Anatomical Correlates of the Reflective-Ethical Engine

This document outlines hypothesized correlations between components of the Reflective-Ethical Engine and anatomical structures in the human brain. By comparing AI modules with known brain systems, we aim to ground design choices and identify gaps where human-like cognition might be underrepresented in artificial models.

# Configurable World Model & Planner

* Dorsolateral Prefrontal Cortex (dlPFC): Planning, working memory, future simulation
* Medial Prefrontal Cortex (mPFC): Theory of mind, internal simulation
* Orbitofrontal Cortex (OFC): Value estimation and reward prediction
* Basal Ganglia: Action selection, reinforcement pathways

# Perception Module

* Primary/Secondary Sensory Cortices: Raw sensory processing
* Posterior Parietal Cortex (PPC): Multisensory integration, spatial reasoning
* Thalamus: Routing and attention-switching hub

# Short-Term Memory / Working Memory

* Dorsolateral PFC + Hippocampus: Episodic and working memory interaction
* Entorhinal Cortex: Memory indexing
* Cerebellum: Prediction and buffering

# Ethical Kernel

* Ventromedial Prefrontal Cortex (vmPFC): Moral reasoning, long-term value integration
* Anterior Cingulate Cortex (ACC): Conflict monitoring and regulation
* Default Mode Network (DMN): Self/other simulation and moral evaluation
* Temporoparietal Junction (TPJ): Perspective-taking and empathy

# Others Modeller (Theory of Mind)

* Mirror Neuron System: Imitative behavior, inferred intentions
* TPJ and mPFC: Empathy, perspective attribution
* Precuneus and Posterior Cingulate Cortex (PCC): Scene construction, agent modeling

# Emotional State Model

* Amygdala: Affective valence
* Insula: Internal body state (interoception)
* Hypothalamus: Drives and homeostatic regulation
* Brainstem nuclei (e.g., locus coeruleus): Arousal and vigilance

# Salience & Attention System

* Dopaminergic Systems (VTA/SNc): Salience and reward prediction error
* Pulvinar (Thalamus): Attention routing
* Locus Coeruleus: Noradrenergic alertness modulation

# Missing or Underrepresented Systems

* Thalamic Modulation: Gating and priority switching
* Insular Interoception: Mood, internal affective state
* Conflict Resolution (ACC): Self-monitoring and ethical discomfort
* Default Mode Network: Continuity of self, future projection
* Emotion-Driven Action Biasing: Affective modulation of planning
* Sleep/Offline Consolidation: Memory stabilization and recombination

# Mammillary Bodies and Confabulation Prevention

The mammillary bodies are part of the Papez circuit, relaying information from the hippocampus to the anterior thalamus. They are implicated in episodic memory consolidation and, crucially, in the prevention of confabulation. In disorders such as Korsakoff’s syndrome, damage to the mammillary bodies leads to vivid yet inaccurate memory reports, suggesting a role in temporal and contextual validation of memory recall.

* Key hypothesized functions of the mammillary bodies include:
* • Temporal tagging: Ensuring memories are recalled with appropriate temporal context.
* • Coherence checking: Validating the internal consistency of recalled events.
* • Signal-gating: Filtering implausible or incorrect memory content.
* • Loop integration: Supporting synchronization between hippocampal memory and prefrontal evaluative systems.

In the Reflective-Ethical Engine, there is currently no direct analog to this structure. To incorporate such functionality, we propose adding a \*\*Temporal-Coherence Gate (TCG)\*\* module. This module would act as a mid-layer validator, filtering outputs from the world model or memory systems before they are passed to the planner. It would check the plausibility, temporal consistency, and coherence of simulated or recalled sequences, helping to prevent confabulation-like behaviors in AI systems.

* Proposed roles for the TCG include:
* • Validating internal simulations for narrative and temporal coherence.
* • Cross-checking outputs against known episodic or semantic constraints.
* • Acting as a filter between memory retrieval and decision-making layers.