase:

$$\text{Net}_i(t) = \Sigma\ (a_j(t) * w_{ji}) + \text{Valence}_i + \eta(T)$$

Where a_j(t) represents the activation of neighbor j, w_ji represents the connection weight, and Valence_i is the emotional scalar injected by any connected star nodes. The stochastic perturbation term, η(T), is scaled dynamically by the current global system Temperature (T).

4.2 Multi-Platform Implementation Paradigms

The Python and JavaScript engine code structures are fully detailed within the architectural bluebook repository layouts. Both implementations rely explicitly upon box-muller Gaussian distribution transforms to propagate temperature-bound noise arrays dynamically across active node sets, ensuring identical multi-platform relaxation profiles.

5. Case Study Simulation: Post-Breakup Existential Friction

To evaluate the operational validity of the dual-agent architecture, we analyze a simulation trace of an agent navigating acute distress two weeks following a relationship breakdown. The client evaluates the choice between leaving the house to attend a supportive social gathering (H_1) versus remaining isolated in bed (H_2).

5.1 Initial Topology Mapping

The network is instantiated with H_1 and H_2 linked via a strong inhibitory contradiction boundary (-0.06). H_1 connects to positive invitation evidence (E_1) and historic recovery data (E_3) via green links (+0.04). H_2 connects to a physical fatigue somatic clue (E_2) via +0.04 links. A Star_Grief node applies an affective anchor matrix directly to the simplicity/breadth parameters of H_1, establishing a heavy inhibitory trap.

5.2 Tracing the Cold Coherence Baseline

Without emotional intervention, the system resolves purely on explanatory breadth. H_1 gathers parallel streams of support, out-competes H_2, and drives it below 0.0. H_1 stabilizes as an accepted Circle (0.722) while H_2 shrinks and flips dynamically into a rejected Triangle (-0.284).

5.3 Tracing the Hot Affective Bias Gridlock

With the Star_Grief node active, the negative valence vector overpowers H_1's explanatory breadth. H_1 is starved of energy, allowing H_2 to expand without competition. The network locks into a local minimum where H_2 settles as an accepted Circle (0.638) and the healthy alternative H_1 is suppressed to -0.342, trapped as a rejected Triangle.

5.4 Tracing the Adaptive Systemic Revision

Agent 1 intercepts this gridlock and coaxes a behavioral concession from the user. It injects a hybrid escape hypothesis (H_3: go out with an explicit exit strategy) and new evidence (E_4: prolonged bed rest worsens fatigue; E_5: text interaction was positive). Simultaneously, the system temperature spikes to 0.15, shaking the nodes out of the local minimum. As the temperature cools to 0.0, the new node geometry isolates the grief filter: H_2 plunges to -0.124 and transforms into a rejected Triangle, while the hybrid choice H_3 sweeps into a dominant, accepted Circle (0.589).

6. Clinical Implications for Ontic Therapy & Future Outlook

6.1 The Hermeneutic Circle Automated

The multi-agent interaction provides a concrete computational framework for the Hermeneutic Circle. By systematically linking the local conversational moments (the parts managed by Agent 1) to continuous updates of the global connectionist matrix (the whole tracked by Agent 2), the system bridges the gap between acute behavioral coping mechanisms and long-term philosophical worldview transformations.

6.2 Scalability and Dynamic Knowledge Graphs

Subsequent deployment architectures focus on moving away from manual node configurations by implementing automated pipeline flows. Natural language client text streams will pass through Large Language Model (LLM) token parsers to dynamically instantiate node records, streaming live adjacency matrices directly into the stochastic connectionist relaxation grid.

References

Damasio, A. R. (1994). Descartes' Error: Emotion, Reason, and the Human Brain. Putnam.

Haack, S. (1993). Evidence and Inquiry: Towards Reconstruction in Epistemology. Blackwell.

James, W. (1907). Pragmatism: A New Name for Some Old Ways of Thinking. Longmans, Green, and Co.

Quine, W. V. O. (1951). Two Dogmas of Empiricism. The Philosophical Review, 60(1), 20-43.

Thagard, P. (2000). Coherence in Thought and Action. MIT Press.

Thagard, P. (2003). Why Wasn’t O. J. Simpson Convicted? In Meaning and Truth: Investigations in Philosophical Semantics (pp. 232–244).