BRAIN: A First-Principles Blueprint for Cognitive Agents

Ana Paola Oviedo Salgado

August, 2025

Abstract

In an era dominated by frameworks, prepackaged architectures, and boilerplate engineering, I decided to develop my agent from the very root: BRAIN, a first-principles blueprint for cognitive agents. BRAIN is defined purely from formal agent theory, emphasizing foundational constructs such as percepts, policies, state representations, and ethical constraints, rather than ad-hoc software patterns. We formalize BRAIN as an agent $\langle S, A, P, \pi, R, M \rangle$, explicitly grounding its behavior in theoretical decision-making models and memory persistence mechanisms. This paper outlines the mathematical underpinnings of each module, proposing a hybrid approach that combines rule-based reasoning with learned policy components. Beyond technical detail, BRAIN serves as a philosophical statement: a return to first principles as the true foundation for engineering intelligence.

1 Introduction

My journey into Computer Science underwent a profound transformation after encountering the Theory of Computation. It was then that my perception shifted from viewing the discipline as a mere collection of frameworks and libraries to understanding it as the rigorous study of intelligence, rationality, and computation itself. This fundamental realization ignited a commitment to ground my work and studies firmly in the enduring principles of theory, mathematics, and the foundational concepts that define this discipline. In today's rapidly evolving technological landscape, the abundance of tools and frameworks often enables the construction of impressive systems without a deep engagement with their underlying foundations. However, I hold a conviction in the pursuit of truth and rigorous logic. I believe that true understanding in Computer Science transcends superficial implementation, revealing a profound connection to universal truths that unify mathematics, philosophy, epistemology, and even neuroscience. While technological trends may emerge and recede, these first principles remain immutable, serving as the timeless bedrock of our field. To reduce Computer Science solely to the development of routine applications feels antithetical to its essence as the polymath of all disciplines. By venturing beyond the confines of prebuilt solutions, we unlock the true potential of our work, transforming it from a shallow, binary exercise into a quest for deeper understanding. This philosophical conviction directly underpins the design and development of BRAIN. Far from being simply another automation script, BRAIN is engineered to embody the very essence of intelligence: perception, memory, decision-making, and rational action. It is conceived as a model-based, hybrid reflex-agent, specifically wired to interact with clients and facilitate car insurance renovations. Rooted entirely in first principles, this paper formalizes BRAIN's architecture, elucidating its mathematical underpinnings and its commitment to foundational cognitive engineering.

2 Formal Agent Design of BRAIN

Before jumping into implementation, it is essential to ground BRAIN in rigorous agent formalism. Coding without a theoretical foundation can lead to fragile or ad-hoc solutions that lack coherence and transparency.

An *agent* is any entity that perceives its environment through sensors and acts upon it through actuators to achieve certain objectives. In the context of BRAIN, the formal components can be described as follows:

- **Agent**: BRAIN itself, conceived as a software program with cognitive capabilities.

- **Environment**: Incoming WhatsApp messages, operational workflows, and external APIs that BRAIN interacts with.

- **Percepts**: Messages received from users via WhatsApp, including all associated metadata.

- **Sensors**: The NLP pipeline that allows BRAIN to interpret and extract meaning from percepts, as well as the memory engine that provides contextual awareness and historical understanding.

- **Actions**: Outgoing responses delivered through API calls or webhooks to the main platform or user interface.

- **Actuators**: The mechanisms through which BRAIN transmits its chosen actions back into the environment, typically through structured API responses.

Grounding BRAIN in this formal agent framework ensures clarity, promotes reproducibility, and aligns the system with first-principles thinking.

3 Formal Agent-Theoretic Model of BRAIN

To rigorously define BRAIN's operational paradigm, we formalize it as a six tuple $\langle S, A, P, \pi, R, M \rangle$, where each component represents a fundamental aspect of intelligent agency.

- S: **State Space**. The set of all possible internal configurations of BRAIN, representing its current understanding of the car insurance renovation dialogue and context. A state $s \in S$ is a comprehensive representation derived from processed percepts and accumulated memory. Formally, $s = (p', m)$, where p' is the processed percept (e.g., extracted intent and entities) and m is the current memory state (e.g., filled slots, policy details).
- A: **Action Space**. The set of all possible actions BRAIN can perform in the context of car insurance renovation. Each action $a \in A$ is a discrete operation executable via actuators, influencing either the customer interaction or internal state. Examples include `ask_for_policy_number`, `provide_renovation_quote`, `confirm_policy_update`, `escalate_to_human_agent`, `request_vehicle_details`.
- P: **Percept Space**. The set of all possible raw inputs (percepts) that BRAIN receives from its environment. A percept $p \in P$ is an observation at a given time t, such as a WhatsApp message string related to insurance renovation along with metadata (sender, timestamp, etc.). For example: *"Hey, I want to renew my car insurance for policy number XYZ123."*
- π: **Policy**. BRAIN's decision-making function, mapping states to actions: $\pi : S \to A$. In BRAIN's hybrid architecture, π is composed of a rule-based component π_R and a learned component π_L, orchestrated by the decision engine. Thus, $\pi(s) = \texttt{DecisionEngine}(\pi_R(s), \pi_L(s))$. This policy determines the most appropriate response or internal operation for an insurance renovation query.

- R: **Reward Function**. A scalar function $R : S \times A \to \mathbb{R}$ that quantifies the desirability of performing an action a in a state s. In the insurance domain, rewards might be tied to successful policy renovations, increased customer satisfaction scores, reduced escalation rates, and overall conversational efficiency.

- M: **Memory**. Represents BRAIN's persistent context across interactions. Formally, M is a dynamic structure that stores historical states, user-specific slots (e.g., policy numbers, vehicle details), and decision traces. It enables BRAIN to conduct coherent multi-turn conversations and maintain continuity over time.

By defining BRAIN as $\langle S, A, P, \pi, R, M \rangle$, we establish a rigorous agent-theoretic foundation that bridges theoretical decision models and practical conversational AI design.

4 Perception and State Representation

To interact meaningfully with its environment, BRAIN must perceive external stimuli and convert them into structured internal representations. This section defines the formal structure through which BRAIN interprets and reasons about the world.

4.1 Percept Space and Processed Percepts

BRAIN does not perceive the world directly, but through structured representations of raw events.

Let $\mathcal{P}$ denote the **percept space**, the set of all possible raw inputs p_t at time t, each representing a user's message and its context:

$$p_t = (U_t, T_t, C_t, M_t)$$

where:

- U_t is the raw text utterance.
- T_t is the timestamp.
- C_t is the communication channel (e.g., WhatsApp).
- M_t is metadata (e.g., user ID, language).

These raw percepts are passed through BRAIN's primary sensor, the NLP pipeline. This sensor function

$$S_{\text{sensor}} : \mathcal{P} \to \mathcal{P}'$$

produces a **processed percept**:

$$p'_t = S_{\text{sensor}}(p_t) = (\text{Intent}_t, \text{Entities}_t)$$

where the user intent and relevant entities are extracted for reasoning.

4.2 State Representation

To make rational decisions, BRAIN maintains a state s_t at each timestep t, which integrates both perception and memory:

$$s_t = (p'_t, m_t)$$

Here:

- p'_t is the most recent processed percept.
- m_t is the current memory state, encapsulating past knowledge and session context.

This state s_t provides a sufficient statistic for decision-making, capturing what BRAIN currently perceives and remembers.

For a detailed discussion of memory architecture and update mechanisms, see Section 6.

5 Policy and Decision-Making Formalism

The policy π represents the core of BRAIN's intelligence—the mapping from states to actions that embodies rational decision-making. Unlike monolithic approaches that rely solely on learned behaviors or rigid rule systems, BRAIN employs a **hybrid policy architecture** that combines the interpretability of rule-based reasoning with the adaptability of learned components.

5.1 Hybrid Policy Composition

BRAIN's policy π is formally decomposed as:

$$\pi(s) = \text{DecisionEngine}(\pi_R(s), \pi_L(s), C(s))$$

where:

- $\pi_R : S \to A$ is the **rule-based policy** encoding domain expertise and explicit logical constraints
- $\pi_L : S \to \Delta(A)$ is the **learned policy** producing a probability distribution over actions
- $C : S \times A \to \{0, 1\}$ is the **constraint function** ensuring ethical and safety compliance
- DecisionEngine orchestrates these components through a priority-based selection mechanism

5.2 Rule-Based Policy π_R

The rule-based component encodes explicit domain knowledge through logical predicates:

$$\pi_R(s) = \arg\max_{a \in A} \sum_{i=1}^{n} w_i \cdot \mathbb{I}[\text{Rule}_i(s) \implies a]$$

where $\text{Rule}_i(s)$ are logical conditions and w_i are rule weights encoding priority.

Example rules in the insurance domain:

- $\text{Rule}_1(s)$: If $\text{Intent}(s) = \text{policy\_inquiry} \wedge \text{PolicyNumber}(s) = \emptyset$, then $a = \text{ask\_policy\_number}$
- $\text{Rule}_2(s)$: If $\text{Intent}(s) = \text{complaint} \wedge \text{Escalation}(s) = \text{required}$, then $a = \text{escalate\_human}$

5.3 Learned Policy π_L

The learned component employs a neural architecture trained on historical interaction data:

$$\pi_L(s) = \text{softmax}(f_\theta(s))$$

where $f_\theta : S \to \mathbb{R}^{|A|}$ is a neural network parameterized by θ, trained to maximize expected cumulative reward:

$$\theta^* = \arg\max_\theta \mathbb{E}_{(s,a)\sim\mathcal{D}}[R(s,a) \cdot \log \pi_L(a|s)]$$

5.4 Decision Engine Integration

The DecisionEngine implements a hierarchical selection mechanism:

1. **Constraint Filtering**: Remove all actions a where $C(s,a) = 0$
2. **Rule Priority**: If $\pi_R(s) \neq \emptyset$ and confidence exceeds threshold τ_R, select $\pi_R(s)$
3. **Learned Fallback**: Otherwise, sample from $\pi_L(s)$ restricted to constraint-compliant actions
4. **Default Handling**: If no valid actions remain, execute safe default action a_{default}

This architecture ensures interpretability through rule-based reasoning while maintaining adaptability through learned components, with safety guaranteed by constraint enforcement.

6 Memory Model and Cognitive Persistence

An intelligent agent must not only perceive and reason, but also *remember*. BRAIN's memory model is designed to simulate human-like cognitive persistence by integrating both short-term and long-term memory systems.

Inspired by findings in neuroscience and cognitive science, we partition the agent's memory M into two interacting modules:

- M_W: **Working memory**, implemented via Redis.
- M_L: **Long-term memory**, implemented via PostgreSQL.

This dual system allows BRAIN to maintain an ongoing conversational context while preserving historical knowledge between sessions.

6.1 Hybrid Memory Architecture and Cognitive Persistence

The agent's memory M is partitioned into *working* and *long-term* memory. This design is inspired by findings in neuroscience and cognitive psychology, which distinguish between short-term (or working) memory - often associated with the prefrontal cortex - and long-term memory stored in more stable and distributed cortical networks. In humans, this separation enables fast, real-time conversational flow while retaining important information over time for learning and context continuity.

BRAIN reflects this biological architecture computationally: Redis acts as a working memory module, providing low-latency access to recent percepts, intents, and conversational slots to support fluid dialogue. PostgreSQL serves as the long-term memory system, preserving user profiles and immutable dialogue histories between sessions. This hybrid design allows BRAIN to reason in the moment while also exhibiting cognitive persistence across interactions, just as a human might recall a previous conversation from memory while engaging in a new one.

6.2 Formal Memory Update Function

We define the memory update function as:

$$\mu : (M, s, a) \to M'$$

where M is the current memory, s is the current state, a is the selected action, and M' is the updated memory.

Memory is updated in two parallel channels:

1. Redis (M_W): Updated at every turn with the latest session state (e.g., intent, extracted entities, status).

2. PostgreSQL (M_L): Updated with immutable logs and mutable profile attributes using an upsert mechanism.

This dual-path design provides fast contextual reasoning during interactions while enabling retrospective analysis, long-term personalization, and alignment auditing. Together, they support cognitive persistence—continuity of identity and memory across agent lifetimes.

7 Ethical Guardrails as Formal Constraints

Ethical behavior cannot be an afterthought—it must be formally integrated into the agent's decision-making process. BRAIN implements ethical guardrails through a **constraint function** that acts as a formal filter on the agent's action space.

7.1 Constraint Function Formalism

We define the constraint function $C : S \times A \to \{0, 1\}$ as:

$$C(s, a) = \prod_{i=1}^{m} C_i(s, a)$$

where each C_i represents a specific ethical or safety constraint. An action is permissible if and only if $C(s, a) = 1$.

7.2 Core Constraint Categories

Privacy Constraints (C_{privacy}):

- Never request unnecessary personal information
- Mask sensitive data in logs and memory
- Respect user consent boundaries

Safety Constraints (C_{safety}):

- Prevent harmful recommendations (e.g., invalid insurance advice)
- Avoid actions that could cause financial harm
- Maintain professional boundaries

Fairness Constraints (C_{fairness}):

- Ensure equitable treatment across demographic groups
- Avoid discriminatory language or decisions
- Provide consistent service quality

Transparency Constraints ($C_{\text{transparency}}$):

- Clearly identify as an AI agent when asked
- Explain reasoning for complex decisions
- Acknowledge limitations and uncertainties

7.3 Dynamic Constraint Evaluation

Constraints are evaluated dynamically at each decision point:

$$\text{ValidActions}(s) = \{a \in A : C(s,a) = 1\}$$

The policy is then restricted to this valid action subset, ensuring that ethical violations are impossible by construction rather than relying on post-hoc filtering.

8 Reclaiming Intelligence: Why Cognitive Engineering Demands First Principles

8.1 The Crisis of Shallow Engineering

Modern software development has fallen into a trap of **shallow engineering**—building systems through composition of frameworks and libraries without understanding the underlying principles. This approach produces brittle systems that fail ungracefully, lack interpretability, and cannot adapt to novel situations.

In the realm of AI agents, this manifests as:

- **Framework Dependency**: Systems built on top of changing APIs and libraries that become obsolete
- **Black Box Reasoning**: Agents whose decision-making processes are opaque even to their creators
- **Ad-hoc Architecture**: Systems cobbled together without theoretical foundation or formal guarantees
- **Narrow Functionality**: Agents that work only within their training distribution

8.2 The First Principles Alternative

BRAIN represents a fundamentally different approach: **cognitive engineering** grounded in formal agent theory. This approach offers several advantages:

Theoretical Grounding: Every component of BRAIN is justified by formal agent theory, ensuring coherence and providing a foundation for reasoning about system behavior.

Interpretability: The hybrid policy architecture makes decision-making transparent, with clear rules and learned components that can be inspected and understood.

Robustness: By building on mathematical foundations rather than ephemeral frameworks, BRAIN's core architecture remains stable even as implementation details evolve.

Extensibility: The formal model provides a clear path for extending BRAIN's capabilities while maintaining theoretical consistency.

8.3 Beyond Implementation: A Philosophy of Intelligence

BRAIN is more than a technical system—it embodies a **philosophy of intelligence** that views cognition as the interplay of perception, memory, reasoning, and action. This perspective elevates AI development from mere engineering to a profound exploration of intelligence itself.

By grounding our work in first principles, we transcend the limitations of current tools and frameworks. We build systems that embody universal truths about intelligence, rather than artifacts of temporary technological constraints.

9 Open Theoretical Challenges

While BRAIN provides a solid foundation for cognitive agents, several theoretical challenges remain open for future research:

9.1 Compositional Reasoning

How can BRAIN's policy components be composed to handle complex, multi-step reasoning tasks that require coordination between rule-based and learned components?

9.2 Memory Consolidation

What formal mechanisms should govern the transfer of information from working memory to long-term memory? How can we ensure important experiences are preserved while preventing memory overflow?

9.3 Ethical Constraint Learning

Can ethical constraints themselves be learned from experience while maintaining safety guarantees? How do we balance adaptability with unwavering adherence to core principles?

9.4 Meta-Cognitive Capabilities

How can BRAIN develop awareness of its own reasoning processes and limitations? What formal frameworks support agent self-reflection and self-improvement?

9.5 Multi-Agent Coordination

How should multiple BRAIN agents coordinate when operating in shared environments? What theoretical foundations support collaborative cognitive agents?

These challenges represent opportunities for advancing both the theoretical understanding of intelligence and the practical capabilities of cognitive agents.

10 Conclusion

BRAIN represents a return to first principles in the design of cognitive agents. By grounding every component in formal agent theory—from the six-tuple specification $\langle S, A, P, \pi, R, M \rangle$ to the hybrid policy architecture and ethical constraint system—we have created a blueprint for intelligent systems that transcends the limitations of framework-dependent engineering.

The philosophical foundation of BRAIN extends beyond technical considerations to embrace a vision of Computer Science as the study of intelligence itself. In an age of rapid technological change, this commitment to first principles provides stability and direction, ensuring that our work contributes to lasting understanding rather than ephemeral solutions.

BRAIN demonstrates that rigorous theory and practical implementation can be unified in the pursuit of genuine artificial intelligence. The formal models presented here provide both a concrete system architecture and a foundation for future research into the nature of cognitive agents.

As we continue to develop BRAIN and explore its theoretical implications, we remain committed to the principle that true intelligence emerges not from the accumulation of frameworks and libraries, but from deep understanding of the mathematical and philosophical foundations that govern rational thought and action.

The journey toward artificial general intelligence requires more than incremental improvements to existing systems—it demands a fundamental reconsideration of what intelligence means and how it can be formally characterized and implemented. BRAIN takes a step in this direction, grounding cognitive engineering in the timeless principles that will endure long after today's frameworks have been forgotten.

References

[1] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, Prentice Hall, 2020.

[2] M. Wooldridge, *An Introduction to MultiAgent Systems*, Wiley, 2009.

[3] M. Sipser, *Introduction to the Theory of Computation*, 3rd ed., Cengage Learning, 2012.

[4] K. H. Rosen, *Discrete Mathematics and Its Applications*, 7th ed., McGraw-Hill Education, 2012.

[5] R. C. Atkinson and R. M. Shiffrin, "Human memory: A proposed system and its control processes," in *The Psychology of Learning and Motivation*, vol. 2, Academic Press, 1968, pp. 89–195.

[6] R. Bellman, *Dynamic Programming*, Princeton University Press, 1957.

[7] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed., MIT Press, 2018.

[8] D. Kahneman, *Thinking, Fast and Slow*, Farrar, Straus and Giroux, 2011.

[9] A. Newell and H. A. Simon, *Human Problem Solving*, Prentice-Hall, 1972.