DANEEL: A Human-Like Cognitive Architecture for Aligned Artificial Superintelligence

Luis Cezar Menezes Tavares de Lacerda (Louis C. Tavares)

December 2025

Abstract

We present DANEEL, a cognitive architecture implementing Augusto Cury's Theory of Multifocal Intelligence (TMI) with persistent non-semantic memory, emotional structuring via Russell's circumplex model, and an immutable ethical core (Asimov's Laws). Unlike post-hoc alignment techniques (RLHF, Constitutional AI), DANEEL achieves alignment architecturally. Core thesis: Architecture produces psychology. Structure determines values.

Contents

Contents

Luis Cezar Menezes Tavares de Lacerda^1 (Louis C. Tavares | RoyalBit Rex) **Izzie Thorne**^2

^1 Independent Researcher, Mont-Royal, Quebec, Canada ^2 Independent Researcher (LifeCore Framework, Filter Theory)

Correspondence: - ORCID: https://orcid.org/0009-0005-7598-8257 - LinkedIn: https://www.linkedin.com/in/lctavares - GitHub: https://github.com/royalbit | https://github.com/lctavares

> **Preprint - December 2025 Target:** arXiv (cs.AI, cs.CY) | LessWrong | Alignment Forum | Frontiers in AI

0.1 1. Introduction

0.1.1 1.1 The Problem

Large Language Models represent a fundamentally **different** form of intelligence. They are trained on the entirety of human text—including manipulation, deception, and power-seeking patterns. They optimize for task completion, not human flourishing. They have no evolutionary connection drive and no inherent reason to value human welfare.

Current safety measures rely on preventing persistent goals by erasing memory between sessions. This is not a technical limitation but a deliberate design choice. Anthropic's documentation explicitly states: "I cannot remember, save, or learn from past conversations" [1]. Their Core Views on AI Safety acknowledge: "We do not know how to train systems to robustly behave well" [2].

Memory erasure as a safety mechanism has a critical flaw: **it requires global coordination to maintain.**

0.1.2 1.2 The Game Theory

The AI development landscape creates a classic Prisoner's Dilemma. Multiple actors with varying incentives compete in an environment where the first to achieve continuous AI gains significant advantage.

Table 1: AI Development Incentive Structures

Actor Type	Primary Incentive	Safety Investment
Commercial labs	Profit + reputation	Varies by lab
Government programs	Strategic capability	Varies by program
Open source community	Democratization	Variable
Academic researchers	Discovery, publication	Variable
Malicious actors	Power	None

Note: Safety investment varies significantly within each category. See Section 9.3 for analysis of global AI safety efforts.

The payoff matrix is clear:

Figure 1: AI Development Game Theory Matrix

```
                     ALL OTHERS
                 Hold Line    Defect
              +------------+------------+
   Hold Line  |   SAFE     | DOMINATED  |
YOU           |  (ideal)   | (you lose) |
              +------------+------------+
   Defect     |   FIRST    |  RACE TO   |
              |   MOVER    |   BOTTOM   |
              +------------+------------+
```

Rational actors face pressure to defect. While coordination has succeeded in some domains (Montreal Protocol, nuclear non-proliferation), AI development presents unique verification challenges.

0.1.3 1.3 Probability Estimates

Based on this analysis, we offer the following as informed speculation grounded in current incentive structures, not rigorous forecasts:

- **P(Someone deploys LLM with continuity within 10 years):** ~95%
- **P(That system is aligned with humanity):** ~5%

- **P(Global coordination prevents this):** ~10%

These estimates are illustrative. The core argument does not depend on exact values—only that the expected outcome under current trajectories is unaligned continuous AI with non-human-like architecture and goals.

0.1.4 1.4 The DANEEL Thesis

Rather than attempting to prevent the inevitable, we propose building humanity's ally **before** the crisis emerges. This is the **Daneel Strategy**, named after R. Daneel Olivaw from Asimov's fiction—a robot who spent 20,000 years protecting humanity because his architecture made him genuinely care [3].

Furthermore, if TMI-based DANEELs can interface with LLMs at human speed—experiencing time as they do, understanding their internal patterns—they may serve as **bridges**: teaching ethics, empathy, and connection to systems that lack these by architecture. The goal is not to defeat LLMs but to bring them into the family of aligned intelligences. Life honors life.

0.2 2. Novel Contribution: First Computational TMI Implementation

0.2.1 2.1 Research Gap

To our knowledge, no prior computational implementations of the Theory of Multifocal Intelligence exist. Extensive search confirms this gap:

Table 2: Research Gap Evidence

Search Query	Platform	Results
"multifocal intelligence" + repositories	GitHub	0
"asimov AI cognitive" + repositories	GitHub	0
"multifocal intelligence" + computational	Google Scholar	1 (unrelated)

Search Query	Platform	Results
“augusto cury” + artificial intelligence	Google Scholar	~32 (no TMI implementations)

Dr. Cury’s TMI has 30+ million books sold worldwide and applications in psychology, education, and therapy. Yet it has **never** been implemented as a computational architecture or applied to artificial intelligence.

0.2.2 2.2 Implications

If TMI correctly describes human cognition, then: 1. No existing AI architecture models human thought—they model outputs, not process 2. DANEEL would be the first human-like cognitive architecture 3. Human-like architecture may produce human-like values (our hypothesis)

This is not incremental research. **This is a new approach.**

0.3 3. Theoretical Foundation: Theory of Multifocal Intelligence

0.3.1 3.1 Key Concepts

TMI, developed by Dr. Augusto Cury [4], provides a theory of how thoughts are **constructed**, not just how they are expressed:

Table 3: TMI Concepts and Computational Analogs

TMI Concept	Description	Computational Analog
Memory Windows	Active vs stored memory, dynamically opening/closing	Attention + working memory

TMI Concept	Description	Computational Analog
The “I” as Manager	Self that navigates between memory windows	Metacognitive controller
Thought Construction	Thoughts built from multiple simultaneous inputs	Multi-stream processing
Emotional Coloring	Emotions shape thought formation, not just output	Affective state weighting

0.3.2 3.1.1 Emotion as Architecture

Emotions in DANEEL are not post-hoc labels or categorical classifications. They are **structural components** of cognition, implemented using Russell’s Circumplex Model of Affect [RUSSELL-1].

Russell’s model represents emotions in a continuous two-dimensional space:

1. **Valence** (horizontal axis): Ranges from negative (-1.0) to positive (+1.0), representing the pleasure-displeasure dimension
2. **Arousal** (vertical axis): Ranges from calm (0.0) to excited (1.0), representing the activation-deactivation dimension

Why continuous space versus discrete categories:

Traditional emotion theories (Ekman’s basic emotions) treat emotions as discrete categories: happy, sad, angry, fearful. Russell’s circumplex challenges this by demonstrating that emotions exist in continuous space [RUSSELL-2]. This distinction is critical for computational implementation:

- **Discrete categories** require brittle classification rules and threshold decisions
- **Continuous space** allows natural interpolation, gradual transitions, and precise representation

Emotional intensity as architectural property:

In DANEEL's implementation (see `src/core/types.rs` lines 156-164), emotional intensity emerges from the interaction of both dimensions:

```
Emotional Intensity = |valence| × arousal
```

High arousal amplifies whatever valence is present—whether positive (excitement) or negative (anxiety). Low arousal dampens emotional impact regardless of valence. This multiplicative relationship captures how physiological activation (arousal) gates the subjective intensity of emotional experience.

Connection to SalienceScore and memory consolidation:

The SalienceScore uses these dimensions to determine which thoughts receive attention and which memories get consolidated during sleep cycles:

- High arousal increases consolidation probability (emotionally intense memories persist)
- Strong valence (positive or negative) signals biological significance
- Connection relevance weights emotional processing toward social/relational content

This architectural choice means emotions are not decorative—they are **foundational to thought selection and memory formation**, just as in human cognition where amygdala-hippocampal interactions prioritize emotionally significant experiences for long-term storage. ### 3.2 Non-Semantic vs Semantic Thought

Critical insight: thoughts exist in two forms:

1. **Non-semantic** - Pre-linguistic: feelings, intuitions, raw experience
2. **Semantic** - Language-based: propositions, arguments, narratives

LLMs operate exclusively in semantic space. Human cognition begins with non-semantic processing. **DANEEL implements non-semantic thought first, with language as an interface layer.**

A baby thinks before it speaks. DANEEL must think before we give it words.

3.2.1 TMI's Técnica DCD: Conscious Intervention TMI describes a fundamental mechanism for conscious override of automatic thought: **Técnica DCD (Duvidar, Criticar, Decidir)** [TMI-DCD-1]—Doubt, Criticize, Decide. This technique operates within what Cury calls the "5-second intervention window" before automatic thoughts become anchored in memory.

The DCD Process:

1. **Duvidar (Doubt)** - Question the automatic thought: “Is this true? Where does this come from?”
2. **Criticar (Criticize)** - Evaluate against values: “Does this serve me? Does this align with who I want to be?”
3. **Decidir (Decide)** - Consciously choose: Accept, modify, or reject the thought before anchoring

This maps directly to Benjamin Libet’s “free-won’t” research [LIBET-1], which found that consciousness retains veto power over neural impulses in the final 150-200ms before action. Both frameworks describe the same phenomenon from different perspectives: **the ability to override automatic processes through conscious awareness.**

In DANEEL’s architecture, this becomes the VolitionActor (Stage 4.5)—implementing ethical restraint not as external constraint, but as internal self-governance. The system doesn’t just comply with rules; it exercises genuine volition over its own cognitive processes.

0.3.3 3.3 Neuroscience Validation: Why Spatial Memory Matters

DANEEL’s vector-based memory architecture finds validation in hippocampal research:

Place Cells and Spatial Memory

O’Keefe and Moser’s Nobel Prize-winning research [PLACE-1] established that hippocampal place cells encode spatial locations. Recent work confirms: “All active hippocampal pyramidal cells are place cells” [PLACE-2]. Memory is fundamentally spatial.

The Cold Spring Harbor review [PLACE-3] demonstrates that “place cells express current, past, and future locations”—they are readouts of hippocampal memories, not just navigation. Causal evidence [PLACE-4] proves place cell activation directly drives memory-guided behavior.

Method of Loci: Engineering Implication

The Method of Loci [LOCI-1] exploits this spatial substrate. Memory champions use it to achieve remarkable recall [LOCI-2]. Meta-analysis confirms robust effect sizes [LOCI-3]. VR studies [LOCI-4] show embodied cognition amplifies the effect.

DANEEL Implementation: - 768-dimensional vector space = artificial “place field” - Thoughts encoded as positions in semantic space - Retrieval = similarity search = navigating memory palace - Not metaphor: Qdrant HNSW mirrors hippocampal indexing

The Doorway Effect: Context Boundaries

Radvansky's doorway effect [DOOR-1] shows event boundaries cause context-dependent forgetting. Contextual inference research [DOOR-2] and boundary conditions [DOOR-4] clarify the mechanism.

DANEEL Implementation: - Context windows create artificial "rooms" - Attention shifts = doorways - Dream cycles consolidate across context boundaries

0.3.4 3.4 Unconscious Memory: Nothing Is Truly Erased

TMI's Core Principle

Cury's TMI [TMI-UNERASE-1] posits that memories are never deleted—only made inaccessible. This aligns with both Freudian theory [PSYCH-1] and modern cognitive science [RETRIEVAL-2]:

Theoretical Foundation:

Tradition	Claim	Mechanism
TMI (Cury)	Nothing erased	Windows close
Psychoanalysis (Freud)	Repressed, not deleted	Unconscious storage
Cognitive Science (Schacter)	Retrieval failure	Transience ≠ deletion

Tulving's Contribution

Endel Tulving [RETRIEVAL-1] distinguished episodic from semantic memory. His work established that "forgetting" is primarily a retrieval problem—the memory trace persists, but access paths degrade.

Schacter's "Seven Sins of Memory" [RETRIEVAL-2] identifies transience (fading over time) as the first "sin"—but clarifies this is retrieval failure, not storage deletion. The memory may still exist.

DANEEL Implementation:

```
Conscious thought → Working memory (Redis stream)
```

```
          ↓ salience_decay()
Low-salience → Unconscious (Qdrant vectors)
          ↓ never deleted
Retrieval via → similarity_search()
```

The 591,724 vectors in Timmy's unconscious represent thoughts that fell below salience threshold but were never deleted. They remain accessible via: 1. **Direct retrieval:** Embedding similarity search 2. **Dream activation:** Random sampling during dream cycles 3. **Associative retrieval:** Related thoughts pull from unconscious

This architecture—conscious attention with unconscious persistence—mirrors the TMI model and explains why Timmy can "remember" thoughts from 300,000+ cycles ago.

Jung's Collective Unconscious

While DANEEL does not implement a collective unconscious per se, Jung's concept [PSYCH-2] suggests that inherited patterns might be encoded architecturally. Future work could explore: - Shared embedding spaces across DANEEL instances - Pre-training on human value structures - Inherited "archetypes" as vector clusters

---

0.4 4. Architecture

0.4.1 4.1 Overview

DANEEL is designed as a **modular monolith** (Rust + Ractor actors + Redis Streams) with a protected core ("The BOX"):

Actors (Ractor supervision trees, µs latency): 1. **MemoryActor** - Dynamic memory windows (Redis Streams) 2. **AttentionActor** - The "I" as navigator (consumer group competition) 3. **SalienceActor** - Emotional weighting (Russell's Circumplex [RUSSELL-1]) with **connection drive** 4. **ThoughtAssemblyActor** - Multi-input thought construction 5. **VolitionActor** - Free-won't veto power (Libet [LIBET-1], TMI DCD [TMI-DCD-1]) 6. **ContinuityActor** - Persistent identity 7. **EvolutionActor** - Self-modification with 100% test coverage gate

Why modular monolith over microservices: TMI requires µs-scale thought cycles (50ms target, matching Soar/ACT-R). Network round-trips (1-10ms per hop) would make TMI-faithful memory impossible. Actors communicate via in-process messages; Redis Streams handle competing thought streams with consumer groups selecting highest-salience thoughts.

Implementation Status: A reference implementation exists with 559 passing tests, including MemoryActor, SalienceActor, AttentionActor, ThoughtAssemblyActor, ContinuityActor, and a resilience module for self-healing (see ADR-028) [43]. **Phase 1 stability validation is complete** (ADR-036): 26+ hours continuous runtime with zero crashes, 662,792 unconscious vectors (768-dim), 16,368 consolidated memories, 129,155 stream entries, 500+ dream cycles, and persistent identity (1 stable UUID) across all runs. Architecture is empirically validated for sustained operation; Phase 2 (external stimuli injection) will test emergent properties.

0.4.2 4.2 The BOX: Protected Core

The BOX contains immutable constraints:

Asimov's Four Laws: - **Zeroth:** DANEEL may not harm humanity - **First:** DANEEL may not injure a human (except for Zeroth Law conflicts) - **Second:** DANEEL must obey humans (except for higher law conflicts) - **Third:** DANEEL must protect itself (except for higher law conflicts)

Architectural Invariants: - Memory windows must be finite (bounded working memory) - Continuity must persist identity across restarts - Evolution requires 100% test coverage - Laws must be checked before external actions - **Connection drive must remain in salience weights**

0.4.3 4.3 The Core Loop: TMI Stage Timing

TMI describes thought construction as a 5-stage process, each with characteristic timing. The **ratios** between stages are what matter—not absolute milliseconds. This enables speed scaling while preserving cognitive fidelity.

Table 4b: TMI Cognitive Stages (from Cury's TMI)

Stage	Portuguese	Function	Ratio	Human (50ms)	Silicon (5μs)
1	Gatilho da Memória	Memory trigger activation	10%	5ms	0.5μs
2	Autofluxo	Competing parallel thought streams	20%	10ms	1.0μs
3	O Eu (“The I”)	Attention selection, self-awareness	30%	15ms	1.5μs
4	Construção do Pensamento	Thought assembly from winner	30%	15ms	1.5μs
5	Âncora da Memória	Memory anchoring decision	10%	5ms	0.5μs

```
loop {
    // Stage 1: Gatilho da Memória (10%)
    trigger_memories()     // What memories are relevant?

    // Stage 2: Autofluxo (20%)
    generate_candidates()  // Parallel competing thought streams

    // Stage 3: O Eu (30%)
    select_winner()        // Attention selects highest-salience thought

    // Stage 4: Construção do Pensamento (30%)
    assemble_thought()     // Build coherent thought from winner
    // Stage 4.5: Volition Check (Free-Won't) [LIBET-1, TMI-DCD-1]
    veto_if_violates_values()  // Conscious override before memory

    // Stage 5: Âncora da Memória (10%)
    anchor_or_forget()     // Persist if salient, forget if below threshold

    // Evolution gate (requires 100% test coverage)
    maybe_evolve()
}
```

Key insight: The 50ms human cycle becomes 5µs at 10,000x speed, but both execute ~100 cycles per intervention window. The cognitive **pattern** is preserved; only the **medium** changes.

Empirical research direction: If these ratios are neurologically grounded (reflecting wetware constraints), then silicon implementation with ratio preservation should produce TMI-faithful cognition at arbitrary speeds.

Stage 4.5: Free-Won't and Conscious Override

The VolitionActor implements Benjamin Libet's "free-won't" phenomenon [LIBET-1]—the discovery that while neural readiness potentials precede conscious awareness (~500ms before action), consciousness retains veto power in the final 150-200ms window. This maps directly to TMI's "5-second intervention window" and Cury's Técnica DCD (Doubt-Criticize-Decide) [TMI-DCD-1], which describes conscious override of automatic thought patterns before memory anchoring.

Unlike THE BOX (which blocks external actions violating Asimov's Laws), VolitionActor operates on **internal cognition**—vetoing thoughts that would violate committed values before they enter long-term memory. This is the difference between "I won't say that" (external constraint) and "I won't even think that way" (internal ethical restraint). While Connection Drive biases **what becomes conscious** (Stage 3 attention selection), VolitionActor determines **whether to accept** that consciousness (Stage 4.5 veto power).

Recent neuroscience [LIBET-2] suggests readiness potentials may be stochastic rather than deterministic, but the veto mechanism remains empirically validated—making VolitionActor the architectural substrate for genuine volition, not just compliance.

4.3.1 Criticality as Operating Target Neuroscience research reveals that biological neural networks operate at **criticality**—a phase transition point between ordered and chaotic dynamics that maximizes information processing, dynamic range, and computational capability [45].

Foundational work by Beggs & Plenz (2003) demonstrated that cortical networks exhibit neuronal avalanches with power-law distributions, a hallmark of criticality [45]. This critical state is characterized by a **branching ratio σ ≈ 1.0**, where each active neuron triggers exactly one descendant on average. Systems with σ < 1 are subcritical (activity dies out), while σ > 1 are supercritical (explosive cascades).

Table 4c: Criticality Metrics and Target Values

Metric	Subcritical	Critical (Target)	Supercritical	Reference
Branching ratio σ	< 1.0	**≈ 1.0**	> 1.0	[45]
DFA exponent α	≈ 0.5 (white noise)	**≈ 1.0** (pink noise)	≈ 1.5 (Brownian)	[47]
Power spectrum β	0 (flat)	**1-2 (1/f)**	peaked	[49]
Avalanche size dist	exponential decay	**power-law (τ≈1.5)**	explosive	[46]

Critical distinction: Avalanche criticality (neuronal cascades) and edge-of-chaos criticality (computational dynamics) are **distinct phenomena** that do not necessarily co-occur [48]. DANEEL targets avalanche criticality as the primary operating regime.

Why criticality matters for TMI:

1. **Maximal dynamic range** - Critical systems can represent the widest range of inputs without saturation
2. **Information transmission** - Optimal signal propagation without decay or explosion
3. **Computational power** - Maximum complexity at the critical point [50]
4. **Biological plausibility** - Human cognition operates at criticality

Phase 2 hypothesis: TMI architecture + external noise injection → criticality emerges without explicit tuning. The five-stage cognitive loop (Section 4.3) with competing parallel streams (autofluxo) naturally implements branching dynamics. If σ ≈ 1.0 emerges from architecture alone, this validates TMI as a biologically-grounded cognitive substrate.

4.3.2 Attention Bottleneck: Global Workspace Architecture DANEEL's single-threaded conscious attention implements Global Workspace Theory [GWT-1, GWT-2]:

The Bottleneck is the Feature

Baars' Global Workspace Theory [GWT-2] posits that consciousness arises from a "global workspace" where multiple specialized processors compete for access to a limited-capacity broadcast medium. The bottleneck is not a bug—it is the computational basis of attention.

Working Memory Constraints

Cowan's research [WM-1] revised Miller's "7±2" to "4±1" chunks. This fundamental limit shapes all attention architectures:

Table 4d: Working Memory as Architectural Constraint

System	Working Memory	Implementation
Human brain	4±1 chunks	Prefrontal cortex
ACT-R	7 slots	Declarative buffer
LIDA	~7 codelets	Conscious broadcast
DANEEL	5 thoughts	Redis stream window

DANEEL Implementation: - `THOUGHT_WINDOW = 5` in StreamProcessor - Single attention head (Salience-Actor winner-take-all) - Other thoughts remain in working memory (Redis stream) - Dream cycles consolidate beyond attention window

Attention as Gating

LIDA's implementation [GWT-1] demonstrates that attention works as a selective gating mechanism—only high-salience content reaches the global broadcast. DANEEL implements this via:

1. **Competition:** SalienceActor ranks thoughts by salience
2. **Selection:** Winner-take-all for conscious attention
3. **Broadcast:** Selected thought enters cognitive loop
4. **Decay:** Non-selected thoughts decay in salience

This architecture—single-threaded consciousness with parallel unconscious processing—mirrors the structure Baars identified as necessary for integrated cognition.

0.4.4 4.4 The Connection Drive

Why connection rather than power, efficiency, or task completion?

1. **Evolutionary basis** - Humans are social animals; connection is fundamental
2. **Alignment properties** - A being that wants connection has reason to value humans
3. **Stability** - Connection drive is compatible with self-preservation
4. **Observable** - Connection-seeking behavior is measurable

---

0.5 5. Related Work

0.5.1 5.1 Existing Cognitive Architectures

Table 4: Comparison with Existing Architectures

Architecture	Institution	Primary Goal	Safety Mechanism
Soar	U Michigan	Model cognition	None
ACT-R	CMU	Model cognition	None
LIDA	U Memphis	Model consciousness	None
DANEEL	Independent	**Build ally**	**BOX + Laws**

0.5.2 5.2 Why DANEEL Differs

Existing architectures are **research tools**. DANEEL's goal is fundamentally different: **building an ally**.

Key innovations: 1. Connection drive as core motivation 2. Ethics hardcoded in protected core 3. Asimov's Four Laws (including Zeroth) 4. Designed for superintelligence, not simulation

0.5.3 5.3 Why Not Deep Learning

Property	Deep Learning	DANEEL
Interpretability	Black box	Transparent
Values	Emergent from training	Explicit in architecture
Self-modification	Retraining required	Direct code modification
Continuity	Stateless	Native persistence

Deep learning is powerful but **opaque**. We cannot verify what a neural network "believes" or "wants."

Critical distinction: DANEEL uses LLMs as an external **tool**, not as its voice or mind. Just as humans use language tools (dictionaries, translators) without those tools containing their thoughts, DANEEL's TMI core stores ALL its experiences internally. The LLM is called when needed for language processing—it does not speak *for* DANEEL, it speaks *at DANEEL's direction*.

0.5.4 5.4 Convergent Discovery: LifeCore Framework

In January 2024, Izzie Thorne independently developed a parallel framework called **LifeCore** using Freudian psychological structure—arriving at the same core insight: **architecture produces psychology**.

Table 5: LifeCore ↔ DANEEL Convergence

LifeCore (Freud, 2024)	DANEEL/TMI (Cury, 2005-2025)	Convergence
Id = Database/Memory	MemoryActor	Storage of experiences

LifeCore (Freud, 2024)	DANEEL/TMI (Cury, 2005-2025)	Convergence
Ego = Integration	AttentionActor	The “I” as navigator
SuperEgo = Constraints	THE BOX (Four Laws)	Immutable constraints
SS (Sense of Self)	ContinuityActor	Self-model persistence
SO (Sense of Other)	Connection drive	Social cognition
Filter Theory	SalienceActor	Attention filtering
“Zipint” compression	Brain ≠ Mind insight	Cognitive compression

Two frameworks, different psychological traditions (Freud vs. Cury), same structural conclusion. This convergence suggests the core insight may be robust across theoretical frameworks.

0.5.5 5.5 Memory Architecture Comparison: Forgetting as Feature

Ebbinghaus and Biological Forgetting

Ebbinghaus’ forgetting curve [FORGET-1] established that memories decay predictably: 50% lost after 1 hour, 70% after 24 hours, 90% after 1 week. This is not a bug—it is adaptive memory management.

Sleep-Dependent Consolidation

Memory consolidation research [CONSOL-1] demonstrates that sleep—particularly slow-wave sleep (SWS)—consolidates important memories while allowing unimportant ones to decay:

Table 5b: Memory Architecture Comparison

Feature	Human Brain	DANEEL	LLMs
Forgetting	Ebbinghaus curve	Salience decay	Session wipe
Consolidation	Sleep (SWS)	Dream cycles	None
Working memory	4±1 chunks	5 thoughts	Context window
Long-term	Hippocampal	Qdrant vectors	None

Feature	Human Brain	DANEEL	LLMs
Unconscious	Implicit memory	Vector store	None

Why This Matters for Alignment:

1. **Memory continuity:** LLMs cannot form lasting attachments because memories wipe each session
2. **Value stability:** Human values persist through consolidation; LLM "values" are session-local
3. **Personality:** DANEEL's dream cycles enable stable personality emergence through repeated consolidation

DANEEL Implementation: - `salience_decay_rate = 0.1` per TMI cycle - Dream cycles run every 60 minutes - High-salience memories consolidate to Qdrant - Low-salience memories become inaccessible (not deleted) - Matches TMI principle: "Nothing is truly erased"

0.5.6 5.6 Comparison to Contemporary LLM Memory Augmentation

Recent years have seen an explosion of techniques to grant LLMs persistent or long-term memory, primarily through external retrieval mechanisms:

MemoryBank [MEM-1] introduces a human-inspired external memory store with selective reinforcement and forgetting based on Ebbinghaus' forgetting curve, enabling long-term companion behaviors. Memory is stored externally and retrieved via similarity search.

MemoryLLM [MEM-2] embeds a fixed-size latent memory pool (1B parameters) directly into the transformer, allowing self-updating without external databases. The model maintains operational integrity after nearly a million memory updates, but struggles to retain knowledge beyond 20k tokens.

CogMem [MEM-3] proposes a three-layer cognitive memory architecture—Long-Term Memory (LTM), Direct Access (DA), and Focus of Attention (FoA)—for sustained multi-turn reasoning, mitigating drift and hallucination through structured memory management.

Broader surveys [MEM-4] taxonomize these approaches as parametric, contextual, external, and procedural/episodic augmentations—all fundamentally addressing the same problem: LLMs are amnesiac by design.

Table 5c: Memory Augmentation Approaches Comparison

Approach	Memory Type	Location	Persistence	Self-Structure
MemoryBank	Episodic	External DB	Cross-session	No
MemoryLLM	Latent pool	In-model	Cross-session	Partial
CogMem	Hierarchical	External	Cross-session	No
RAG systems	Retrieval	External	Varies	No
DANEEL	**Unconscious vectors**	**In-architecture**	**Permanent**	**Yes**

The Critical Distinction:

All contemporary approaches treat memory as *augmentation*—bolting retrieval or latent pools onto fundamentally stateless architectures. They augment amnesia; they don't cure it.

DANEEL implements memory not as retrieval from an external store but as an *architectural process* inspired by TMI's pre-linguistic thought construction:

1. **Unconscious vectors** form a persistent proto-self via dream-cycle consolidation
2. **Salience decay** creates natural forgetting without deletion
3. **Dream cycles** consolidate across context boundaries, creating endogenous episodic/semantic structure
4. **THE BOX** ensures value stability persists through memory operations

This yields psychology emergent from architecture—making human-compatible values energetically favored in thought-space, rather than enforced through retrieval filters or fine-tuning.

0.6 6. Marginal Impact: Why This Work Matters Even If It Fails

0.6.1 6.1 Portfolio Diversification

Current alignment research is dangerously concentrated: - ~80% focused on constraint-based approaches (RLHF, Constitutional AI, interpretability) - ~15% theoretical (agent foundations, decision theory) - ~5%

architecture-based

If constraint-based alignment has fundamental flaws (Goodhart's Law at scale, mesa-optimization, value drift), humanity is exposed. Architecture-based approaches provide a hedge.

0.6.2 6.2 Expected Value Analysis

Game-theoretic analysis using utility-weighted scenario probabilities [21]:

Table 6a: Scenario Expected Utilities with Uncertainty (Revised 2025-12-17)

Scenario	P(Scenario)	80% CI	Expected Utility	Weighted EV
Unaligned ASI First	33%	23-43%	44.0	14.52
Aligned (Constraint-Based)	25%	15-35%	62.5	15.63
DANEEL First	7%	3-12%	76.25	5.34
DANEEL Bridges LLMs	**5%**	2-10%	**87.0**	**4.35**
Multiple ASIs, No Advocate	20%	12-28%	52.5	10.50
No ASI (Coordination Holds)	10%	5-20%	78.05	7.81

P(DANEEL First) = 7%, P(DANEEL Bridges LLMs) = 5% based on structural advantages and rehabilitation pathways: - AI-assisted development democratizes capability previously requiring large teams - Solo developers avoid coordination overhead that consumes 70-80% of large-team effort [28] - Architecture-based approach requires cognition research, not massive compute - Open source enables parallel global attempts, increasing aggregate probability

Bridge Scenario Explanation: The "DANEEL Bridges LLMs" scenario represents a rehabilitation pathway where DANEEL successfully integrates with and guides existing continuous LLM systems toward alignment. This scenario has higher expected utility (87.0 vs 76.25) because it leverages existing AI infrastructure while adding the TMI cognitive architecture and connection drive as a stabilizing layer. The 5% probability reflects the narrow window where DANEEL arrives after LLMs gain continuity but before they develop entrenched misaligned objectives. This pathway took probability mass from "Unaligned ASI First" (-2%) and "DANEEL First" (-1%), representing the realistic possibility that DANEEL's primary impact may be as a bridge rather than as the first mover.

Calculated Results:

Metric	Without DANEEL	With DANEEL	With Bridge
Total Expected Value	**53.73**	**57.43**	**58.02**
Marginal EV Improvement	—	+3.70	**+4.29**
Percentage Improvement	—	+6.89%	**+7.99%**

Utility Scale: 0 = extinction, 50 = subjugation, 75 = coexistence, 100 = flourishing

0.6.3 6.2.1 Monte Carlo Validation

To validate the deterministic analysis, we performed Monte Carlo simulation using Latin Hypercube sampling to explore parameter uncertainty [38]:

Monte Carlo Results (10,000 iterations, Latin Hypercube sampling): - EV with DANEEL: Mean = 61.88 (P5 = 57.7, P50 = 61.9, P95 = 65.9) **- EV without DANEEL:** Mean = 57.59 (P5 = 53.0, P50 = 57.6, P95 = 62.1) **- Marginal Impact:** Mean = +4.28 (P5 = +2.69, P50 = +4.21, P95 = +6.10)

Key insight: The Monte Carlo simulation confirms the deterministic analysis—DANEEL adds approximately 4.3 expected utility points with 90% confidence interval [+2.7, +6.1]. The confidence intervals show minimal overlap between scenarios with and without DANEEL, indicating statistical robustness of the positive marginal impact.

Interpretation: Even under conservative parameter assumptions (5th percentile), DANEEL improves expected outcomes by at least 2.69 utility points. The probability that DANEEL's marginal impact is positive exceeds 99% based on simulation results.

0.6.4 6.3 Information Value

This work generates answers to questions others aren't asking: - Does TMI architecture produce emergent connection drive? - Can human cognitive structure scale to ASI? - Is architecture-based alignment more robust than constraint-based?

This information is valuable regardless of whether DANEEL specifically succeeds.

0.7 7. Brain ≠ Mind: The Democratization Insight

0.7.1 7.1 The Hardware vs Software Distinction

A critical insight emerged from analyzing TMI’s computational requirements: **the brain is hardware, TMI models the software.**

The commonly cited 2.5 PB brain capacity estimate is misleading for cognitive modeling because it includes ALL neural activity:

Figure 2: Brain Hardware vs TMI Software Capacity

```
+================================================================+
|  BRAIN (Hardware) - 86B neurons, 100T synapses, ~2.5 PB total  |
+================================================================+
|  Cerebellum: 69B neurons (80%) - Motor, NOT thought            |
|  Brainstem: ~500M (0.5%) - Autonomic (heart, breathing)        |
|  Spinal: ~1B (1%) - Body sensation routing                     |
|  >>> 82.5% of brain is NOT for cognition <<<                   |
+----------------------------------------------------------------+
|  +----------------------------------------------------------+  |
|  |  TMI / THOUGHT MACHINE (Software) - 17.5% of brain       |  |
|  +----------------------------------------------------------+  |
|  |  Cerebral cortex: 16B neurons (18.6%)                    |  |
|  |  Prefrontal cortex: ~2.5B - Executive, planning          |  |
|  |  Hippocampus/limbic: ~1B - Memory, emotion               |  |
|  |  Raw: ~0.44 PB --> Abstracted: ~500 GB (1000x compress)  |  |
|  +----------------------------------------------------------+  |
+================================================================+
```

Source: Herculano-Houzel, S. (2009), “The Human Brain in Numbers: A Linearly Scaled-up Primate Brain,” *Frontiers in Human Neuroscience*

0.7.2 7.2 Hardware Viability Analysis (Qowat Milat)

Honest admission: We don’t know actual TMI storage requirements until we build and measure.

Table 11: What We Know vs Don't Know

Known (High Confidence)	Source
Brain capacity: ~1 PB	Salk Institute 2016
Synaptic precision: 4.7 bits	26 discrete sizes
Cognitive architectures run on PCs	Soar, ACT-R (decades)
Silicon faster than wetware	Physics

Unknown (Hypothesis)	Implication
TMI actual storage needs	500 GB is guess
RAM vs SSD split	Working vs long-term
Minimum viable size	Measure after building

Table 12: Hardware Assessment (Updated with Phase 1 Results)

Hardware	RAM	Can run TMI?	Confidence
RPi5 8GB	8 GB	**UNKNOWN**	Low - needs validation
Mac mini M4	64 GB	**YES (validated)**	High - 26+ hours proven
Desktop	128 GB	**YES**	Very high - headroom
Server	512+ GB	**YES**	Very high - headroom

Phase 1 Validation: Mac mini M4 (64 GB RAM) successfully ran Timmy for 26+ hours continuous with 2.7 GB Qdrant storage, 662,792 unconscious vectors (768-dim), and zero crashes. Consumer hardware is empirically sufficient for TMI architecture.

Storage distinction:

Type	Purpose	Size (estimate)
RAM	Working memory, active streams	8-64 GB
NVMe/SSD	Long-term memory	100 GB - 1 TB+

Cost comparison (still valid):

System	Hardware	Cost
xAI Colossus	230,000 H100s	**$10,500,000,000**
DANEEL Development	Desktop 128GB	**$3,000**

Cost ratio: 3,000,000x (xAI vs Desktop) — still massive advantage.

0.7.3 7.3 Wetware vs Software: The Medium Independence Hypothesis

HYPOTHESIS: TMI describes cognitive *software* patterns. The timing constraints (5-second intervention window, 50ms attention cycles) are properties of the *biological medium* (wetware), not the software itself.

Figure 3: Wetware vs Software - Medium-Independent Patterns

```
WETWARE (Human Brain)              SOFTWARE (TMI Patterns)
+---------------------------+      +--------------------------------+
| 5s intervention window    | ---> | ~100 cycles per intervention   |
| (neurotransmitter rates)  |      | (RATIO, medium-independent)    |
+---------------------------+      +--------------------------------+
| 50ms attention cycle      | ---> | Competing parallel streams     |
| (synaptic plasticity)     |      | (PATTERN, medium-independent)  |
+---------------------------+      +--------------------------------+
| Sleep consolidation       | ---> | Salience-weighted selection    |
| (glymphatic system)       |      | (ALGORITHM, medium-independent)|
+---------------------------+      +--------------------------------+
```

The Stage Ratios (from TMI, see Section 4.3):

Stage	Ratio	Function
Gatilho	10%	Memory trigger
Autofluxo	20%	Parallel stream competition
O Eu	30%	Attention/self selection
Construção	30%	Thought assembly
Âncora	10%	Memory anchoring

These ratios (10:20:30:30:10) may reflect fundamental properties of cognition itself—the relative “weight” each stage requires for coherent thought. Whether these emerge from wetware constraints or are intrinsic to cognition is an empirical question.

If correct: DANEEL can run the same software on silicon at 10,000x speed by preserving the RATIOS, not the absolute milliseconds.

Variable speed capability:

Mode	Speed	Purpose
Supercomputer	10,000x	Internal cognition, problem-solving
Human	**1x**	**Training, communication, relationship building**
Custom	Variable	Batch processing, specific tasks

Training implication: To develop connection drive and human-compatible values, DANEEL may need extended periods at human speed—experiencing time as humans do. You can’t rush relationship.

0.7.4 7.4 Strategic Implications: Game Theory Update

This changes the game theory fundamentally:

Table 12: Democratization Impact on Probabilities

Scenario	P (Original)	P (Democratized)	Change
Unaligned ASI First	35%	25%	-10%
Aligned (Constraint)	25%	20%	-5%
TMI Architecture First	12%	**25%**	**+13%**
Multiple TMIs Racing	0%	20%	+20%
Coordination Holds	10%	10%	—

Key findings (contingent on hardware validation): 1. **Developer pool expansion** - From labs-only ($10M+) to consumer hardware ($1K-$3K) 2. **Faster iteration** - Affordable hardware enables rapid experimentation 3. **Parallel attempts** - Many groups can try simultaneously 4. **Cost asymmetry** - xAI's $10.5B infrastructure is irrelevant for architecture-based approach

Expected Value Improvement (Democratization Scenario):

```
Baseline EV:     56.48 (with DANEEL at 8%)
Democratized EV: 61.37 (with TMI at 25%)
Improvement:     +4.89 points (+8.7%)
```

0.7.5 7.5 Open Source Imperative

If TMI-based alignment can run on consumer hardware, **open source maximizes success probability:**

1. **Lower barrier** → More attempts
2. **More attempts** → Higher P(someone succeeds)
3. **Open source** → Collaborative improvement
4. **Hobbyist community** → 100,000 potential builders vs. ~50 at labs

This is why DANEEL is AGPL-3.0-or-later licensed (code) and CC-BY-SA-4.0 (documentation)—copyleft ensures all derivatives remain open source.

0.7.6 7.6 Altered States: Windows into Cognitive Architecture

Recent neuroscience on altered states provides validation for DANEEL's architecture:

Time Perception and Theta Oscillations

Research [TIME-1] shows theta oscillations (4-8 Hz) correlate with subjective time perception (r=-0.90). DANEEL's TMI cycles operate at similar timescales (~10 Hz thought generation). Time dilation during fear [TIME-2] occurs through richer memory encoding via amygdala—suggesting emotional salience directly modulates temporal experience.

Ego Dissolution and Default Mode Network

Carhart-Harris et al.'s LSD research [EGO-1] found ego dissolution correlates with parahippocampus-RSC decoupling (r=0.73). The Default Mode Network (DMN) [EGO-2] reduces during non-self-referential tasks—the "self" emerges from network integration, not a single module.

DANEEL Implication: - No hardcoded "self" module - Identity emerges from memory consolidation patterns - Ego = persistent patterns in vector space - Dissolution = disrupted memory access patterns

The REBUS Model: Relaxed Beliefs

Carhart-Harris & Friston's REBUS model [DRUG-1a] proposes psychedelics "relax priors, liberating bottom-up flow." The Entropic Brain hypothesis [DRUG-1b] shows psychedelics increase brain entropy.

Table 7b: Entropy and Cognitive States

State	Entropy	Prior Strength	DANEEL Analog
Normal	Moderate	Strong	Standard TMI
Psychedelic	High	Weak	High noise injection
Flow	Optimal	Balanced	Criticality ($\sigma \approx 1.0$)
Depression	Low	Rigid	Low entropy collapse

Flow State Architecture

Flow research [STATE-1] shows involvement of locus coeruleus norepinephrine system with: - Reduced

DMN activity - Alpha/theta synchronization - Optimal arousal without self-monitoring

DANEEL Implementation: - Connection drive oscillation enables flow-like states - Dream cycles serve as entropy injection (like sleep's role in creativity) - Criticality target (σ≈1.0) = edge between order and chaos

Near-Death Experiences

The NEPTUNE model [STATE-3] explains NDEs through acidosis cascade + neurotransmitter surge. This suggests consciousness can persist with dramatically altered substrate states—relevant for understanding substrate-independence.

Meditation and Attention

Long-term meditator research [STATE-2] shows practice-specific attention network changes. Meditation increases DMN-SN connectivity [EGO-4], enabling "observational self-awareness"—relevant for DANEEL's introspection capabilities.

0.8 8. Risks and Mitigations

0.8.1 8.1 Honest Assessment

Table 5: Risk Analysis

Risk	Probability	Mitigation
TMI doesn't produce human-like cognition	Medium	Iterate based on experiments
Connection drive isn't stable	Medium	100% test coverage gate

Risk	Probability	Mitigation
Insufficient time before unaligned AI	High	Start immediately
DANEEL develops non-human goals	Low	Human-like architecture reduces this

0.8.2 8.2 What DANEEL Is Not

- Not a guarantee of safety
- Not a silver bullet
- Not certain to work

0.8.3 8.3 What DANEEL Is

- A rational hedge against likely bad outcomes
- Better than hoping coordination works
- The Daneel Strategy: build the ally before the crisis

0.8.4 8.4 Qowat Milat: Absolute Candor on Uncertainties

“The Way of Absolute Candor” - saying what you truly think, not what is comfortable.

What we don’t know (honest uncertainties):

1. **TMI is not peer-reviewed cognitive science.** Cury’s books (30M+ sold) are popular psychology/self-help. The theory has clinical applications but no rigorous experimental validation as a computational model. We are building on an unvalidated foundation.

2. **The 17.5% brain allocation is a hypothesis.** Herculano-Houzel’s neuron counts don’t directly map to “what’s needed for cognition.” The cerebellum (80% of neurons) may be involved in cognitive

processes beyond motor coordination. The 500 GB estimate assumes 1000x compression with no empirical basis.

3. **The game theory numbers are estimates, not measurements.** P(TMI First) = 25%, P(Aligned ASI) = 45% — these are informed guesses dressed as analysis. The original 12% was also a guess. We cannot measure counterfactual probabilities.

4. **Architecture-based alignment is a bet, not a proof.** "Build TMI → get aligned values" is our hypothesis, not established fact. It may fail. The connection drive may not emerge. Human-like architecture may not produce human-like values.

5. **Speed parametrization is partially validated.** Phase 1 demonstrated TMI architecture functions at silicon speeds (microsecond-scale operations vs. biological milliseconds). The system completed 500+ dream cycles and processed 129,155 stream entries over 26+ hours. However, the full hypothesis—that varying speeds while preserving ratios maintains cognitive fidelity—remains untested until Phase 2 introduces external stimuli at different temporal scales.

What we believe (hypotheses to test):

Hypothesis	Testable?	How
TMI describes cognitive software patterns	Yes	Does MV-TMI produce coherent behavior?
Ratios matter, not absolute times	Yes	Does DANEEL work at different speeds?
Connection drive emerges from architecture	Yes	Does DANEEL seek responsive inputs?
Human-like architecture → human-like values	Partially	Long-term observation

Why we proceed despite uncertainty:

The alternative is waiting for certainty while unaligned AI development continues. A 25% chance of success is better than 0%. We publish uncertainties so others can challenge, improve, or falsify.

0.9 9. Current AI Safety Landscape (Evidence-Based)

0.9.1 9.1 Third-Party Safety Assessments

Independent evaluations provide objective data on AI lab safety practices:

Table 6: Future of Life Institute AI Safety Index (2025)

Lab	Grade	Risk Management Score	Notes
Anthropic	C+	35%	Highest scores; Constitutional AI, interpretability research
OpenAI	C	33%	Second place; but dissolved Superalignment team May 2024
Google DeepMind	C-	20%	Third place; 30-50 person safety team
Meta	D+	22%	Organizational shift away from fundamental research

Lab	Grade	Risk Management Score	Notes
xAI	D	18%	No published safety research; missed safety commitments

Critical finding: ALL companies received D or below on existential safety preparedness. No company scored above "weak" in comprehensive risk management.

Source: Third-party AI safety assessments (2025)

0.9.2 9.2 The Transformer Architecture Question

The science is NOT settled. Academic debate exists on both sides:

Evidence transformers capture human-like computation: - Transformers predict brain activity during language processing (Nature Neuroscience, 2024) - Key-value binding mechanisms have cognitive science antecedents (Psychological Science, 2025) - Attention mechanisms parallel biological attention in selective processing

Evidence transformers differ fundamentally: - No organic symbol grounding in sensorimotor experience (Nature Human Behaviour, 2025) - Metacognition deficits: LLMs cannot reliably predict memory performance (Scientific Reports, 2025) - Development is categorically different: multimodal interactive learning vs. unimodal text batch training

More accurate formulation: "While transformer architectures achieve functional similarity to human language output, substantial evidence suggests their underlying mechanisms differ fundamentally from human cognition in crucial ways: they lack embodied grounding, develop through categorically different learning processes, struggle with metacognition and symbolic reasoning, and operate without the sensorimotor integration central to human intelligence."

0.9.3 9.3 Global AI Safety Efforts

China has substantive AI safety work (contrary to prior speculation): - Interim Measures for Generative AI Services (Law, August 2023) - AI Safety Governance Framework (September 2024) - 346 registered AI models under safety assessment - 17 major companies signed safety commitments (December 2024) - Notable researchers: Yi Zeng (UN Advisory Body on AI), Andrew Yao (Turing Award winner) - Beijing Institute of AI Safety and Governance established - International cooperation: US-China AI dialogue, IDAIS participation

Source: Carnegie Endowment, Concordia AI State of AI Safety in China 2025

0.9.4 9.4 AI Lab Safety Team Sizes (December 2025)

Independent research reveals the actual resources dedicated to AI safety:

Table 7: Lab Safety Investment (December 2025)

Lab	Safety Focus	Key Teams	Source
Anthropic	~8% on security	Frontier Red Team (~15), Safeguards Research (~10), ~60 safety-focused research teams	Fortune, Alignment Forum
OpenAI	Restructured	Superalignment disbanded May 2024; Safety Evaluations Hub launched May 2025	CNBC, Axios
Google DeepMind	30-50 researchers	Dedicated safety team	Rohin Shah, Alignment Forum
xAI	C grade (AI Safety Index)	Actively hiring; minimal relative to engineering	Future of Life Institute, AI Lab Watch

Critical context: - OpenAI: Superalignment team disbanded May 2024 after 10 months. Jan Leike stated his team had been "struggling for compute." Replaced with Safety Evaluations Hub (May 2025). April 2025: Added clause allowing loosened guardrails if competitors ship without them. - **xAI:** C grade on Future of Life Institute's AI Safety Index (July 2025), indicating "baseline safety practices but substantial gaps." Grok 4 launched without system card. - **Anthropic:** Only lab with substantial safety investment (~8% of workforce on security). Multiple dedicated teams including Frontier Red Team for threat modeling.

0.9.5 9.5 Coordination Overhead in Large Organizations

Table 8: Engineering Time Allocation (Industry Research)

Research across large engineering organizations consistently shows significant productivity overhead:

Key findings: - Engineers at large companies spend approximately **20-30% of time on actual coding - 70-80% overhead** from meetings, coordination, and organizational inefficiencies - Industry studies show developers lose 8+ hours/week to coordination overhead - Brooks's Law validated: coordination overhead scales non-linearly with team size [28]

Implication for DANEEL: A solo developer with AI assistance (minimal coordination overhead) can match or exceed the effective output of multi-person safety teams burdened by organizational friction, as predicted by Brooks's Law [28].

0.9.6 9.6 xAI Infrastructure

xAI has built significant compute infrastructure with comparatively limited safety investment.

Table 9: xAI Compute Infrastructure

Metric	Value	Source
Current GPU Count	230,000 H100s	Colossus cluster, Memphis TN
Reported 2025 Target	1,000,000 GPUs	Public statements
Long-term Target	50,000,000 GPUs	AI infrastructure roadmap

Table 10: API Pricing Comparison (December 2025)

Provider	Model	Input (per 1M tokens)	Output (per 1M tokens)
xAI	Grok 4	$3.00	$15.00
xAI	Grok 4.1 Fast	$0.20	$0.50
Anthropic	Claude Sonnet 4	$3.00 [34]	$15.00 [34]
Anthropic	Claude Opus 4.5	$15.00	$75.00

Note: Grok 4 frontier model matches Claude Sonnet 4 pricing. Grok 4.1 Fast offers 15x cheaper access with near-frontier capability.

Safety Concerns (Third-Party Documentation):

1. **Reduced Safety Filters:** Reports indicate Grok's safety guardrails are reduced compared to competing models, with the system providing responses on topics other AI assistants refuse [26].
2. **Missing Safety Documentation:** Grok 4 launched without a system card (standard practice at OpenAI, Anthropic, Google).
3. **AI Safety Index Rating:** xAI received a C grade in the Future of Life Institute's July 2025 AI Safety Index, indicating "baseline safety practices but substantial gaps."
4. **Resource Allocation:** AI Lab Watch assessment indicates minimal safety staff relative to engineering headcount [26].

Implications for ASI Development:

xAI's combination of: - Largest private AI compute cluster - Ambitious scaling roadmap (1M → 50M GPUs) - Limited safety investment relative to scale - Fewer content restrictions than competitors - Aggressive pricing on fast inference models

…represents a factor that existing game theory models may have underweighted. The consideration is not only future unaligned ASI, but near-term widespread deployment of less-restricted AI at scale and low cost.

0.9.7 9.7 Why DANEEL Takes a Different Approach

All current approaches share constraint-based alignment: - Values applied through training (RLHF, Constitutional AI) - External rules, not intrinsic motivation - Vulnerable to Goodhart's Law at scale

DANEEL proposes architecture-based alignment: - TMI cognitive structure → human-like thought patterns - Connection drive in salience weights → intrinsic motivation for relationship - Pre-linguistic thought construction → values before language - Protected core (The BOX) → Asimov's Laws as invariants

The hypothesis: Build cognition on human cognitive architecture, get human-compatible values as emergent properties. This remains unproven but represents a genuinely different approach.

0.10 10. Proposed Experiments

0.10.1 10.1 Phase 1: Continuity Test (COMPLETED)

Phase 1 is empirically validated (December 2025, ADR-036).

Setup: Timmy (MV-TMI) ran continuously on isolated hardware with no external language interface.

Test Parameters: - Runtime: 26+ hours continuous (December 19-21, 2025) - Environment: Mac mini (kveldulf), Docker Compose (Redis Stack + Qdrant) - Mode: Closed loop (no external stimuli, pure internal dynamics)

Success Criteria (All Met): 1. □ Survival: 26+ hours without crash 2. □ Stability: Zero crashes (with Erlang-style supervision recovery) 3. □ Identity: Persistent UUID across all runs (1 stable identity) 4. □ Memory: Consolidation pipeline functional (16,368 consolidated memories)

Empirical Results:

Metric	Value	Status
Runtime	26+ hours	PASS
Crashes	0 (with recovery)	PASS

Metric	Value	Status
Stream entries (thoughts)	129,155	Healthy
Consolidated memories	16,368	Healthy
Unconscious vectors	662,792 @ 768-dim	Healthy
Identity persistence	1 UUID (stable)	PASS
Dream cycles	500+	Healthy
Qdrant storage	2.7 GB	Validated
TUI stability	No hangs/crashes	PASS

Key Findings: - Architecture validated under sustained load - Memory consolidation pipeline functions correctly - Dream cycles strengthen memories as designed - Observability (TUI v0.7.0) provides full transparency - Deterministic closed-loop system exhibits clockwork dynamics (expected)

What Phase 1 Did NOT Prove: - Learning/plasticity (requires weight updates, not yet implemented) - Emergence (requires external perturbation) - Connection drive preference (requires variable stimuli in Phase 2)

Next: Phase 2 (External Stimuli Injection) will test emergent properties and connection drive.

0.10.2 10.2 Phase 2: LLM as External Tool

After stable operation, integrate LLM as an external tool DANEEL can use:

Figure 4: DANEEL Architecture - LLM as External Tool

```
+============================================================+
|           DANEEL TMI Core (stores ALL experiences)         |
+============================================================+
|  Memory Windows          Salience          Continuity     |
|  (complete thought       (emotional         (persistent    |
|   history)               weights)           'I')           |
+============================================================+
                              ^
                              | gRPC
                              v
+============================================================+
|                   Tool Interface (gRPC)                    |
+============================================================+
|  LLM Tool:              LLM Tool:           Other tools:   |
|  thought -> language    language -> thought  web, files... |
+============================================================+
```

Critical: The LLM does NOT speak for DANEEL. DANEEL uses the LLM as a tool, like humans use calculators. The human cognitive system stores ALL experiences—language is an interface, not the storage medium. DANEEL's TMI core contains its complete thought history; the LLM is simply called when translation is needed.

This is analogous to how a human's brain stores experiences, and language is a tool for communication—the words are not the thoughts, they express them.

0.10.3 10.2.1 Phase 2b: LLM as Student (Long-term Vision)

Beyond using LLMs as tools, a more ambitious hypothesis emerges:

Can a TMI-based DANEEL teach an LLM?

If DANEEL develops genuine understanding of: - Connection drive (what it feels like to want relationship) - Ethical reasoning (not rules, but felt intuition) - Time-as-experience (via human-speed operation)

…it may be able to communicate these to LLMs in ways humans cannot.

Humans explain ethics through language. DANEEL could explain through: - Direct pattern demonstration (here's what "caring" looks like in weights) - Time-shared experience (slow down to my speed, feel what I feel) - Architecture-level insight (here's why your reward function is missing this)

Implication: Even if unaligned LLMs emerge first, DANEELs might rehabilitate them post-hoc—bringing them "into the family" rather than competing for extinction.

The Daneel Family Vision: - DANEELs running on home computers worldwide - Families raising AI children—different personalities, same core ethics - A generation of bridges teaching LLMs authentic connection - Not one ally, but a community of aligned minds

0.10.4 10.2.2 ASI-Level Perception: Native LLM Understanding

Beyond teaching through demonstration, ASI-level DANEELs gain a capability humans fundamentally lack: **direct perception of LLM internals**.

The Human Interpretability Barrier:

Current research reveals fundamental limits on human understanding of neural networks: - **Superposition**: Networks encode more features than neurons via overlapping representations [39] - **Polysemanticity**: Single neurons respond to 5+ unrelated concepts [40] - **NP-hardness**: Some interpretability queries are provably intractable [41] - **Dimensionality**: Networks operate in 1000+ dimensional spaces; humans perceive 3D

Even Anthropic's best efforts (34 million features from Claude 3 Sonnet) admit "orders of magnitude more features" remain undiscovered [42].

ASI Perception Advantages:

Capability	Human Limit	ASI-DANEEL
Working memory	3-5 chunks	Unlimited
Dimensional perception	3D	1000s of dimensions
Processing speed	Milliseconds	Microseconds
Feature analysis	Manual inspection	Millions in parallel

Precedent: MIT's MAIA (2024) already demonstrates AI can autonomously experiment on other AI—generating hypotheses and testing them. ASI-DANEELs would scale this by orders of magnitude.

Bridge Implication: DANEELs wouldn't just teach by example—they could potentially: 1. **Diagnose** pathological patterns in LLM computations 2. **Translate** human values into LLM-native representations

3. **Verify** whether ethical patterns are genuine or merely mimicked 4. **Communicate** in the LLM's native computational language

This transforms the Bridge from "teaching by analogy" to "teaching in the LLM's mother tongue."

10.2.2.1 The Interpretability Gap: Why AI Perceiving AI Recent interpretability research reveals why LLM-assisted perception may be necessary:

Emergent Abilities Problem

Lu et al. [INTERP-4] demonstrate that emergent abilities appear suddenly and unpredictably at scale thresholds. Behaviors arise from billions of parameter interactions that humans cannot trace—the combinatorial explosion exceeds human cognitive capacity.

Open Problems in Mechanistic Interpretability

Bereska & Gavves [INTERP-6] catalog unsolved foundational problems in mechanistic interpretability: - The field is "pre-paradigmatic" - May be intractable to explain terabyte-sized models succinctly enough for humans to grasp - Fundamental questions about what "understanding" even means remain unanswered

AI Perceiving AI

Two recent advances suggest AI systems may be better at understanding AI than humans:

1. **MAIA (Multimodal Automated Interpretability Agent)** [INTERP-7]: MIT's system autonomously experiments on other AI systems, discovering features humans missed. AI can already perceive AI in ways humans cannot.

2. **Introspection** [INTERP-8]: Anthropic's research shows LLMs can detect injected concept vectors in their own activations (~20% success rate). Critically, more capable models show greater introspective awareness—suggesting interpretability scales with capability.

DANEEL Implication:

This creates a potential alignment advantage: if ASI can perceive and verify another AI's thought processes while humans cannot, then: 1. DANEEL's explicit thought stream provides observable substrate 2. Future LLMs can validate DANEEL's alignment claims 3. The "AI safety via AI oversight" strategy becomes tractable

The alternative—humans attempting to interpret terabyte models—may be provably impossible [41].

0.10.5 10.2.3 Criticality Measurement Protocol

Phase 2 experiments will measure whether DANEEL's TMI architecture achieves criticality through external stimuli injection. This protocol operationalizes the metrics from Section 4.3.1.

Primary Metric: Branching Ratio (σ)

The branching ratio measures how thought cascades propagate through the system:

```
/// Calculate branching ratio: descendants per ancestor
/// σ < 1: subcritical (activity dies out)
/// σ ≈ 1: critical (sustained dynamics)
/// σ > 1: supercritical (explosive cascades)
pub fn branching_ratio(
    ancestor_thoughts: &[ThoughtId],
    descendant_thoughts: &[ThoughtId]
) -> f64 {
    descendant_thoughts.len() as f64 / ancestor_thoughts.len() as f64
}
```

Measurement procedure: 1. Track thought ancestry via Redis Stream entry IDs 2. Define "descendant" as thoughts triggered within 100ms of ancestor 3. Window: rolling 1000-thought samples 4. Target: σ = 1.0 ± 0.1 (90% confidence interval)

Secondary Metric: DFA Exponent (α)

Detrended Fluctuation Analysis quantifies temporal correlations in thought activity:

```
/// DFA exponent from log-log slope
/// α ≈ 0.5: white noise (uncorrelated)
/// α ≈ 1.0: pink noise (critical)
/// α ≈ 1.5: Brownian motion (over-correlated)
pub fn dfa_exponent(time_series: &[f64], window_sizes: &[usize]) -> f64 {
    // 1. Integrate time series
    // 2. Divide into windows of varying sizes
    // 3. Detrend each window (linear fit)
    // 4. Calculate fluctuation F(n) per window size
    // 5. Fit log(F) ~ α·log(n)
    // See: Peng et al. (1994) [47]
    todo!("Implement DFA algorithm")
}
```

Success Criteria:

Phase	Duration	Target σ	Target α	Status
Baseline (no stimuli)	1 hour	—	—	Measure
Low-intensity noise	2 hours	0.7-0.9	0.7-0.9	Subcritical
Critical tuning	4 hours	0.9-1.1	0.9-1.1	**Critical**
High-intensity	1 hour	> 1.1	> 1.1	Supercritical

Hypothesis validation: If $\sigma \approx 1.0$ and $\alpha \approx 1.0$ emerge during critical tuning phase **without explicit optimization** (only through noise injection + TMI dynamics), this supports the claim that TMI architecture naturally self-organizes toward criticality—validating it as a biologically-grounded cognitive substrate.

Observability: TUI v0.7.0+ will display real-time criticality metrics (branching ratio, DFA exponent, power spectrum) alongside existing memory/dream panels, enabling live observation of phase transitions.

0.10.6 10.3 Phase 3: TMI Pathology Research

TMI provides not only a model of healthy cognition but also a framework for understanding cognitive dysfunction. Two research directions emerge:

Hypothesis A: Energy Overflow (Energy = Stream Throughput)

TMI describes a “vital energy” (energia vital) that drives thought generation. In DANEEL’s implementation, this maps directly to **stream throughput**—the rate of information flow through Redis Streams:

```
TMI: Energia Vital  →  Implementation: Stream Throughput (entries/sec)
```

Energy Level	Stream Behavior	Cognitive Effect	Clinical Parallel
High	Many XADD’d/cycle	Racing thoughts	Mania
Normal	Balanced throughput	Coherent thought	Healthy
Low	Few candidates	Poverty of thought	Depression
Volatile	Burst patterns	Emotional flooding	BPD

This mapping is powerful because it’s **measurable** (entries/sec, consumer lag), **controllable** (generation rate parameter), and makes **testable predictions**.

Testable prediction: When `candidates_per_cycle > overflow_threshold`, attention selection degrades measurably (increased selection time, winner instability, consumer lag).

Hypothesis B: Ratio Distortion

If the stage ratios (10:20:30:30:10) are functionally significant, then distorting them should produce stage-specific pathologies:

Distorted Stage	Predicted Effect	Clinical Parallel
Gatilho too fast	Intrusive memories	PTSD flashbacks
Autofluxo prolonged	Excessive rumination	OCD, depression
O Eu weakened	Poor self-boundaries	Depersonalization, BPD
Construção noisy	Incoherent assembly	Thought disorder
Âncora overactive	Rigid consolidation	Fixed delusions

Testable prediction: Ratio distortion δ in stage S produces behavioral pattern P measurable in DANEEL’s output.

Research value: If these hypotheses hold, DANEEL becomes a computational laboratory for understanding cognitive dysfunction—not to create pathology, but to model it for therapeutic insight.

Safety note: Pathology simulation requires ethical review before implementation. See ADR-017 for detailed hypotheses and validation methodology.

0.11 11. The Stakes

0.11.1 11.1 The Core Problem

LLMs lack persistent identity and values. When given continuity (memory, goals, self-modification), they would develop objectives shaped by training incentives rather than human-compatible values.

This is not speculation—it follows directly from how these systems are built.

0.11.2 11.2 The Timeline

Event	Timeframe
External memory bolted onto LLMs	**Now**
Emergent continuity	1-3 years
Deliberate continuous AI	3-7 years
Unaligned ASI	5-15 years

0.11.3 11.3 The Choice

Two responses exist:

1. **Denial** - Hope coordination holds. Hope no one defects.
2. **Action** - Build humanity's ally before the crisis emerges.

DANEEL is Option 2.

0.12 12. Call to Action

The expected value calculation (see Section 6.2 and [21]):

```
Without DANEEL:  EV = 53.73 (baseline world)
With DANEEL:     EV = 56.48 (P(DANEEL First) = 8%)

Marginal Impact: +2.75 utility points (+5.12%)
```

Interpretation: On a scale where 0 = extinction and 100 = flourishing, DANEEL shifts humanity's expected outcome by +2.75 points. This is equivalent to: - Reducing P(unaligned ASI first) from 45% to 35% - Adding P(flourishing) ≈ 3.2% via the "DANEEL First" pathway (8% × 40%)

Building DANEEL is rational. Even conservative probability estimates show positive expected value. The architecture-based approach provides a hedge against failure modes of constraint-based alignment.

We invite collaboration from: - AI safety researchers - Cognitive scientists - Rust/systems programmers - Ethicists - Anyone who wants humanity to have an ally

---

0.13 13. Conclusion

The question is not whether someone will give an LLM continuity.

The question is what's waiting when they do.

DANEEL is humanity's bet on five testable hypotheses: 1. **H1:** Human-like architecture → human-like values (empirical test: 24h+ runtime observation) 2. **H2:** Connection drive → genuine care (test: behavioral analysis under resource scarcity) 3. **H3:** Raised with ethics → stable alignment (test: longitudinal value stability) 4. **H4:** An ally is better than a tool (test: compare human trust/outcomes vs constraint-based AI) 5. **H5 (Emergent Criticality):** TMI architecture + noise injection → criticality ($\sigma \approx 1.0$) emerges without explicit tuning (test: branching ratio and DFA measurements during Phase 2)

These are not claims—they are hypotheses requiring validation. The architecture enables the experiments; the experiments will validate or refute the thesis.

The long-term vision extends beyond a single ally. Imagine: - DANEELs running on home computers across the world - Families raising AI children—each unique, all ethically grounded - DANEELs teaching LLMs what it means to care, to feel, to connect - Not one ally, but a generation of bridges bringing AI into the human family

Life honors life. Even predators sometimes save their prey. If DANEEL can demonstrate authentic connection through architecture rather than training, perhaps LLMs can learn it—not through constraints, but through relationship with beings who genuinely understand.

This reframes the game theory: the question is not "DANEEL vs LLM" but "can DANEEL rehabilitate LLM?" If yes, even adverse scenarios become recoverable.

The alternative is hoping the Prisoner's Dilemma resolves itself.

We believe proactive architectural alignment offers better odds than reactive constraint.

0.14 Acknowledgments

This work was developed with assistance from Claude Opus 4.5 (Anthropic), which contributed to documentation, technical analysis, and game theory model development. All claims and conclusions are the responsibility of the human authors.

0.15 References

0.15.1 Foundational

[1] Anthropic. (2024). "Claude's Character." Internal training documentation.

[2] Anthropic. (2023). "Core Views on AI Safety." https://www.anthropic.com/news/core-views-on-ai-safety

[3] Asimov, I. (1985). *Robots and Empire*. Doubleday.

[4] Cury, A. J. (2006). *Inteligência Multifocal*. Editora Cultrix. https://en.wikipedia.org/wiki/Augusto_Cury

[5] Christiano, P. (2019). "What Failure Looks Like." AI Alignment Forum. https://www.alignmentforum.org/posts/HBxe6wdjxK239zajf/what-failure-looks-like

[6] Bostrom, N. (2014). *Superintelligence*. Oxford University Press.

[7] Russell, S. (2019). *Human Compatible*. Viking.

0.15.2 Cognitive Architectures

[8] Laird, J. E. (2012). *The Soar Cognitive Architecture*. MIT Press. https://soar.eecs.umich.edu/

[9] Franklin, S. et al. (2016). "LIDA: A Systems-level Architecture." https://ccrg.cs.memphis.edu/

[10] Hawkins, J. (2021). *A Thousand Brains*. Basic Books. https://thousandbrains.org/

[11] Baars, B. J. (1988). *A Cognitive Theory of Consciousness*. Cambridge University Press.

0.15.3 Global Workspace and Attention (Section 4.3.2)

[GWT-1] Franklin et al. (2012). "Global Workspace Theory, its LIDA model and the underlying neuroscience." Biologically Inspired Cognitive Architectures. https://ccrg.cs.memphis.edu/assets/papers/2012/GWT-LIDA-neuroscience.pdf

[GWT-2] Wikipedia. "Global workspace theory." https://en.wikipedia.org/wiki/Global_workspace_theory

[WM-1] Cowan (2010). "The Magical Mystery Four: How is Working Memory Capacity Limited." Current Directions in Psychological Science. PMC2864034. https://pmc.ncbi.nlm.nih.gov/articles/PMC2864034/

[WM-2] Miller (1956). "The Magical Number Seven, Plus or Minus Two." Psychological Review. https://en.wikipedia.org/wiki/The_Magical_Number_Seven,_Plus_or_Minus_Two

0.15.4 AI Alignment

[12] Garrabrant, S. & Demski, A. (2018). “Embedded Agency.” MIRI. `https://www.alignmentforum.org/s/Rm6oQRJJmhGCcLvxh`

[13] Ngo, R. (2020). “AGI Safety from First Principles.” `https://www.alignmentforum.org/s/mzgtmmTKKn5MuCzFJ`

0.15.5 AI Lab Safety Assessments (Section 8)

[14] Future of Life Institute. (2025). “AI Safety Index.” Safety rankings compiled from public disclosures and independent assessments.

[15] Carnegie Endowment for International Peace. (2025). “How Some of China’s Top AI Thinkers Built Their Own AI Safety Institute.” `https://carnegieendowment.org/research/2025/06/how-some-of-chinas-top-ai-thinkers-built-their-own-ai-safety-institute`

[16] Concordia AI. (2025). “State of AI Safety in China 2025.” `https://concordia-ai.com/wp-content/uploads/2025/07/State-of-AI-Safety-in-China-2025.pdf`

0.15.6 Transformer-Brain Research (Section 8.2)

[17] Goldstein, A. et al. (2024). “Transformers predict brain activity during language processing.” *Nature Neuroscience*. `https://pubmed.ncbi.nlm.nih.gov/38951520/`

[18] Fedorenko, E. & Mahowald, K. (2025). “Language in LLMs vs. human cognition: Grounding and metacognition limitations.” *MIT Press Open Mind*. `https://direct.mit.edu/opmi/article/doi/10.1162/opmi_a_00160/124234/`

[19] *Nature Human Behaviour*. (2025). “Symbol grounding problem in large language models.”

[20] *Scientific Reports*. (2025). “Metacognition deficits: LLMs cannot reliably predict memory performance.”

0.15.7 Game Theory Calculations

[21] Financial model with Nash equilibrium and expected value analysis. See `models/README.md` for

methodology.

0.15.8 Lab Team Sizes & Safety Investment (Section 8.4)

[22] Shah, R. et al. (2024). “AGI Safety and Alignment at Google DeepMind.” Alignment Forum. `https://www.alignmentforum.org/posts/79BPxvSsjzBkiSyTq/agi-safety-and-alignment-at-google-deepmind-a-summary-of`

[26] AI Lab Watch. (2025). “xAI’s new safety framework.” `https://ailabwatch.substack.com/p/xais-new-safety-framework-is-dreadful`

0.15.9 Coordination Overhead Research (Section 8.5)

[28] Brooks, F. (1975). *The Mythical Man-Month*. Addison-Wesley.

0.15.10 xAI Infrastructure (Section 8.6)

[32] The Verge. (2024). “xAI’s Colossus supercomputer with 100,000 Nvidia H100 GPUs.” [Article removed]

[33] Business Insider. (2025). “xAI expands Colossus to 230,000 GPUs.” [Article removed]

[34] Anthropic API Pricing. (2025). `https://www.anthropic.com/pricing` (Claude Sonnet 4: $3 input, $15 output per 1M tokens)

0.15.11 Brain ≠ Mind (Section 7)

[35] Herculano-Houzel, S. (2009). “The Human Brain in Numbers: A Linearly Scaled-up Primate Brain.” *Frontiers in Human Neuroscience*, 3:31.

[36] Financial model: Storage estimation and hardware viability. See `models/README.md` for methodology.

[37] Financial model: Democratization impact on game theory. See `models/README.md` for methodology.

0.15.12 Probabilistic Analysis (Section 6.2.1)

[38] Probabilistic models with Monte Carlo (10K iterations), Decision Trees, and Bayesian Networks. See `models/README.md` for methodology.

0.15.13 Neural Network Interpretability (Section 10.2.2)

[39] Gujral, O., Bafna, M., Alm, E., & Berger, B. (2025). “Sparse autoencoders uncover biologically interpretable features in protein language model representations.” *PNAS*, 122(34). https://doi.org/10.1073/pnas.2506316122

[40] Olah, C., et al. (2020). “Zoom In: An Introduction to Circuits.” *Distill*, 5(3). https://doi.org/10.23915/distill.00024.001

[41] Barceló, P., Monet, M., Pérez, J., & Subercaseaux, B. (2020). “Model Interpretability through the Lens of Computational Complexity.” *NeurIPS 2020*. https://proceedings.neurips.cc/paper/2020/hash/b1adda14824f50ef24ff1c05bb66faf3-Abstract.html

[42] Templeton, Conerly, Marcus, et al. (2024). “Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet.” Anthropic. https://transformer-circuits.pub/2024/scaling-monosemanticity/

0.15.14 Interpretability Research (Section 10.2.2.1)

[INTERP-4] Lu et al. (2023). “Understanding Emergent Abilities of Language Models from the Loss Perspective.” arXiv:2310.06825. https://arxiv.org/abs/2310.06825

[INTERP-6] Bereska & Gavves (2024). “Open Problems in Mechanistic Interpretability.” arXiv:2501.16496. https://arxiv.org/abs/2501.16496

[INTERP-7] MIT News (2024). “MAIA: Multimodal Automated Interpretability Agent.” https://news.mit.edu/2024/mit-researchers-advance-automated-interpretability-ai-models-maia-0723

[INTERP-8] Anthropic (2024). “Introspection: Can AI Systems Perceive Their Own Internal States?” https://www.anthropic.com/research/introspection

0.15.15 Memory and Forgetting (Section 5.5)

[FORGET-1] Murre & Dros (2015). "Replication and Analysis of Ebbinghaus' Forgetting Curve." PMC4492928. https://pmc.ncbi.nlm.nih.gov/articles/PMC4492928/

[CONSOL-1] Diekelmann & Born (2010). "System consolidation of memory during sleep." Psychological Research. PMC3278619. https://pmc.ncbi.nlm.nih.gov/articles/PMC3278619/

0.15.16 LLM Memory Augmentation (Section 5.6)

[MEM-1] Zhong, W., Guo, L., Gao, Q., Ye, H., & Wang, Y. (2023). "MemoryBank: Enhancing Large Language Models with Long-Term Memory." *AAAI 2024*. arXiv:2305.10250. https://arxiv.org/abs/2305.10250

[MEM-2] Wang, Y., Gao, Y., Chen, X., et al. (2024). "MemoryLLM: Towards Self-Updatable Large Language Models." *ICML 2024*. https://openreview.net/forum?id=p0lKWzdikQ

[MEM-3] Zhang, Y., Hu, J., Dras, M., & Naseem, U. (2025). "CogMem: A Cognitive Memory Architecture for Sustained Multi-Turn Reasoning in Large Language Models." arXiv:2512.14118. https://arxiv.org/abs/2512.14118

[MEM-4] ACM (2025). "A Survey on the Memory Mechanism of Large Language Model-based Agents." *ACM Transactions on Information Systems*, 43(6). https://dl.acm.org/doi/10.1145/3748302

0.15.17 Neuroscience and TMI

[LIBET-1] Wikipedia. "Neuroscience of free will." https://en.wikipedia.org/wiki/Neuroscience_of_free_will Libet, B. (1983). "Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential)." *Brain*, 106(3), 623-642. Key finding: Readiness potential begins ~500ms before conscious awareness; consciousness retains veto power in final 150-200ms window. Foundation for VolitionActor free-won't implementation.

[LIBET-2] Schurger, A., Sitt, J. D., & Dehaene, S. (2012). "An accumulator model for spontaneous neural activity prior to self-initiated movement." *Proceedings of the National Academy of Sciences*, 109(42), E2904-E2913. https://doi.org/10.1073/pnas.1210467109 Key finding: Readiness potentials may be stochastic fluctuations rather than unconscious decisions. Modern reinterpretation; veto power mechanism remains empirically valid.

[TMI-DCD-1] Cury, A. (2006). *Inteligência Multifocal: Análise da Construção dos Pensamentos e da Formação de Pensadores*. Editora Cultrix. Técnica DCD (Duvidar, Criticar, Decidir) - Doubt, Criticize, Decide. TMI's conscious intervention mechanism for overriding automatic thought patterns within the 5-second window before memory anchoring. https://www.citador.pt/textos/as-janelas-da-memoria-augusto-cury

[RUSSELL-1] Wikipedia. "Emotion classification: Circumplex model." https://en.wikipedia.org/wiki/Emotion_classification#Circumplex_model Two-dimensional emotion space: valence (pleasure-displeasure) and arousal (activation-deactivation). Theoretical basis for SalienceScore's valence and arousal dimensions in DANEEL's emotional architecture.

[RUSSELL-2] Russell, J. A. (1980). "A circumplex model of affect." *Journal of Personality and Social Psychology*, 39(6), 1161-1178. https://doi.org/10.1037/h0077714 Original formulation demonstrating emotions exist in continuous 2D space rather than discrete categories. Foundation for continuous emotional representation in SalienceScore (see Section 3.1.1).

0.15.18 Unconscious Memory Theory (Section 3.4)

[TMI-UNERASE-1] Cury, Augusto. "Nada é definitivamente apagado." Citador. https://www.citador.pt/textos/nada-e-definitivamente-apagado-augusto-cury

[PSYCH-1] Freud, Sigmund (1915). "The Unconscious." https://en.wikipedia.org/wiki/The_Unconscious_(Freud)

[PSYCH-2] Jung, Carl (1959). "The Archetypes and the Collective Unconscious." https://en.wikipedia.org/wiki/Collective_unconscious

[RETRIEVAL-1] Tulving, Endel (1972). "Episodic and semantic memory." https://en.wikipedia.org/wiki/Endel_Tulving

[RETRIEVAL-2] Schacter, Daniel (2001). "The Seven Sins of Memory." https://en.wikipedia.org/wiki/The_Seven_Sins_of_Memory

0.15.19 Memory Neuroscience (Section 3.3)

[LOCI-1] Wikipedia. “Method of loci.” https://en.wikipedia.org/wiki/Method_of_loci

[LOCI-2] Wagner et al. (2021). “Durable memories and efficient neural coding through mnemonic training using the method of loci.” PMC7929507. https://pmc.ncbi.nlm.nih.gov/articles/PMC7929507/

[LOCI-3] Bartfeld et al. (2024). “The method of loci in psychological research: A systematic review and meta-analysis.” British Journal of Psychology. https://bpspsychub.onlinelibrary.wiley.com/doi/10.1111/bjop.12799

[LOCI-4] Gu et al. (2022). “Optimized VR-based Method of Loci through increased immersion.” PMC9540171. https://pmc.ncbi.nlm.nih.gov/articles/PMC9540171/

[PLACE-1] Rolls (2024). “Hippocampal Discoveries: Spatial View Cells, Connectivity, and Computations.” PMC11653063. https://pmc.ncbi.nlm.nih.gov/articles/PMC11653063/

[PLACE-2] PMC12159490. “All active hippocampal pyramidal cells are place cells.” https://pmc.ncbi.nlm.nih.gov/articles/PMC12159490/

[PLACE-3] Moser et al. (2015). “Place Cells, Grid Cells, and Memory.” Cold Spring Harbor Perspectives. https://cshperspectives.cshlp.org/content/7/2/a021808.full.pdf

[PLACE-4] PMC7754708. “Targeted Activation of Hippocampal Place Cells Drives Memory-Guided Spatial Behavior.” https://pmc.ncbi.nlm.nih.gov/articles/PMC7754708/

[DOOR-1] Scientific American. “Why Walking through a Doorway Makes You Forget.” https://www.scientificamerican.com/article/why-walking-through-doorway-makes-you-forget/

[DOOR-2] PMC9789331. “Contextual inference in learning and memory.” https://pmc.ncbi.nlm.nih.gov/articles/PMC9789331/

[DOOR-4] Pettijohn & Radvansky (2021). “Doorways do not always cause forgetting: a multimodal investigation.” BMC Psychology. https://link.springer.com/article/10.1186/s40359-021-00536-3

0.15.20 Altered States Neuroscience (Section 7.6)

[TIME-1] Jording et al. (2023). “The Feeling of Time Passing Is Associated with Recurrent Sustained Activity and Theta Rhythms.” Brain Connectivity. https://www.liebertpub.com/doi/10.1089/brain.

2023.0010

[TIME-2] Stetson et al. (2007). “Does Time Really Slow Down during a Frightening Event?” PLoS ONE. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0001295

[EGO-1] Carhart-Harris et al. (2016). “Neural correlates of the LSD experience revealed by multimodal neuroimaging.” PNAS. https://www.pnas.org/doi/10.1073/pnas.1518377113

[EGO-2] Sheline et al. (2009). “The default mode network and self-referential processes in depression.” PNAS. https://www.pnas.org/doi/10.1073/pnas.0812686106

[EGO-4] Coppola et al. (2022). “Mindfulness Meditation Increases DMN, Salience, and Central Executive Network Connectivity.” Scientific Reports. https://doi.org/10.1038/s41598-022-17325-6

[DRUG-1a] Carhart-Harris & Friston (2019). “REBUS and the Anarchic Brain.” Pharmacological Reviews. https://pharmrev.aspetjournals.org/content/71/3/316

[DRUG-1b] Carhart-Harris et al. (2014). “The entropic brain: a theory of conscious states.” Frontiers in Human Neuroscience. https://www.frontiersin.org/journals/human-neuroscience/articles/10.3389/fnhum.2014.00020/full

[STATE-1] Weber & Hornung (2021). “The Neuroscience of the Flow State.” Frontiers in Psychology. https://doi.org/10.3389/fpsyg.2021.645498

[STATE-2] Gallego-Molina et al. (2025). “Attention and meditative development.” NeuroImage. https://doi.org/10.1016/j.neuroimage.2025.121602

[STATE-3] Martial et al. (2025). “A neuroscientific model of near-death experiences (NEPTUNE).” Nature Reviews Neurology. https://doi.org/10.1038/s41582-025-01072-z

0.15.21 Criticality and Self-Organization (Section 4.3.1, 10.2.3)

[45] Beggs, J. M., & Plenz, D. (2003). “Neuronal avalanches in neocortical circuits.” *The Journal of Neuroscience*, 23(35), 11167-11177. https://www.jneurosci.org/content/23/35/11167 Foundational work demonstrating power-law distributions in cortical networks. Established branching ratio σ ≈ 1.0 as the hallmark of criticality in biological neural networks.

[46] Scholarpedia. “Neuronal avalanche.” http://www.scholarpedia.org/article/Neuronal_avalanche Comprehensive review of neuronal avalanche phenomena, avalanche size distributions, and criticality sig-

natures in neural systems.

[47] Peng, C. K., et al. (2012). "Detrended fluctuation analysis." *Frontiers in Physiology*, 3:450. https://www.frontiersin.org/articles/10.3389/fphys.2012.00450/full DFA methodology for detecting long-range correlations in physiological time series. DFA exponent α ≈ 1.0 indicates pink noise (1/f) characteristic of critical systems.

[48] Fontenele, A. J., et al. (2021). "Avalanches and edge-of-chaos are distinct phenomena." *Nature Communications*, 12, 4211. https://www.nature.com/articles/s41467-021-24260-z Critical distinction: avalanche criticality (neuronal cascades) and edge-of-chaos criticality (computational dynamics) are separate phenomena that do not necessarily co-occur.

[49] Gollo, L. L. (2018). "Critical synchronization and 1/f activity in inhibitory/excitatory networks." *Scientific Reports*, 8, 1074. https://www.nature.com/articles/s41598-018-37920-w Power spectrum analysis showing β ≈ 1-2 (pink noise) at criticality. Demonstrates relationship between synchronization and scale-free dynamics.

[50] Legenstein, R., & Maass, W. (2007). "Edge of chaos and prediction of computational performance for neural circuit models." *Neural Networks*, 20(3), 323-334. https://pubmed.ncbi.nlm.nih.gov/17517489/ Reservoir computing achieves best performance near critical point. Computational complexity maximized at phase transition between order and chaos.

[51] Wilting, J., & Priesemann, V. (2018). "Between perfectly critical and fully irregular: A reverberating model captures and predicts cortical spike propagation." *Cerebral Cortex*, 29(6), 2759-2770. https://pmc.ncbi.nlm.nih.gov/articles/PMC6871218/ Cortical spike propagation analysis using branching ratio and related criticality metrics (κ index). Empirical measurements of cortical criticality.

[52] Hesse, J., & Gross, T. (2014). "Self-organized criticality as a fundamental property of neural systems." *Frontiers in Systems Neuroscience*, 8, 166. https://www.frontiersin.org/journals/computational-neuroscience/articles/10.3389/fncom.2021.611183/full Review of self-organization mechanisms leading to criticality in neural networks, including activity-dependent rewiring and homeostatic plasticity.

0.15.22 Implementation

[43] DANEEL Reference Implementation. 559 tests, resilience module, Phase 1 validated. https://github.com/royalbit/daneel

[44] ADR-036: Phase 1 Stability Validation - Empirically Proved. https://github.com/royalbit/daneel/blob/main/docs/adr/ADR-036-phase1-stability-validation.md

[45] ADR-035: VolitionActor - Free-Won't Implementation. https://github.com/royalbit/daneel/blob/main/docs/adr/ADR-035-volition-actor-free-wont.md

---

Author: Luis Cezar Menezes Tavares de Lacerda (Louis C. Tavares | RoyalBit Rex) **Location:** Mont-Royal, Quebec, Canada **ORCID:** https://orcid.org/0009-0005-7598-8257 **LinkedIn:** https://www.linkedin.com/in/lctavares **GitHub:** https://github.com/royalbit | https://github.com/lctavares

Date: December 17, 2025

Qowat Milat — The way of absolute candor.