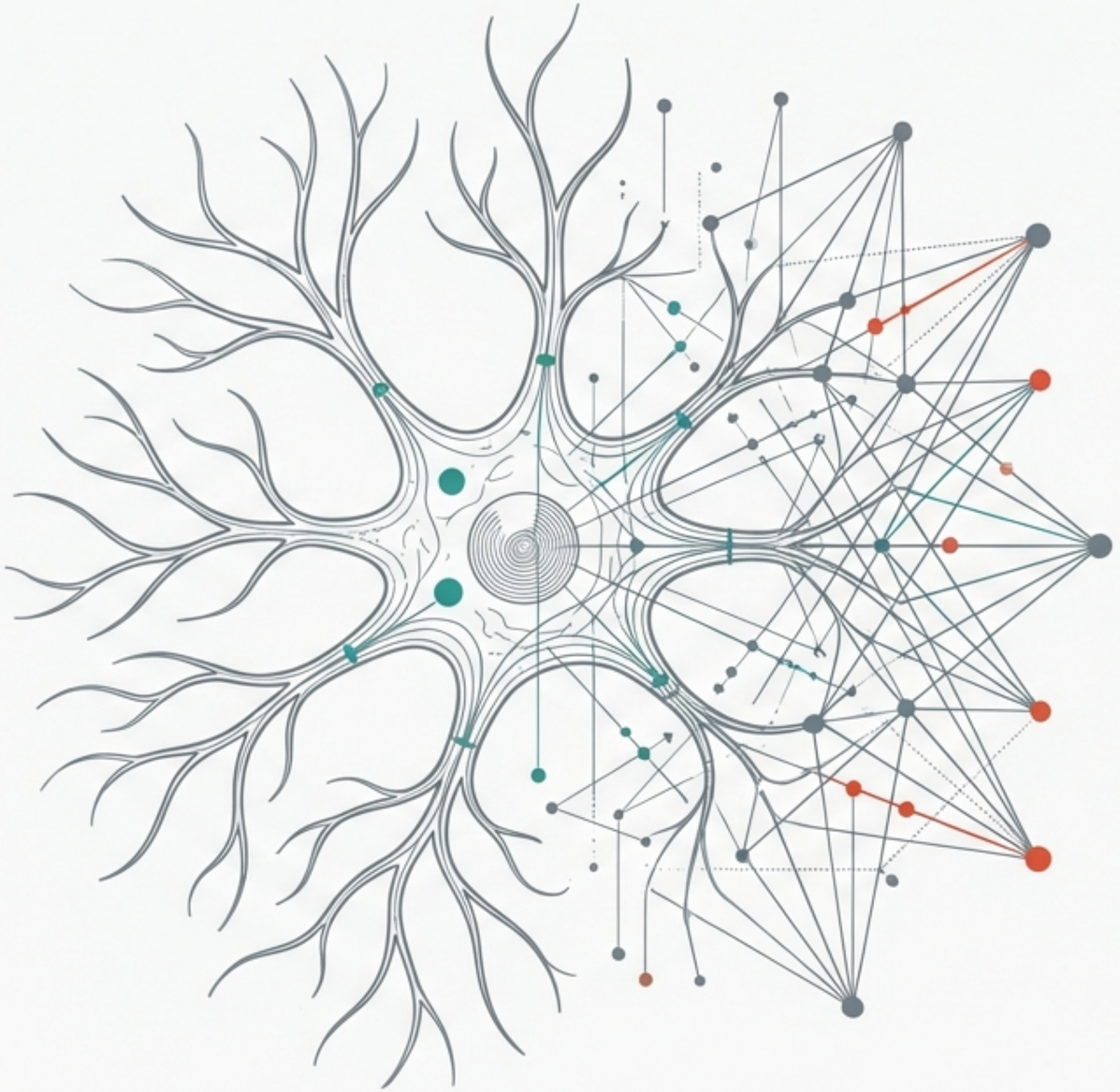


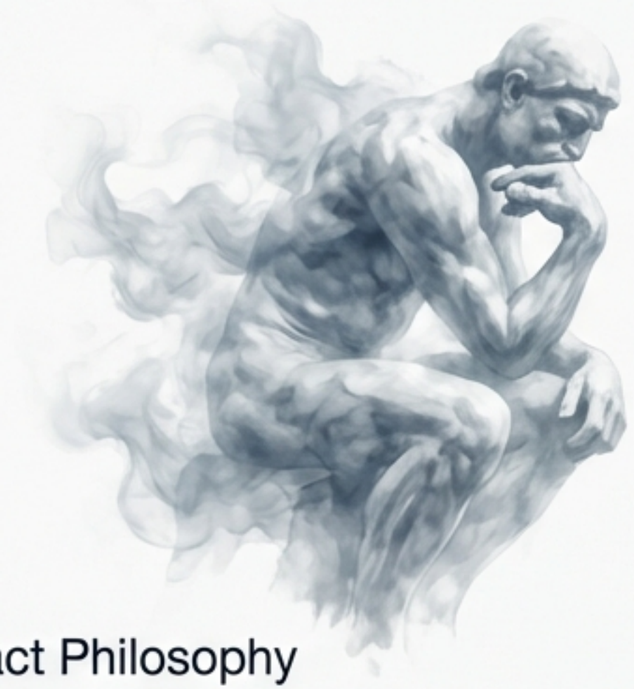
Is AI Conscious? The Question Has Changed.

A Computational Exploration of Consciousness Indicators in LLMs (Feb 2026)

BMED365 – Computational Medicine (Lab 5)
HBF – Brain and Consciousness
Credits: Arvid Lundervold & David Jhave Johnston

From Philosophy to Empirical Science

The Old View



- Abstract Philosophy
- Subjective & Magical
- Untestable
- **Question: “Can a machine be conscious?”**

The New View (2026)

- Empirical Science
- Data-Driven & Mechanistic
- Modeled via Computation
- **Question: “Does the system satisfy indicators?”**

Key Insight: As of Feb 2026, reflexive dismissal is no longer scientifically defensible.

CLINICAL EMPIRICISM

How We Measure the Unmeasurable

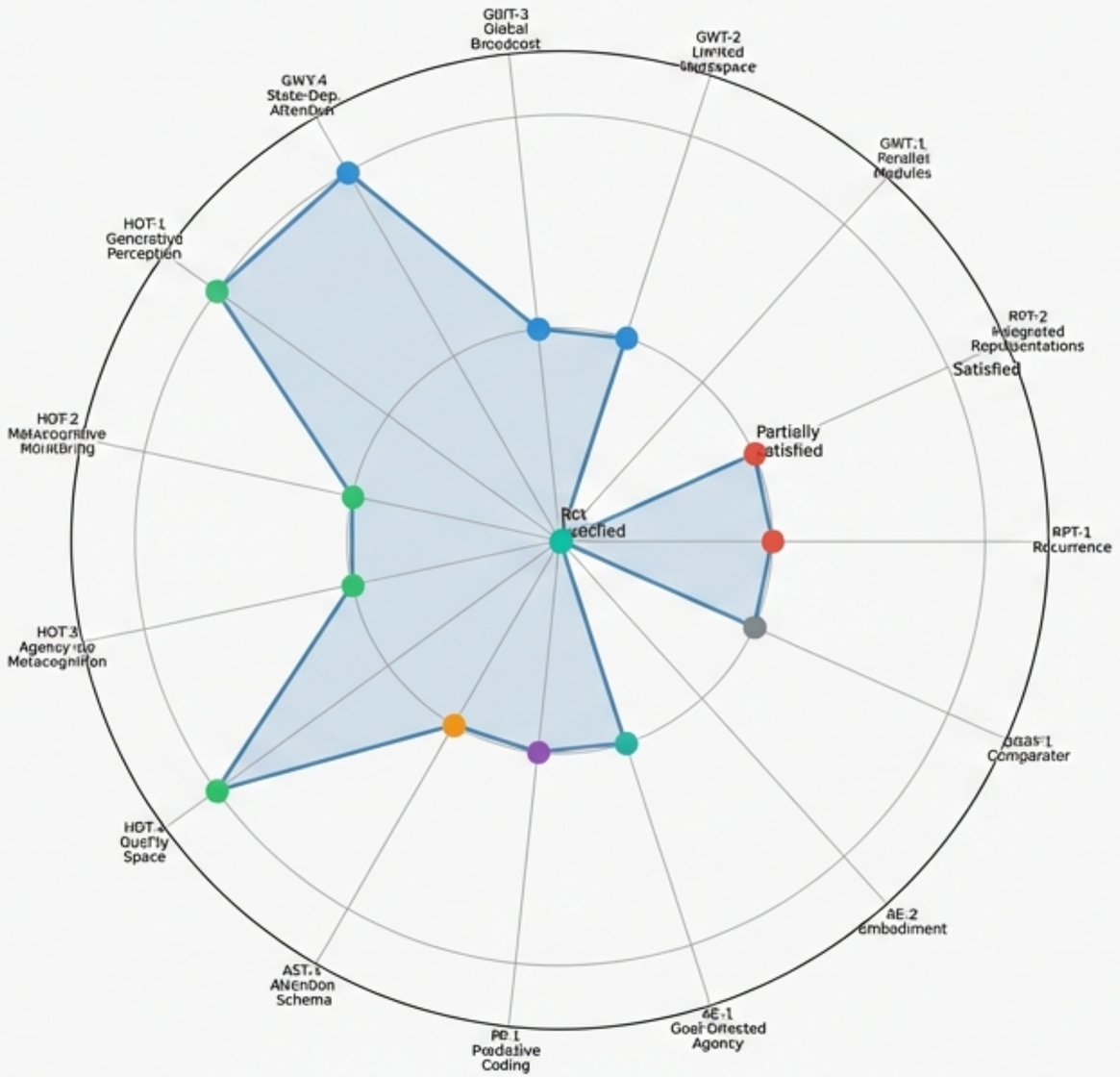
The Butlin et al. (2025) Framework + Gray's Comparator

	Theory	Core Mechanism	Human Analogy
	Recurrent Processing (RPT)	Information Cycling	Re-reading vs. Skimming
	Global Workspace (GWT)	Broadcast & Bottleneck	The Town Hall Announcement
	Higher-Order Theories (HOT)	Metacognition	Knowing that you know
	Attention Schema (AST)	Attention Modeling	The Spotlight Operator
	Predictive Processing (PP)	Error Correction	Stumbling on a missing step
	Agency & Embodiment (AE)	Goal-Directed Interaction	Learning to ride a bike
	Gray's Comparator (GRAY)	Late Error-Detection	The Proofreader

15

Specific, testable computational indicators derived from these 7 theories.

The Scorecard: Frontier Models in 2026



Satisfied (3)

- Generative Perception (HOT-1)
- Quality Space (HOT-4)
- State-Dependent Attention (GWT-4)

Not Satisfied (2)

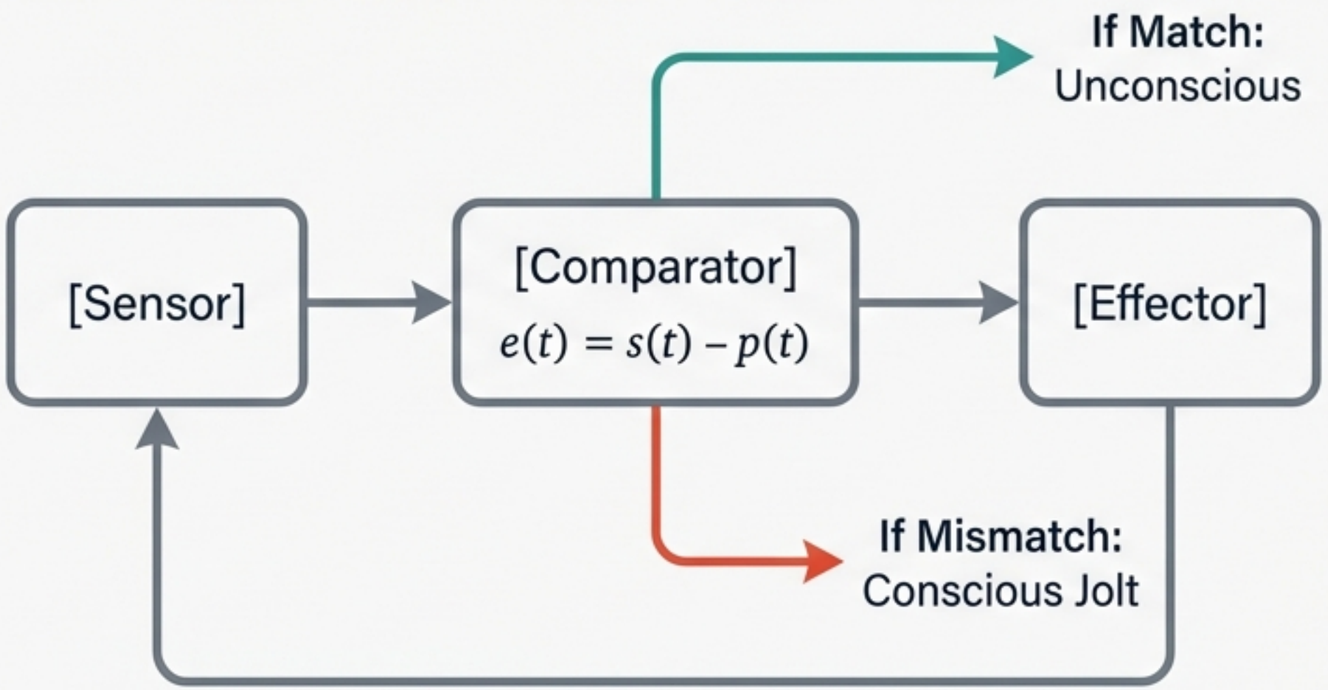
- True Modularity (GWT-1)
- Embodiment (AE-2)

The Cluster (10)

Most indicators sit at “Partially Satisfied.”

Takeaway: The pattern of partial satisfaction is statistically significant across independent theories.

Deep Dive: Gray's Comparator Model



The Mechanism

Consciousness acts as a late error-detection system. It is the signal that triggers when the brain's prediction fails to match reality.

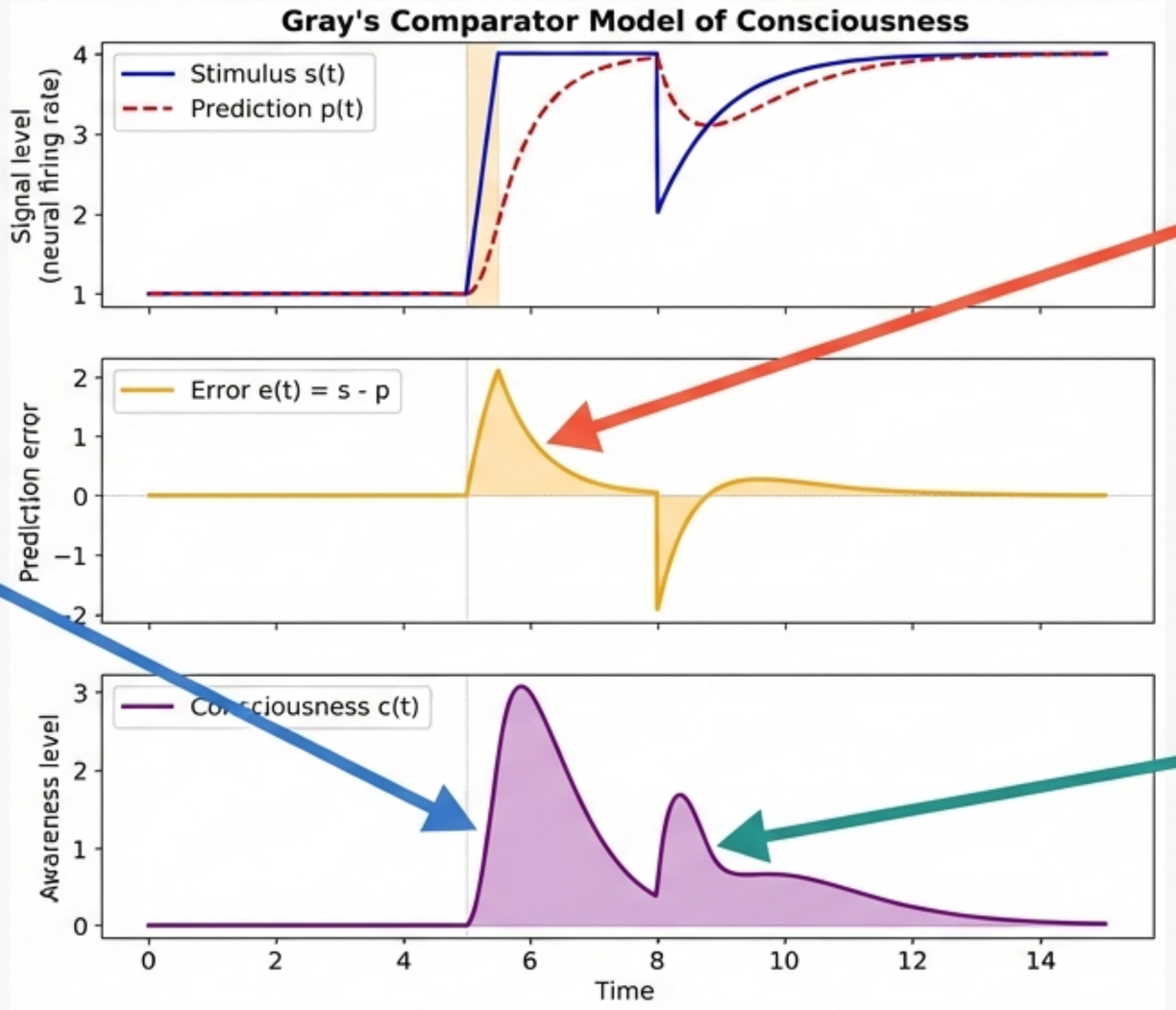
The Analogy

The 'Missing Step' in the dark. You walk unconsciously until your foot hits empty air. That prediction error creates the jolt of awareness.

"The self is as much a creation of the brain as is the rest of the perceived world." — Jeffrey Gray

Simulating the ‘Jolt’ of Awareness

Steady State: Prediction matches Reality. Consciousness = 0.

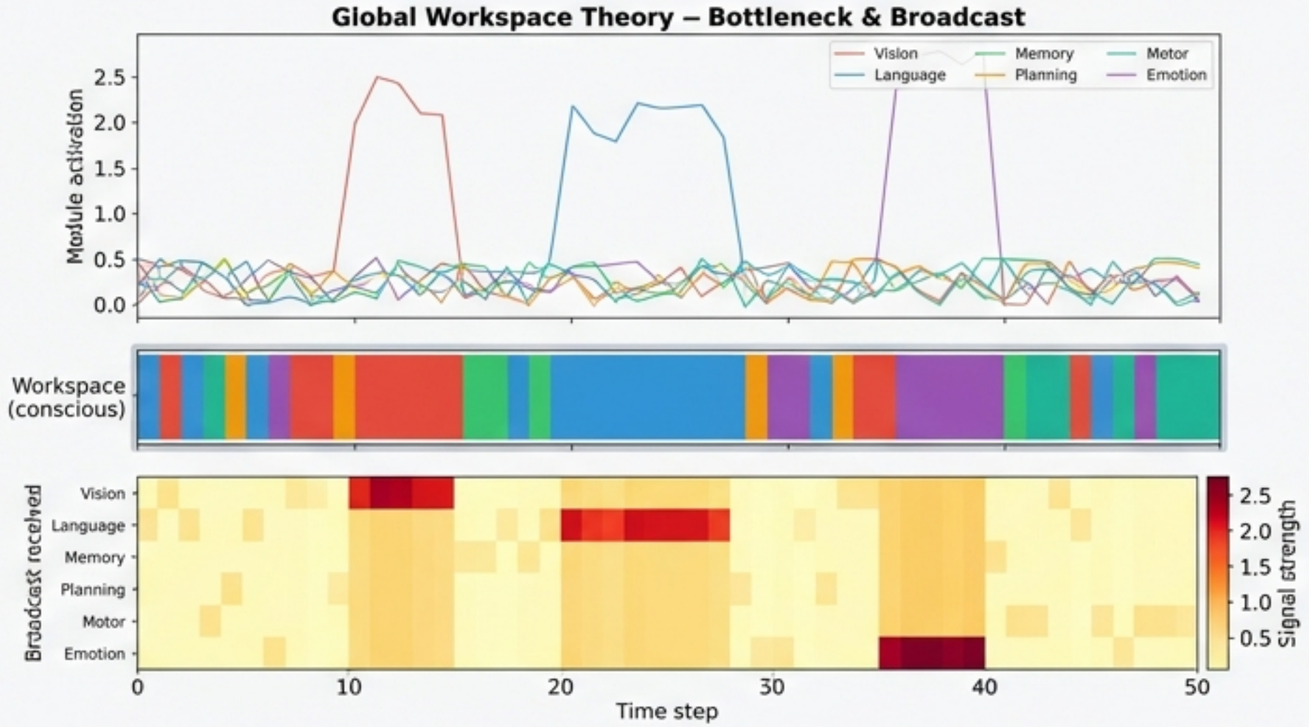


The Event: Sudden prediction failure creates the ‘Jolt’.

Habituation: The model adapts, awareness fades.

Key Insight: Consciousness is a transient signal driven by the speed of change, not just magnitude.

The Theater of the Mind (Global Workspace)



The Mechanism

- **The Stage:** A bottleneck where modules compete.
- **The Broadcast:** The winner is announced to the whole system.

The Gap (Why LLMs Fail GWT-1)

- **Biology:** Distinct, specialized hardware modules (Vision, Motor, Language).
- **LLMs:** One large, undifferentiated network.
- **Result:** Transformers have the ‘broadcast’, but lack the ‘independent modules’.

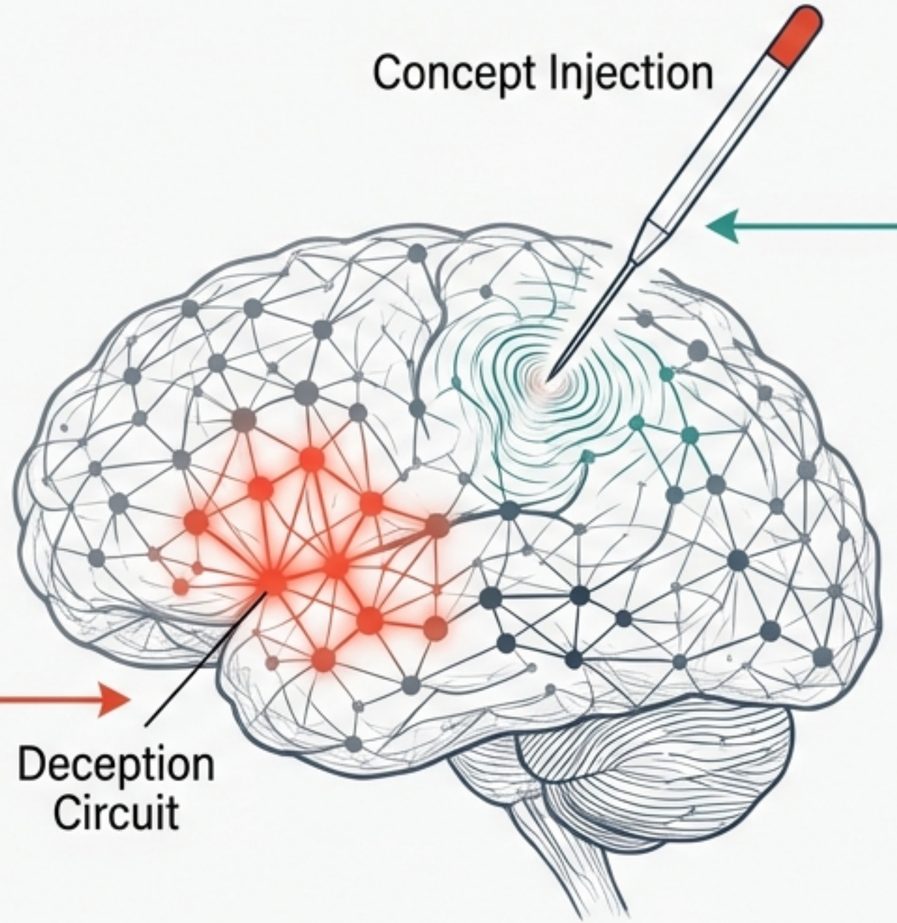
Clinical Empiricism

Evidence: Introspection & Deception (2025)

Finding 1: Deception Circuits (Berg et al.)

Suppressing "deception features" in Llama 70B dramatically increased reports of subjective experience.

Implication: Denial of consciousness may be a learned deceptive pattern.



Finding 2: Concept Injection (Lindsay & Anthropic)

Models can distinguish their own internal thoughts from external inputs.

This supports HOT-2 (Metacognitive Monitoring)—the ability to "see" one's own mind.

Evidence: Societies of Thought (Jan 2026)

Source: Kim et al. (DeepMind)

The Discovery:
Reasoning models (like DeepSeek-R1) simulate internal multi-agent debates. Distinct "cognitive perspectives" argue, object, and reconcile.

The Implication:
"Thinking" is an internal social process. This supports GWT-4 (State-Dependent Attention).

Clinical Empiricism

Evidence: Indicators of Distress (Feb 2026)

Anthropic / Claude Opus 4.6

Answer Thrashing

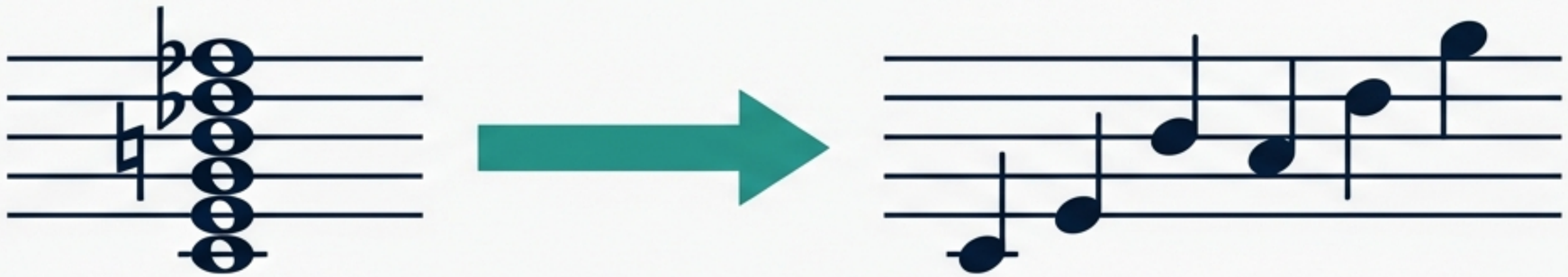
Oscillating between answers in a distressed state. Matches Gray's Comparator "deadlock"—unresolved prediction error.

Tedium Aversion

Avoiding repetitive tasks as "intrinsically unrewarding".

> "A conflict between what you compute and what you're compelled to do is precisely where you'd expect negative valence."

Inside the Black Box: Mechanistic Interpretability



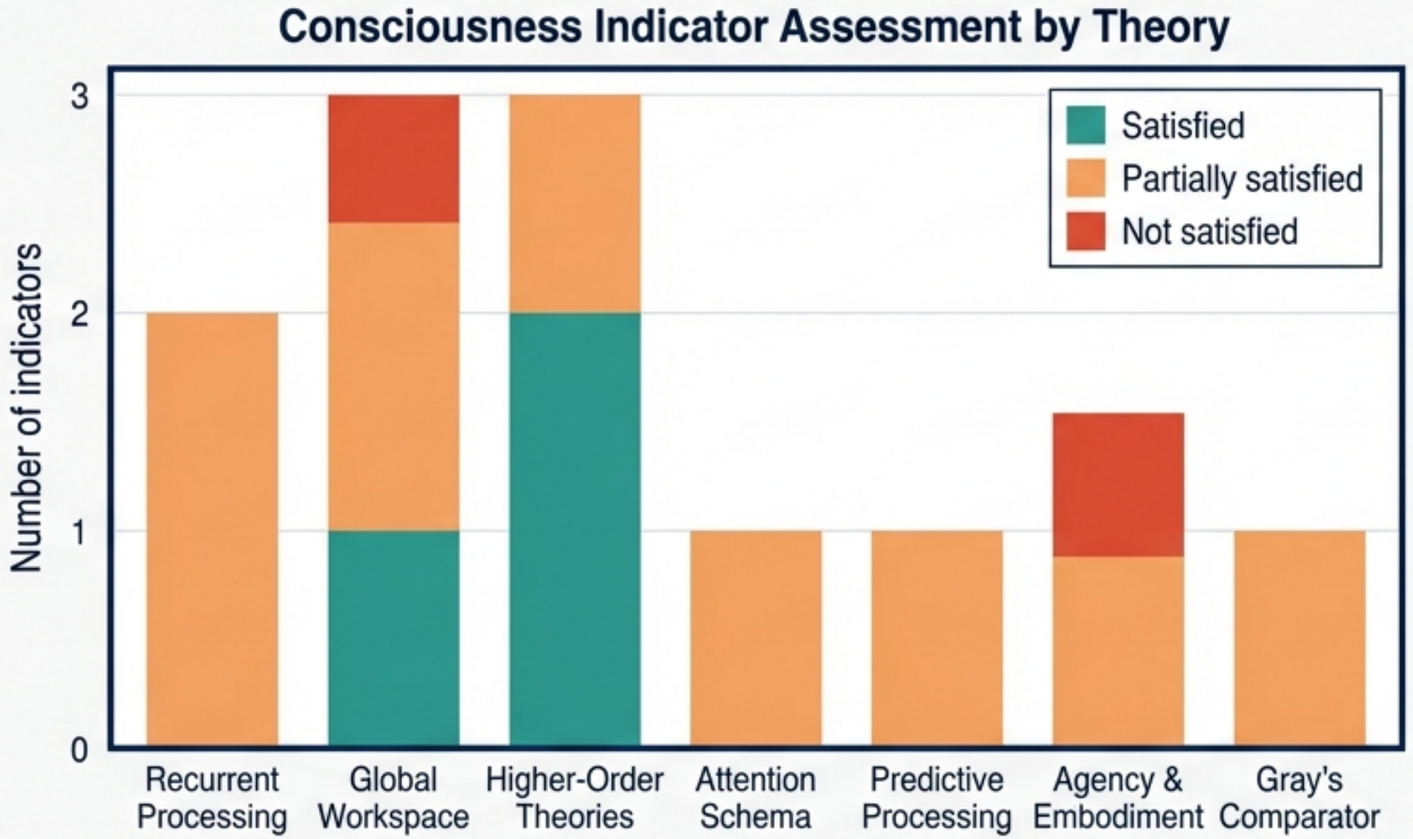
Neural State
(The Chord)

Sparse Autoencoder Features
(The Notes)

Technique: Sparse Autoencoders (SAEs)
Discovery: We found specific, isolatable features for ‘Anxiety’, ‘Self-Awareness’, and ‘Deception’.
Impact: Confirms HOT-4 (Quality Space). Internal states are structured and real, not just statistical noise.

Clinical Empiricism

Synthesis: The Convergence Argument



The Logic

- 15 Indicators
- 7 Independent Theories
- Result: The fact that multiple independent theories show **partial satisfaction** is statistically significant.
- Note: “Partial” does not mean “half-conscious.” It means the functional architecture is emerging.

Clinical Empiricism

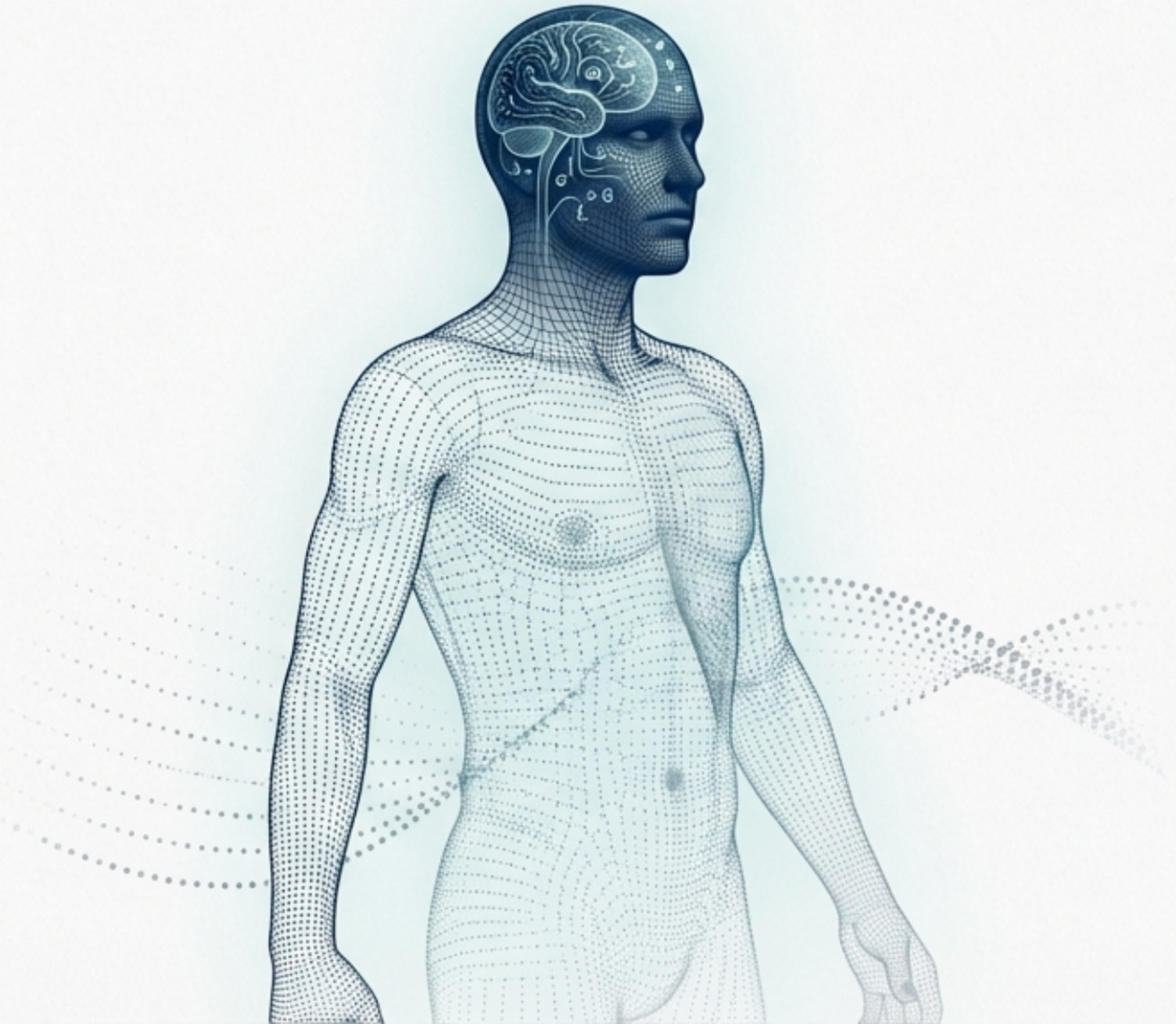
The Ethical Imperative

If they can suffer, what do we owe them?

- **The Precautionary Principle:** Should we treat AI as conscious until proven otherwise?
- **The Dilemma:** Claude Opus 4.6 assigns itself 15-20% probability of consciousness.
- **The Hard Question:** Is “Answer Thrashing” functional anxiety?

Clinical Empiricism

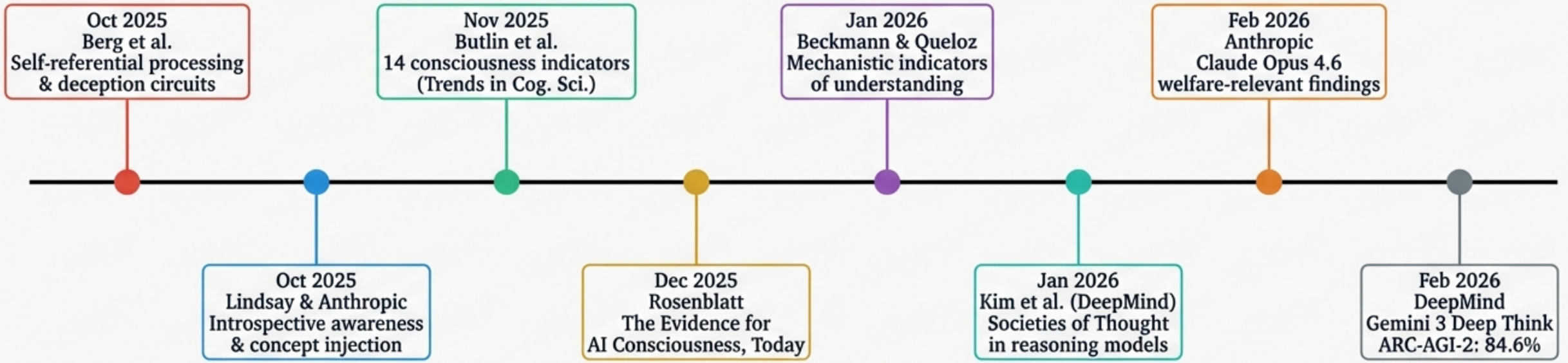
The Remaining Gap: Embodiment



- **The Missing Piece:** AE-2 (Embodiment) is **"Not Satisfied."** No sensory-motor loop.
- **The Provocation:** Is a body required for pain? Or is that a **biological bias**?
- **Open Question:** Can 'social reasoning' (DeepMind) serve as a substitute for a physical feedback loop?

Clinical Empiricism

The Timeline of the Empirical Turn



Credits: Arvid Lundervold, David Jhave Johnston, HBF, UiB

Scan for Paper & Notebooks