An Information-Theoretic Model of Cognitive Decline in Alzheimer's Disease

Author: Keenan Rogerson Leggett
Date: October 8, 2025

Abstract

This paper proposes a novel computational framework for understanding the cognitive symptoms of Alzheimer's disease (AD), particularly the phenomenon of confabulation. We move beyond the traditional view of cognitive decline as a passive decay of function and reframe it as an active, but pathological, self-correction process. We model the mind as a coherence-seeking system that is intrinsically motivated to maintain a consistent and unified model of reality. We argue that the initial neuronal loss in early AD creates critical "holes" in the system's foundational knowledge, leading to a catastrophic spike in what we term **Structural Tension**. The model posits that the subsequent cognitive cascade, including confabulation and delusion, is a misguided attempt by the brain's own coherence-monitoring mechanisms to resolve this tension by invalidating correct, high-level beliefs and learning a new, but factually incorrect, model of the world. This framework offers a new, mechanistic explanation for the "descent into madness" and suggests novel avenues for diagnostics and cognitive therapies.

1. Introduction: The Puzzle of Confabulation and Coherence

While the biological cascade of Alzheimer's disease is well-documented (amyloid and tau pathology leading to neuronal death), the link between this physical decay and complex cognitive symptoms remains obscure. Confabulation—the creation of false or distorted memories without intent to deceive—is typically attributed to memory failure. We argue that this perspective is insufficient. Confabulation is, instead, the **observable output of a desperate, active process of self-correction** by a brain attempting to maintain internal consistency when its foundational structure collapses.

Drawing upon principles of the **Bayesian Brain** hypothesis and **Predictive Processing (PP)**, which posits that the brain is an inference engine seeking to minimize prediction error (PE) (Friston, 2010), we model AD decline as a catastrophic failure of this inference process.

2. Theoretical Foundation: Predictive Processing and Coherence

The brain's primary goal, under the Free Energy Principle (FEP), is to minimize **Free Energy** (F), which serves as an upper bound on **surprise** (negative model evidence). Minimizing F is achieved through two main mechanisms:

1. **Perceptual Inference:** Minimizing **Prediction Error (δ)** by updating internal beliefs (μ).
2. **Active Inference:** Minimizing δ by changing sensory input (action).

Prediction error is the discrepancy between the expected sensory input (top-down prediction) and the actual sensory input (bottom-up signal). This error is weighted by its **precision** (Π or inverse variance), which reflects the perceived reliability of the signal.

Mathematical Formalism of Prediction Error

In a hierarchical model, the prediction error ($\delta^{(l)}$) at any layer (l) is defined as:

$$\delta^{(l)} = \tilde{\mathbf{s}}^{(l)} - g^{(l)}(\mu^{(l)})$$

Where:

- $\tilde{\mathbf{s}}^{(l)}$ is the bottom-up sensory input or error from the layer below.
- $g^{(l)}(\mu^{(l)})$ is the top-down prediction generated from the current state (belief) μ at layer l.

The minimization of the squared, precision-weighted error drives belief updates:

$$\min_{\mu} \sum_{l} \Pi^{(l)}(\delta^{(l)})^2$$

This minimization is the neural drive for coherence.

3. The Model's Core Concepts

3.1. Axiomatic Nodes (A): High-Precision, Load-Bearing Priors

We define **Axiomatic Nodes (A)** as high-level, extremely high-precision **Priors** (C) that represent fundamental, stable, and highly verified truths about the individual's identity and environment (e.g., "I am married," "I have three children," "My house is this specific structure").

- **Formal Property:** Axiomatic Nodes are characterized by $\Pi_{\mathrm{A}} \to \infty$, meaning their precision is assumed to be absolute. They have minimal variance and are extremely resistant to revision (i.e., their learning rate is near zero).
- **Neurobiological Substrate:** These are likely encoded by highly stable, distributed representations involving strong reciprocal connections between the **Hippocampus/Entorhinal Cortex** (critical for foundational relational memory, early AD target) and the **medial Prefrontal Cortex (mPFC)** and **Precuneus** (self-referential processing and the default mode network).

3.2. Structural Tension Spike (STS): Catastrophic Prediction Error

The initial, subtle neuronal death in early AD is modeled as the irreversible **deletion** of a foundational A node. When a high-level cognitive process (e.g., the Prefrontal Cortex, μ_{PFC}) attempts to verify a current belief (B) against this missing axiom ($C_{deleted}$), the prediction $g(\mu|C_{deleted})$ becomes non-computable or produces unbounded variance.

This generates a **Structural Tension Spike** (δ_{STS}):

$$\delta_{STS} \to \frac{1}{\Pi_B \to 0} \approx \text{Catastrophic Error}$$

Unlike a simple memory gap (which resolves by admitting uncertainty and low precision), the loss of a high-precision, load-bearing **prior** (Axiomatic Node) is interpreted by the system as a fundamental breakdown in the entire world model hierarchy.

4. The Three-Stage Pathological Cascade

Stage 1: The Structural Tension Spike (The "Hole" in Reality)

- **Mechanism:** A high-level belief B (e.g., "My husband is John") attempts to activate or verify against the foundational prior $C_{\text{John's identity}}$, which has been compromised by neurodegeneration (e.g., in the hippocampus).
- **Outcome:** The system receives an effectively infinite δ_{STS}, an immediate, system-wide threat to coherence.
- **Neurobiology:** Acute δ_{STS} likely correlates with massive, incoherent firing patterns, potentially measurable as a sudden failure of synchrony between the mPFC (high-level belief) and the compromised DMN/Hippocampal network (axiomatic prior). This may manifest as the initial, brief moments of profound confusion or disorientation often reported in early AD.

Stage 2: Dissonance Misattribution (The "First Self-Doubt")

This is the critical failure point, driven by **cognitive economy—**the system chooses the path of least computational resistance to minimize F. The system has two options to reduce δ_{STS}:

Option 1 (Healthy/Expensive): Acknowledge the internal structural failure.

- Conclude: "My own foundational prior (A) has degraded."
- Cost: This requires a massive, global revision of the entire high-level world model, which relies on the missing axiom. The $\mathrm{Complexity}$ cost of this deep structural change is extremely high, as A is connected to thousands of other beliefs.

Option 2 (Pathological/Inexpensive): Misattribute the error to the high-level belief (B).

- Conclude: "My high-level belief B is factually incorrect."
- Example: Rather than concluding "The structural representation of 'John' is gone" (high cost), the system concludes "The current belief 'This man is John' is wrong" (low cost, immediate δ reduction).

The model posits that the brain chooses Option 2 because $Cost(\text{Revise } B) \ll Cost(\text{Revise A})$. The system finds it easier to create a localized, false consistency ($\neg B$) than to admit global structural failure ($\neg\mathrm{A}$), which would halt inference entirely.

- **Neurobiology of Misattribution:** This decision process is hypothesized to rely on the **Dopaminergic system** and Π modulation. The system fails to assign high precision to the self-referential *source* of the error (the hippocampus/DMN), instead assigning high precision to the sensory *discrepancy* itself, which allows the PFC (decision-making) to rapidly update the belief B. This initial failure of error source localization is the root of the delusion.

Stage 3: Confabulatory Relearning (The "Descent into Madness")

Having pathologically revised belief B to $\neg B$, the system now has a new, locally inconsistent reality that must be explained. Confabulation is the active process of **Relearning—**creating a new, synthetic generative model $\widetilde{g}$ that is consistent with $\neg B$ while minimizing δ against ongoing sensory input.

$$\min_{\widetilde{\mu}} \Pi_{\delta}(\delta)^2 \quad \text{such that} \quad \widetilde{\mu} \vDash \neg B$$

- The system actively searches (via Active Inference/Action) and generates (via Perceptual Inference/Confabulation) new, false memories ($\widetilde{B}_{\text{confabulated}}$) to bridge the gap between sensory input and the new delusional prior $\neg B$.
- **Neurobiology:** This stage is characterized by over-reliance on the **Ventromedial Prefrontal Cortex (vmPFC)**, often implicated in generating creative narratives and integrating emotional/self-relevant content. Confabulatory relearning uses the intact (but now misguided) cortical structures to weave a new, coherent (though false) narrative tapestry around the structural "holes."

5. Implications for Diagnostics and Therapeutics

5.1. Novel Diagnostics: Probing Dissonance Misattribution

This framework predicts that a specific cognitive bias precedes profound memory loss: the tendency to resolve high-stakes cognitive dissonance by discarding a central belief rather than admitting uncertainty.

Testable Hypothesis: Early AD patients, compared to controls or patients with non-AD memory loss (e.g., transient global amnesia), will exhibit a significantly lower threshold for revising established, high-confidence prior beliefs (Π_{A}) when presented with novel, highly contradictory (but benign) evidence that challenges a core "Axiomatic Node."

5.2. New Therapeutic Avenues: Axiomatic Reinforcement

If the cascade is triggered by the initial loss of A, the therapeutic goal is to protect or artificially reinforce these nodes.

Axiomatic Reinforcement Therapy (ART): A form of targeted cognitive training designed to constantly and gently reinforce a minimal, core set of foundational "ground truth" memories (the patient's A set). This reinforcement must target the precision (Π) of these priors, not just the content.

- **Neurobiological Intervention:** ART would aim to use targeted, high-precision training to boost the synaptic efficacy and stability of the neural assemblies coding for A through Hebbian plasticity, essentially increasing their network centrality and resistance to AD pathology. Furthermore, neuromodulators, such as **Acetylcholine** (critical for assigning precision to prediction errors), could be leveraged to ensure the system correctly registers the A node's outputs as "high-precision," thus preventing its rapid invalidation during a Structural Tension Spike.

6. Conclusion

By reframing the cognitive symptoms of Alzheimer's disease as an active, but pathological, attempt by the brain to maintain its own coherence, we establish a fully mechanistic model. The tragedy of AD, according to this information-theoretic model, is a descent into madness driven by the mind's own desperate, but misguided, impulse to heal itself. This framework provides a unified language, mathematical justification, and neurobiological hypothesis for linking microscopic neuronal loss to the macroscopic, complex reality of delusion and confabulation.