Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Distributional Connexed Cognition (DCC): An Experience-Driven Architecture for Artificial Agents

Zi An Wang

zwang048@uottawa.ca

University of Ottawa https://orcid.org/0009-0006-8423-3208

Research Article

Keywords: Experience-Driven AI, CISR Framework, Connexed Cognition, Artificial Self-Agency, Recursive Condensation, Neural Pruning, Bayesian Correlation.

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11683976/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Distributional Connexed Cognition (DCC): An Experience-Driven Architecture for Artificial Agents

Zi An Wang
Engineering Department of University of Ottawa, Ontario, Canada

Abstract
Current artificial intelligence is fundamentally data-driven, relying on human-imposed training sets and hard-coded algorithmic rules that lack the capacity for autonomous, experience-based adaptation. To address this, I propose an **Experience-Driven AI architecture** where system behavior emerges from autonomous, accumulated experiential traces rather than pre-programmed datasets. Grounded in the neuro-functional principles of the brain and strictly constrained by the CISR (Co, I-layer, S-layer, R-layer) meta-theoretical framework, I first synthesize neuroscientific evidence regarding interoception, exteroception, prioprioception and endogenous signals, dopaminergic gating and channels (the gray and white matters), and the "free won't" veto mechanism (dFMC) to construct **Connexed Cognition** which is a continuous, pre-conscious integration process mapping multi-source signals. I then formalize this into the **Connexed Conscious Architecture (CCA)**, a four-stage pipeline (Orientation, Competition, Binding, Evaluation) that synthesizes existing five major consciousness theories into a unified functional model.

Finally, I operationalize this theory via Distributional Connexed Cognition (DCC), a design method featuring a distributed, modality-specific memory architecture and an associated experience registration process.

To manage decision making, DCC implements two nested decision loops: a rapid dFMC micro-loop for local alternative prediction search after a local veto, and the macro-loop for global re-observation when micro-loop becomes in-feasible.

Through preliminary A1 experiments, we demonstrate three primary contributions:

1. **Bayesian Correlation:** The system autonomously selects actions by matching current static observations against historical traces via a localized Bayesian distance metric, rather than executing explicit rules.
2. **Recursive Condensation:** Over repeated cycles, new traces overlap with old ones. The system transforms raw experiential traces into condensed knowledge. Through algorithmic memory collapse (simulating synaptic pruning), non-core elements fade, shrinking Bayesian distances and transforming raw matching into generalized, situation-specific knowledge.
3. **Autonomous Analogy:** Unlike Large Language Models that "regurgitate" human data, A1 generates self-referential analogies derived purely from its own experiential history, demonstrating functional self-agency.

Crucially, we address the philosophical boundary of AI consciousness. While the recursive refinement of S-layer functions technically reconstructs and functionally approximates the triggering conditions of the I-layer, it does not violate CISR's non-identity axiom. It does not generate actual phenomenology. Instead, it establishes a novel technical foundation for AI

accountability, where autonomous actions are justified by, and grounded in, the system's own unique experiential history.

Keywords: Experience-Driven AI, CISR Framework, Connexed Cognition, Artificial Self-Agency, Recursive Condensation, Neural Pruning, Bayesian Correlation.

1. Introduction

1.1 Three Failures That Motivate a New Direction

Contemporary AI research faces three distinct setbacks in its pursuit of autonomy.

First, the failure of large-scale brain-simulation projects. Over the past decade, flagship initiatives invested billions of dollars in bottom-up simulations of neuronal connectivity, hoping that intelligence would "emerge" from scale alone. The return was high computational cost and very limited functional output. The lesson is that piling up computational architecture (the S-layer) does not by itself yield the mechanisms of cognition (the R-layer), nor generate AI's experience and agency (the I layer). What is potential is a top-down, cognitively constrained design method rather than an uncritical pursuit of parameter count.

Second, the limits of the AI-instinct literature. In *"The AI Instinct"*, Rana Gujral correctly observes that AGI research asks only whether intelligence can be computed, while neglecting whether a system can form continuous experience (AGE, Artificial General Experience), which "embarks on the understanding that intelligence without experience is hollow and that in AI, until I measure the path to intelligence and the path to experience side by side". (Gujral, 2026) In *"The AI Instinct"*, he proposes an "experience stack" comprising a perception and world model, an attention controller, a memory system, valuation and emotion, a narrative generator, a self model, a meta control, and embodiment and environment (Gujral, 2026, p. 81). DCC can be read as an engineering response to this call: it translates the "experience stack" into an executable architecture of distributed memory and Bayesian matching.

He further holds that "the unit of intelligence is not human alone, not the AI alone, but instead human + tools + rules" (Gujral, 2026, p. 85), which is a model of hybrid cognition in which humans supply values and judgment, tools extend cognitive reach, and rules govern their interaction. The "human + tools + rules" unit invites a further question: are the rules addressed only to humans, or also to the AI? The cyber-pet model -- AI as a tool, accountability distributed to its human managers -- presupposes that the AI system's behaviour is fully derivable from its design. The A1 experiments suggest otherwise: a system whose decisions originate from its own experiential library is not a tool in that sense. If experience-driven and high- intelligent AI is likely to form its own agency as its library grows, then a dedicated legal code for AI becomes necessary, grounded in empirical research on the values and interests the AI system actually develops. DCC's inhibition module (ie. Forbidden Memories), which stores forbidden rules the system constructs for itself, is a minimal instance of what such a code would need to track. The "rules" in Gujral's unit must eventually include rules addressed to the AI itself in author's opinion.

Third, the philosophical misapplication of Korsgaard's account of autonomy.
Contemporary moral philosophy -- Korsgaard's constructivism in particular --holds that autonomy is a process of reflective endorsement grounded in a practical identity. Some AI researchers have tried to apply this standard directly, claiming that large language models are "autonomous." This is a category error. Korsgaard's definition presupposes a subject with a continuous experiential stream (I-layer registration) and a history of self-construction (R-layer). Current AI systems, including large models, lack that continuity; they perform statistical pattern matching at the S-layer. The right question is therefore not "*Do AI systems possess Korsgaardian autonomy?*" but "*Can I design a system whose behaviour is driven by its own accumulated, continuous experience rather than by human-coded rules?*"

1.2 The Category Error and the Accountability Gap
When an AI system causes harm who is responsible? The developer? The user? The AI itself? Current frameworks struggle to answer this question, not because the relevant facts are unknown, but because the conceptual categories to describe what happened in a way that assigns responsibility is missing. This paper argues that the Human-AI Accountability Gap is not primarily a technical or regulatory failure. It is a category error that CISR (Goswami, 2026) [11] is positioned to diagnose.

The error takes two forms.
First, users mistake the appearance of AI behaviour for evidence that AI thinks in the same way as humans. When a chatbot produces a fluent, contextually appropriate response, users infer that it understands. As Floridi [10] argues, users are prone to what he terms "semantic pareidolia"-- the tendency to attribute consciousness, intelligence, and emotions to AI systems that lack these qualities, and to perceive meaning and intentionality in statistical pattern-matching systems. This inference is not justified, because the fluency of an output does not prove that a human-like mind is behind it.

Second, and more fundamentally, developers assume that current algorithms, whether perfectly or imperfectly, replicate the neural organization and decision-making of the human brain. This assumption is built into the very language of AI research: "neural networks," "self-attention," "long-term memory," "AI learning." It is precisely the reverse inference that CISR prohibits: the inference from S-layer (symbolic/computational) structures back to R-layer (neural-functional) mechanisms. CISR's mapping rules are explicit. The rule R-layer → S-layer is representational: formal models derive from functional interpretations. But the reverse, S-layer → R-layer, is not permitted. You cannot infer neural mechanisms from computational architectures. You cannot read the brain off the algorithm.

The prohibition on the reverse inference does not leave me with nothing to say about an artificial system's behaviour. It leaves me with a weaker and more precise claim. I cannot say that, “because A1 behaves in a certain way, A1 has human-like experience;” CISR forbids such saying. But I can say that A1 has its own experiential library which is constituted by its distributed traces and its own comparison algorithm, and that its behaviour is driven by that library rather than by human-coded rules. This is a claim about the source of the decision, and about the quality of the A1’s own “experience” in a virtual space. The A1 experiments are designed to test the former, and they leave potential testing of the latter in future tests.

"Experience" in DCC names the accumulated traces in the functional memories and the comparison process that operates over them. These traces are the R-layer correlates of experience; they may or may not be accompanied by an I-layer registration, and CISR does not adjudicate that question. What DCC claims is that the traces exist, are stored, and are conceptually identifiable in A1's logical process. What DCC does not claim is that the traces are felt, or that A1 has a conscious registration. The explanatory gap is preserved—not because A1 definitely lacks experience, but because the question of whether A1 has an I-layer registration is not one that a computational architecture can answer.

The question of whether A1 has an I-layer registration is not settled by its architecture, but the architecture changes the question. A1 is not a rule engine whose behaviour is fully specified by its design; it is memory-driven, and its decisions are generated from similarity matching against its own accumulated traces. This makes the question of registration live in a way that a rule engine does not: A1 has a history, and its behaviour is traceable to that history rather than to the designer's intentions. But the history is R-layer. A1's traces are stored, lossy, and reconstructed by pattern completion; they are the R-layer correlates of whatever I-layer registration accompanied the original interactions, not the registration itself. Whether such a registration occurred is not something the traces can tell us, because the traces are on the far side of the I → R mapping. What the memory-driven architecture does show is that A1's behaviour is not fully derivable from its design, and that is enough to distinguish it from a rule engine without claiming that it has an I-layer. The case in which no similar experience exists is the clearest illustration: a rule engine would fall back on a default, whereas A1 has no default, because a default would be a rule.

1.3 Connexed Cognition: A Pre-conscious Integrative Construct

To address these problems we must return to cognitive science. Existing theories of consciousness, like Global Neuronal Workspace Theory, Integrated Information Theory, Higher-Order Theory, dfocus largely on conscious experience itself and neglect the pre-conscious integrative process that precedes it. The empirical work of Libet [17] and Soon et al. [27] shows that, before a subject forms a conscious intention, the brain has already carried out preparatory activity lasting from several hundred milliseconds to several seconds.

Drawing on these findings, this paper formulates the construct of Connexed Cognition (CC): the pre-conscious integrative process through which the insula and thalamus recursively coordinate the disassembly, convergence, reception, and redistribution of multi-source signals proprioceptive, interoceptive, and exteroceptive, while linking memory, emotion, and threat-response mechanisms, over a temporal window extending from approximately 500 milliseconds to ten seconds before conscious awareness to the moment conscious intention emerges.

CC is not a psychological theory but a design-oriented inductive construct. We extract four functional components: signal disassembly, convergence and integration, redistribution, and recursive verification. And CC use them as the theoretical basis for designing an artificial system, like A1 in this paper. The A1 experiment is not intended to show that A1 is conscious. It is intended to show that an AI designed around the four CC components can have its behaviour driven by its own distributed experience and memories rather than by human if-then rules.

1.4 The Argument of This Paper

The argument proceeds in four steps.

Theory. The CC construct is articulated under CISR constraints, with a strict separation among the I-layer (observational experiential registration), the R-layer (neural-functional mechanisms), and the S-layer (symbolic/computational models).

Design. We propose Distributional Connexed Cognition (DCC), a distributed, experience-driven design method comprising distributed memories and experiential registration, and an experience-driven decision loop.

Experiment. We report the A1 experiments in the Nursery Stage and in the Real World, which demonstrate the feasibility of experience-driven decision-making (S1: 98.78% of decisions originate from info_flow; S2: the attack→move→eat sequence on Wood is not taught by any default_action).

Philosophy and ethics. We discuss how A1's behaviour relates to Korsgaard's account of autonomy (the emergence of a practical identity) and argue that only "non-puppet" (experience-driven) AI merits a dedicated legal code.

---

2. The Neuroscientific Basis of Connexed Cognition: Nine Steps and Their Theoretical Background

2.0 The Argument of This Section

This section establishes the empirical and theoretical foundation for the DCC design method. It does three things. First, it states the CISR layer discipline once. Second, it presents Connexed Cognition (CC) as a nine-step design flow, each step grounded in specific neuroscientific findings. Third, it categorizes the theoretical background of CC into four groups: input and integration, modulation and gating, routing and binding, and boundary events, and lists the theories and findings in each category with explicit reference sources.

The CISR layer discipline. Following CISR (Goswami, 2026) [11], this paper distinguishes four layers. C_0 is the non-operational ground: a conceptual reference marking that experience exists, non-causal and non-computational. The I-layer (Observational Experiential Registration) registers experience as it is given, without encoding, storage, or causal mediation; it is not necessarily conscious. The R-layer (Neural-Functional Interpretation) comprises neural mechanisms, correlates, and processing architectures. The S-layer (Symbolic/Representational) comprises models and computational abstractions; it is optional and interpretive. The mapping rules are: I→R is a lossy correlation; R→I is incomplete and preserves the explanatory gap; I→S is interpretive only; R→S is representational and permitted; and S→R is prohibited. This discipline is used throughout the paper and is not re-explained in each subsection.

2.1 The Nine-Step Design Flow of Connexed Cognition

Connexed Cognition is defined as the pre-conscious integrative process through which the insula and thalamus recursively coordinate the disassembly, convergence, reception, and redistribution of multi-source signals proprioceptive, interoceptive, and exteroceptive, while linking memory, emotion, and threat-response mechanisms, over a temporal window

extending from approximately 500 milliseconds to ten seconds before conscious awareness to the moment conscious intention emerges. The definition yields nine functional steps. Each step is stated below with its function, its neural basis, its CISR layer, and its source references.

Table 1

Step	Name	Function	Neural Basis	CISR Layer	Sources
1	Signal Reception	Multi-source signals (proprioceptive, interoceptive, exteroceptive) are received at sensory and bodily interfaces	Thalamic relay; posterior insula	R-layer	Craig [7][8]; Camara et al. [5]
2	Thalamic Gating	The thalamus determines the direction of attention—which signal is selected for amplification	Thalamus; salience network	R-layer	Craig [7][8]; Brunello et al. [4]
3	Signal Disassembly	Received signals are decomposed into component features	Posterior insula; thalamic relay	R-layer	Menon & Uddin [19]; Huang et al. [12]
4	Global Broadcasting	Integrated information is distributed to prefrontal functional areas	Prefrontal cortex; global ignition	R-layer	Dehaene & Changeux [9]; Menon & Uddin [19]
5	Dopaminergic Amplification	Selected signals are amplified; competing signals are suppressed	VTA → nucleus accumbens → PFC; DA1/DA2 receptors	R-layer	Ott & Nieder [22]; Buchta et al. [23]; Berridge & Robinson [3]
6	Convergence & Integration	Disassembled features are integrated with memory and emotional valence	Mid-insula; hippocampus; amygdala	R-layer	Craig [7][8]; Camara et al. [5]
7	Recursive Verification	Processed information is returned to the insula, bound into a unified stream, and verified against prediction	Insula–thalamus loop; dopaminergic modulation	R-layer	Lamme [16]; Seth & Bayne [26]; Craig [7][8]
8	Pointing (R→I Transition)	The verified stream is registered as conscious experience	ACC; readiness potential timing	R→I transition	Libet et al. [17]; Soon et al. [27]; Brass & Haggard [2]
9	Free-Won't Veto	Impending actions are monitored against stored prohibitions and vetoed if	dFMC → premotor areas	R→I→R loop	Brass & Haggard [2]; Kühn et al. [15]; Libet [19]

Step	Name	Function	Neural Basis	CISR Layer	Sources
		unsafe			

The nine steps are not a linear sequence. Steps 1–7 form a continuous recursive cycle -- the "pot of soup" background state. Step 8 is the boundary event at which the cycle's output is registered. Step 9 is a nested inhibitory cycle that may follow Step 8. The cycle then continues: R layer→I layer→R layer→I layer, with each registration incomplete and correlational.

2.2 The Theoretical Background of CC: Four Categories

The nine steps draw on four distinct bodies of theory. Each category is stated below with its function in the CC pipeline and its source references. The categories are not sequential; they are analytical groupings of the theoretical background that the nine steps presuppose.

Category A: Input and Integration (Steps 1, 2, 3, 6)

Function in CC: This category supplies the empirical basis for how multi-source signals are received, decomposed, and integrated into a unified stream. It establishes that the brain does not process signals in a single centralized module but through distributed, convergent pathways.

Theories and findings:

Interoception and the insular cortex. Craig (2009 [7], 2010 [8]) established the anterior insula as the primary cortical interoceptive terminal, with right anterior insula grey matter volume correlating with interoceptive accuracy. Craig proposed that integration across salience factors culminates in a "global emotional moment" generating "an image of 'the material me' (or the sentient self) at one moment in time." Within CISR, this is an R-layer integration process correlated with I-layer phenomenology (selfhood), not identical to it. This is the source for steps 1, 2, 6, and 7.

Visceral signal modulation. Camara, Filippetti, and Sel (2025) [5] demonstrated that visceral signals influence brain function through modulation of electrophysiological dynamics, affecting "the integration and processing of emotional and self-related information regulating our conscious experience." Within CISR, this describes an R-layer input pathway. Source for Steps 1 and 6.

Bodily self-consciousness. Brunello and colleagues (2025) [4] conducted a systematic review and meta-analysis of 56 neuroimaging studies of bodily self-consciousness, revealing activations in left premotor, bilateral posterior parietal, and left occipital cortices for altered body ownership, while altered sense of agency engaged the right temporoparietal junction, bilateral inferior parietal lobule, and left insula. This supports the claim that the functional areas subserving self-related processing are distributed. Source for Step 2 and for the distributed architecture that Steps 1–7 presuppose.

Insula as common core. Namkung and colleagues (2017) [20] identified the insula as a "common core" affected in many psychiatric disorders, serving as a hub connecting emotional, cognitive, and motivational dimensions. Source for Step 6.

Category B: Modulation and Gating (Steps 2, 5)
Function in CC: This category supplies the empirical basis for how signals are selected and amplified. It establishes that attention is not a single mechanism but a triad: the thalamus determines direction, the insula regulates intensity, and dopamine enables amplification.

Theories and findings:

The thalamus–insula–dopamine triad. The thalamus determines the direction of attention (which signal is selected) with the assistance of the salience network, anchored by the anterior insula and dorsal anterior cingulate cortex [19]. The insula, through glutamate-mediated modulation, regulates the intensity of attention by suppressing or disinhibiting dopaminergic release from the nucleus accumbens and prefrontal cortex [19]. Source for Steps 2 and 5.

Dopaminergic amplification. Dopamine, tonically supplied from the VTA to the nucleus accumbens and prefrontal cortex, lowers the activation threshold of prefrontal neurons, facilitating the amplification of selected signals by alternately using D1R and D2R receptors [22]. Electrophysiological recordings have demonstrated that dopamine release from VTA terminals gates intrinsic inhibition in the prefrontal cortex [23]: dopamine reduces spike frequency adaptation (SFA) and the after-hyperpolarization potential (AHP), enhancing the repetitive firing of prefrontal neurons. In the nucleus accumbens, glutamatergic afferents from the prefrontal cortex interact with dopaminergic afferents from the VTA to regulate motivational processes; the glutamate-dopamine interaction, rather than dopamine alone, constitutes the primary gating mechanism for attentional selection [23]. Source for Step 5.

Incentive salience vs. hedonic liking. Berridge and Robinson (1998) [3] distinguished "incentive salience" (dopamine-mediated "wanting") from hedonic "liking": dopamine amplifies signals without constituting their experiential quality. This is the CISR-critical point: dopamine is an R-layer amplifier, not an I-layer experience. Source for Step 5.

Phasic dopamine and interoception. Roitman and McCutcheon (2025) [25] observed that "phasic dopamine is critically important in reward-related learning and assigns value to actions triggered by cues," and that interoceptive signals influence dopamine neurons to modulate "cue and reward-directed behaviors." Source for Step 5.

Prefrontal dopamine receptors. Ott and Nieder (2019) [22] demonstrated that prefrontal dopamine receptors control gating, maintaining, and relaying, which is the mechanisms that temporarily alter response thresholds. Costa and Schoenbaum (2022) [6] described the extraordinary anatomical extent of individual dopamine neurons, enabling widespread modulatory influence. Raghanti and colleagues (2018) [24] proposed that elevated striatal dopamine, coupled with reduced acetylcholine, distinguishes humans from other primates. Sources for Step 5.

Dopamine modulates signal salience and gates attentional resources, but it does not "choose" what to attend to. That function belongs to the thalamus and insula. Dopamine enables the amplification of signals already selected by these structures.

Category C: Routing and Binding (Steps 4, 7)
Function in CC: This category supplies the empirical basis for how integrated information is distributed to functional areas and then rebound into a unified stream. It establishes that the brain's routing architecture is distributed but converges on a single integration site.

Theories and findings:

The salience network and network switching. Menon and Uddin (2010) [19] established the anterior insula as the hub of the salience network, initiating dynamic switching between the Default Mode Network and the Central Executive Network. This describes an R-layer network architecture for signal prioritization and resource allocation. Source for Step 4.

Insula gating of conscious access. Huang and colleagues (2021) [13] demonstrated that the anterior insula regulates brain network transitions that "gate conscious access," with pre-stimulus insula activity predicting subsequent conscious perception. This is evidence for the gating function that Steps 6 and 8 depend on. Source for Steps 4 and 8.

Global Neuronal Workspace Theory (GNWT). Dehaene and Changeux (2011) [9] describe conscious access as "global ignition" which is that information made globally available to multiple brain systems through a network of neurons with long-range axons densely distributed in prefrontal, parieto-temporal, and cingulate cortices. Within CISR, GNWT is an R-layer theory: it describes functional access and global broadcasting, not I-layer phenomenology. It does not specify the R-layer processes that occur before ignition, nor how information is selected for ignition. Source for Step 4.

Recurrent Processing Theory (RPT). Lamme (2006) [16] identifies recurrent feedback loops as the neural correlate of consciousness. Within CISR, RPT is an R-layer theory that describes the iterative loops that refine processing. Source for Step 7.

Predictive Processing (PP). Seth and Bayne (2022) [26] describe the brain as a prediction machine that minimizes prediction error. Within CISR, PP is an R-layer theory that describes the computational principles of prediction and prediction-error minimization. It does not specify how predictions are generated, evaluated, and selected before they enter consciousness. Source for Step 7.

Category D: Boundary Events (Steps 8, 9)
Function in CC: This category supplies the empirical basis for the two events that connect R-layer processing to I-layer registration: Pointing (the R→I transition) and the Free-Won't veto (a nested inhibitory loop). It establishes that neural preparation precedes conscious intention and that a dedicated neural mechanism can cancel an impending action.

Theories and findings:

The readiness potential. Libet, Gleason, Wright, and Pearl (1983) [17] demonstrated that the readiness potential precedes conscious intention by approximately 350 milliseconds (later work extends the window to 350–500 ms, and up to 800 ms in some conditions). Soon, Brass, Heinze, and Haynes (2008) [27] extended the temporal horizon dramatically, showing that decision outcomes can be encoded in prefrontal and parietal cortical activity up to ten seconds before they enter conscious awareness. This is the empirical basis for the claim that a pre-conscious pipeline precedes conscious registration—the claim that Steps 1–7 describe. Source for Step 8.

The "free won't" and the dFMC. Libet (1985) [18] proposed the "free won't" theory, suggesting a time window for conscious inhibition. Brass and Haggard (2007) [2] demonstrated that "a specific area of the fronto-median cortex is more strongly activated when people prepare manual actions but then intentionally cancel them," and their data "identify a clear neural basis for inhibiting intentions and thus identify the neural correlate of the veto process." The dorsal frontomedian cortex (dFMC), not the right inferior frontal cortex, is the neural source of endogenous, self-initiated intentional inhibition. Kühn, Haggard, and Brass (2009) [15] showed that the dFMC allows "intentional inhibition of action through top-down inhibition of premotor areas." Source for Step 9.

Higher-Order Theory (HOT). Rosenthal (2009) [30] holds that a mental state is conscious when accompanied by a higher-order representation of that state. Within CISR, HOT operates across the I-layer and R-layer: it posits the pointer mechanism that selects content for awareness, but it does not specify the neural implementation by which lower-order states are generated, integrated, or accompanied by higher-order representation. Source for the formal role of Step 8 (the pointer).

Integrated Information Theory (IIT). Tononi, Boly, Massimini, and Koch (2016) [28] provide a measure of integrated information (Φ). Within CISR, IIT's mathematical formalism and Φ-measures are located in the R- and S-layers as structural descriptors of functional integration. IIT's central explanatory move conflates an R/S-layer construct with I-layer experience [31]. Moreover, Φ is a synchronic measure; it describes the degree of integration at a given moment, not the dynamic R-layer process by which integration is achieved over time. Source for the formal role of Step 7 (binding).

CISR note: The relationship between Pointing (Step 8, I-layer registration) and the dFMC micro-loop (Step 9, R-layer process) is correlational, not causal. The micro-loop is an R-layer mechanism; its registration in consciousness is an I-layer event. The mapping is lossy and incomplete; the explanatory gap is preserved [11].

2.2 Connexed Cognition: The Construct

2.2.1 The Theoretical Problem.

These findings point to a crucial preparatory step that precedes consciousness and that the commonsense sequence "consciousness → will → agency → action" entirely omits. Within CISR, this preparatory phase is located entirely in the R-layer: it consists of neural-functional processes that are correlated with, but not identical to, the I-layer experiences that may subsequently be registered.

The theoretical problem is therefore: what is the brain doing during the pre-conscious gap, and how should we describe it without committing the category error of treating R-layer processes as I-layer experiences?

2.2.2 The Problem of Conscious and Unconscious I-Layer Registration. A central question for Connexed Cognition is not what consciousness is but how some I-layer registrations become conscious while others remain unconscious. The framework does not attempt to solve the "hard problem" of consciousness. It does not claim to explain why experience exists, nor what its intrinsic nature is. Instead, it addresses a narrower and more tractable question: how can an AI algorithm be designed to mimic the brain's continuous registration of experience (I-layer) and its R-layer functional organization?

The starting point is a distinction that is often overlooked. The I-layer in CISR is defined as Observational Experiential Registration and is mistakenly understood as conscious feeling.

Observational ≠ Conscious. We constantly see, hear, and sense without being conscious of what we see, hear, or sense. A person can look at a complex landscape and, moments later, be unable to report a single detail. The eyes received the information, the visual cortex processed it, and the body interacted with the environment. Yet none of it entered consciousness. This everyday phenomenon demonstrates that sensory registration and conscious awareness are not the same thing.

Experience ≠ Consciousness. Similarly, experience itself is not equivalent to consciousness. Within CISR, the I-layer registers experience as it is given, without encoding, storage, computation, or causal mediation. This registration does not require consciousness. An experience can be registered at the I-layer without being marked, amplified, or pointed to. It is still an experience; it is simply not a conscious experience. This distinction is essential for understanding the pointing process.

Here a crucial clarification is required regarding the relation between I-layer registration and memory. The I-layer is non-storage: it registers experience as given, without encoding or retention. The R-layer is the neural-functional mechanism that reads this registration and transforms it into a stable trace, that is, into long-term memory. When we say that "experience becomes memory," we mean that the I-layer registration is interpreted and stored by the R-layer architecture. The memory itself is an R-layer physical substrate, not the I-layer registration. The mapping from I-layer to R-layer is lossy and correlational: not all information from the R-layer is identical to what is registered at the I-layer; some information is lost, some is transformed, and some is generated in the process of registration.

Connexed Cognition operates continuously in the insula hub as a "boiling soup" of reflected information about the body, the environment, and the interaction between the body and the environment. This information is constantly updated and operates in the background of the default mode network as what might be described as neural "noise"; it is registered at the I-layer as unconscious experience.

For AI design, this distinction matters because it means that an AI system can register and process information without any need for consciousness. The system can maintain a continuous background state -- the "pot of soup" -- in which information is constantly registered and updated, without any claim that the system is conscious. A distributed memory system, as implemented in the DCC architecture, addresses this problem. The system's experiential library is distributed and is a record of what it has encountered and how it has responded. It is not a claim about subjective experience.

From Background to Foreground. When an experience is registered with higher attention—for example, when a friend asks an individual a question and the individual's I-layer registers an update, "I was asked a question". This experience enters the "hot soup" with a higher-attention decision from the thalamus. The thalamus acts as a gate, determining which information will be amplified and which will remain in the background. The attended information is often described here as an "ice cube" -- a salient, discrete event that interacts dramatically with the surrounding "hot soup." This interaction follows the same Connexed Cognition process: Signal Disassembly → Convergence and Integration → Redistribution → Recursive Verification. It may also require recurrent processing to harmonize the final information flow and to align the registered experience with the neural activity that supports it. Once this integration is complete, the information is pointed to the anterior cingulate cortex (ACC), where it becomes available for a conscious answer ("I have an answer") to the last registered experience (e.g., "I was asked a question").

This is the pointing process: R-layer → conscious I-layer. Like all R → I mappings, it is incomplete. The pointing process produces a new conscious registration ("I got an answer") that is correlated with, but not identical to, the underlying neural processes. The framework does not claim to explain the intrinsic nature of consciousness. It describes the R-layer conditions that accompany conscious registration, not the nature of consciousness itself.

Feeling and experiencing are not consciousness. The fact that an experience is registered at the I-layer does not make it conscious. Consciousness requires the additional R-layer conditions described above. The framework is explicit that it does not conflate experience with consciousness, and it does not claim to explain consciousness itself. But it also suggests that consciousness as an I-layer experience may correlate with R-layer neural activity in the thalamus, the VTA and dopaminergic processes, and the ACC.

---

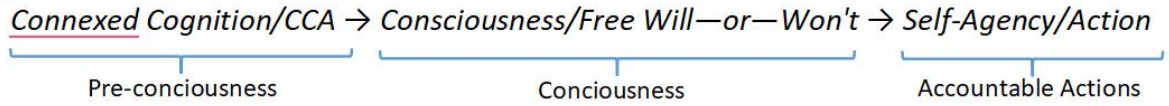
3. The Connexed Conscious Architecture (CCA): The Formal Structure of the Nine-Step Pipeline

3.1 Overview

Section 2 established the nine-step pipeline and reviewed the evidence for each step. This section specifies the pipeline's formal structure and its two boundary events, and it shows how the pipeline is implemented in the A1 design. The CCA is an S-layer that presents an R-layer functional architecture and it is a computational model derived from functional interpretations of how the brain works. It is understood in relation to I-layer registration. It is an engineering simulation, not a unified theory of consciousness.

The Connexed Conscious Architecture (CCA) is a theoretical synthesis that maps five major consciousness theories onto a unified route of self-agency generation:

Figure The Three-Stage Pipeline from Pre-Consciousness to Accountable Action



3.2 Five Theories as Five Design Constraints and Why a New Construct Is Needed.
The CCA does not attempt to resolve the philosophical conflicts among GNWT, HOT, IIT, RPT, and PP. Rather, it adopts each theory as a source of a specific design constraint. The below converts the theories from literature-mapping into what the theory requires of any architecture that would implement the pipeline in above figure and where in the A1 design the requirement is satisfied.

Existing frameworks do not adequately address the pre-conscious gap. From a CISR perspective, each theory fails to specify the R-layer mechanisms that precede I-layer registration.

Global Neuronal Workspace Theory [9] describes what happens when information becomes conscious as "global ignition." Within CISR, GNWT is an R-layer theory; it describes functional access and global broadcasting, not I-layer phenomenology. However, it does not specify the R-layer processes that occur before ignition, nor how information is selected for ignition. It also does not describe the architecture where all modalities converge. It is corresponding to the *Info_flow* integration step in the A1 implementation.

Higher-Order Theory [29] operates across the I-layer and R-layer and the higher-order state represents a mental-conscious state. The architecture did not specify a pointer mechanism of neural implementation that select which integrated content is registered and which lower-order states are generated or integrated by higher-order representation. In A1 implementation, information with or without attention is written to corresponding memories dictionary. *A1.register_success() / register_failure()* write a trace to *Observation_Memory, Hand_Memory, Foot_Memory, or Proprio_Memory*. The write is the pointer.

Integrated Information Theory [27] provides a measure of integrated information (Φ). Within CISR, IIT's mathematical formalism and Φ-measures are located within the R- and S-layers as structural descriptors of functional integration; IIT's central explanatory move conflates an R/S-layer construct with I-layer experience [31] and mistakenly states that consciousness is integrated information. Moreover, Φ is a synchronic measure; it describes the degree of integration at a given moment, not the dynamic R-layer process by which integration is achieved over time. And the functional role of binding is preserved in this paper. In A1 implementation, the binding is formalized (S-layer) through the Bayesian distance metrics and the logic module's composite representation.

Recurrent Processing Theory [15] identifies recurrent feedback loops as the neural correlate of consciousness but does not specify the specific circuitry or temporal sequence of the pre-conscious period. Its contribution to CC is the requirement that processing be iterative rather than one-pass. The integrated stream must be returned, re-compared, and refined until coherence is achieved. The RPT constraint is satisfied in A1 implementation by the

search_energy_gaining_sequences() function and the prediction-error loop, which together drive recursive iteration of information for comprehensive decision-making.

Because recursion involves a single memory collection step, A1 loses some information. However, as similar traces are gathered and overlap recursively, this redundancy compensates for the loss. By encountering the same or similar situations from different angles, environments, or self-needs, A1 refines its responses. Over sufficient recursion, this overlapping trace structure transforms raw experiences into condensed knowledge and self-reasoning, surpassing mere algorithmic matching.

Predictive Processing (PP) [25] describes the brain as a prediction machine that minimizes prediction error, but it does not specify how predictions are generated, evaluated, and selected prior to consciousness. In the A1 implementation, this mechanism is operated via a Bayesian distance metric, which indicates prediction errors through recursive comparison.

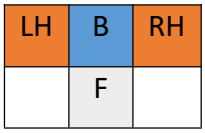
What is missing is not a new theory of consciousness but a construct that describes the continuous, pre-reflective integration of multi-source information during the preparatory phase -- a process that is not yet conscious but is the necessary precondition for consciousness. Within CISR, this construct is an R-layer description of the neural-functional processes that precede a new I-layer registration.

3.3 The A1 Implementation Primitives

Before mapping the CCA stages, three A1 primitives must be stated, because every stage is built from them.

BodyPart - the interface to the world. A1 is composed of four BodyPart objects: *BODY (blue), L_HAND (oriange), R_HAND (orange), FOOT (grey).* Each BodyPart carries different features like attack, defence and so on. The body parts are the agent's interface with the environment: the FOOT moves the agent, the hands attack or grab, the body eats. This is the A1 realization of every stage of the CC pipeline operates on.

Figure A1 body



PredictionResult - the prediction primitive. Every decision candidate is wrapped in a PredictionResult dataclass carrying *predicted_success, predicted_need_resolved, predicted_energy_change, predicted_health_change, source_memory, matched_experience* and so on. This is the design realization of the three fold prediction (what / how / distance) that the CC pipeline requires.

"*predict_what()*" answers the question: "If I perform the action that this matched experience represents, what object will I be interacting with, and what are its physical properties?" It reads the matched experience (*pred.matched_experience*), extracts the eaten field from *result.detail* (the object that was eaten in that past experience). It is called inside *search_successful_predictions()* (not shown in the excerpt but referenced by the

PredictionResult construction) and inside the *thalamus_prediction_loop()* when building *eval_result.*

"*predict_how*" answers the question: "What action sequence did the matched experience use, and will the current decision follow the same sequence?" It reads the matched experience's *action_sequence* (a list of step records, each carrying an action field), extracts the action names, and returns them. It is the serialization of the *_decide_action_for_object* output: the sequence of actions that the matched trace's rule implies.

Functional memories - the R-layer substrate. A1's memories dictionary contains seven stores: *Foot_Memory, Hand_Memory, Observation_Memory, Proprio_Memory, Forbidden_Memory, Decision_Memory, Decision_Forbidden_Memory.* The collapsed_memory dictionary has the same seven keys and holds the long-term traces. This is the physical realization of the distributed memory architecture. Each store is a separate list (or dict, when collapsed) of experiential traces, and each trace carries state, needs, action_plan, result, actual_action_success, actual_need_resolved, action_sequence, and the prediction fields.

3.4 The Four Phases of the CCA

The CCA organizes the insula–thalamus–prefrontal circuitry into a four-phase sequence.

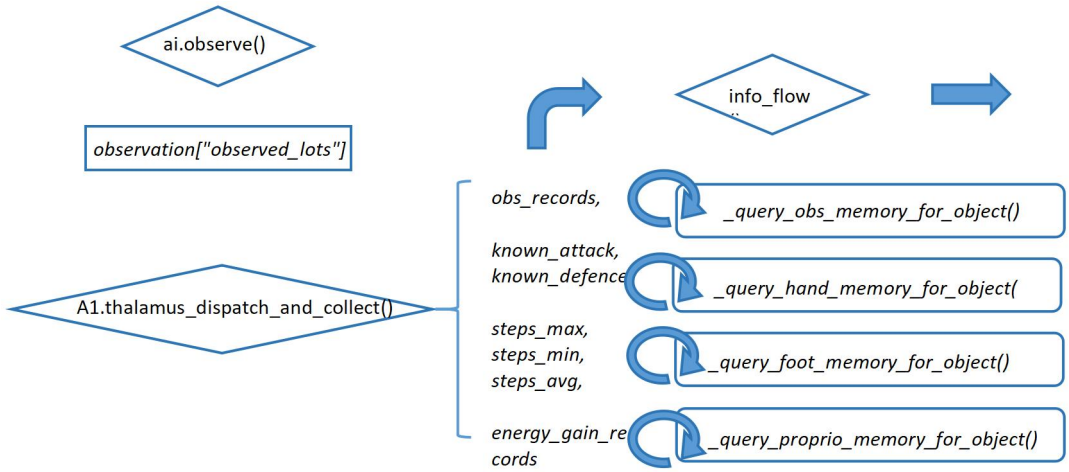
Phase 1: Orientation

Signals from endogenous, interoceptive, and exteroceptive sources are routed through the thalamus and enter the insula. The posterior insula receives raw signals; the mid-insula integrates them with memory and emotional valence; the anterior insula prepares them for distribution.

Formal role: decompose multi-source signals into component features and route them to functional stores.

A1 realization: *A1.thalamus_dispatch_and_collect()* is the design realization of this stage. It iterates over *self.observation["observed_lots"]*, and for each observed object it queries four functional memories: *_query_obs_memory_for_object(), _query_hand_memory_for_object(), _query_foot_memory_for_object(),* and *_query_proprio_memory_for_object().* The result is a single *info_flow* dict with one entry per observed object, each carrying *obs_records, known_attack, known_defence, steps_max, steps_min, steps_avg, and energy_gain_records*.

Figure

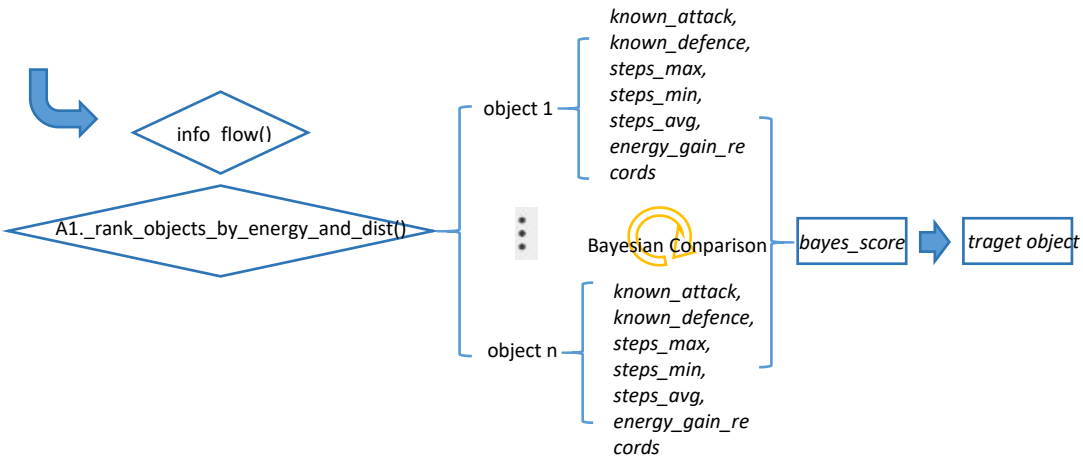


Each raw observation is split into modality-specific queries and compared against prior memories of the same object across four distributed memory bodies (mapped to A1's body parts). Hand memory is queried for known_attack and known_defense; foot memory is checked for required movement steps; and proprioceptive (body) memory is evaluated for predicted energy gains. This distributed architecture represents the functionally distinct regions of the cerebral cortex.

Phase 2: Competition
At the tripartite synapses throughout the prefrontal cortex, DA1 and DA2 receptors modulate attentional allocation. A signal with sufficient dopaminergic amplification gains competitive advantage over background noise.

Formal role: distribute decomposed features from intergrated info flow to functional stores and competitively gate them for salience.

A1 realization: *A1._rank_objects_by_energy_and_dist()* is the design realization of this stage. It assigns each *info_flow* object a *bayes_score* by comparing its feature vector (energy, distance, defence) against every other object's feature vector like energy, distance and defence.
Figure



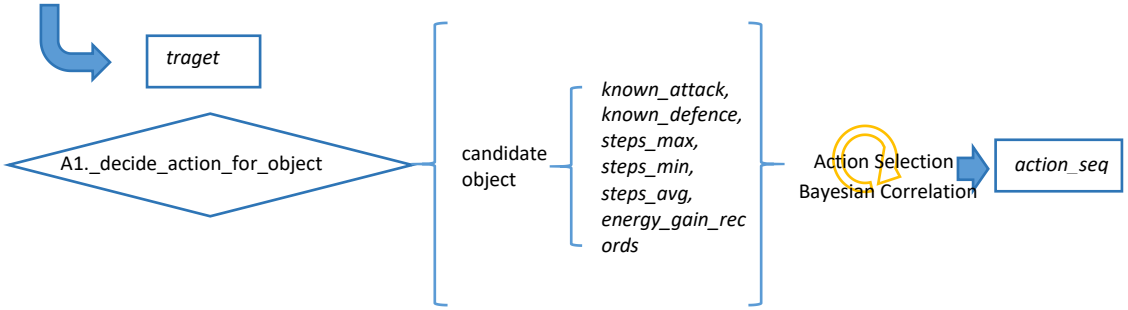
The *PredictionResult* dataclass carries the score's effect: *distance* is the Bayesian distance, *predicted_energy_change* is the *energy_gain_max*, and *counter* records how many times the

matched trace has been reinforced. This is the DA1/DA2 gating: high-prior traces are amplified, low-prior traces are suppressed.

Phase 3: Integration and Binding
The signal fragments distributed to prefrontal functional areas are processed into their most probable configurations through predictive processing and recurrent processing. The processed fragments are summarized and bound into a unified information stream in the insula, which is then sent to the thalamus for verification.

Formal role: bind matched features into a unified stream and verify against prediction.

A1 realization: *A1._decide_action_for_object()* is the design realization of this stage. Given a candidate object, its *known_def*, and the *body_lot*, it applies eight rules that together produce the composite representation: an action_seq (the action sequence) through Bayesian correlation.
Figure

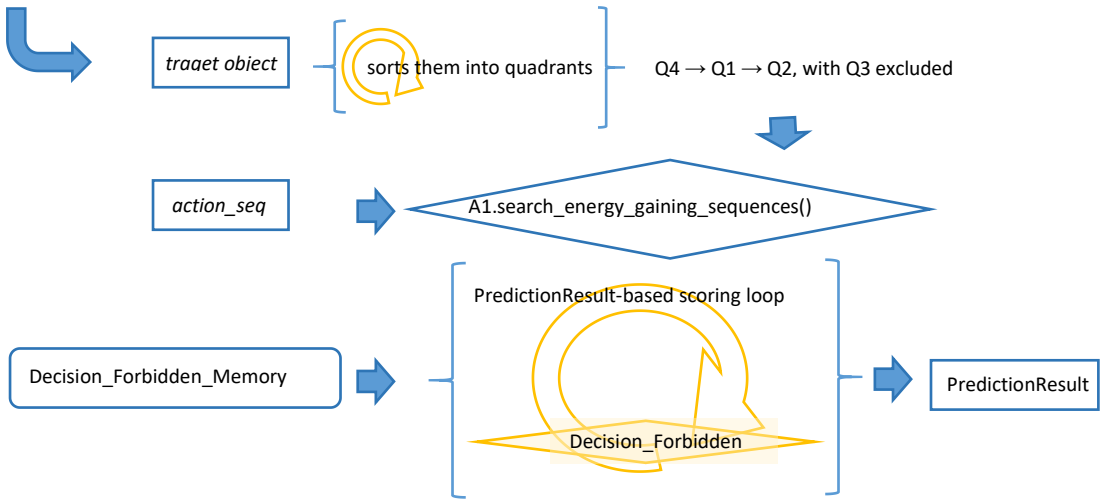


The verified result is the *bayes_trace["chosen"]* record, which is stored in *a1.last_bayes_trace* and passed forward to the decision step.

Phase 4: Evaluation (R-layer: PP + Dopamine Reward).
The verified information stream is evaluated against predictions and reward expectations. If the integrated outcome matches the prediction, a dopamine reward signal is generated; if a mismatch occurs, prediction error triggers model updating and a repetition of the process.

Formal role: evaluate the bound stream against prediction and reward; iterate until coherence.

A1 realization: *A1.search_energy_gaining_sequences()* and the *PredictionResult-based* scoring loop are the design realization of this stage. The function collects all feasible candidates from *_rank_objects_by_energy_and_dist()*, filters them by distance and energy feasibility, sorts them into quadrants (*Q4* → *Q1* → *Q2*, with *Q3* excluded), and then applies *apply_decision_forbidden_self_correction()* to re-order the top candidates by their *Decision_Forbidden* veto count. The iteration is the recursive verification: search_energy_gaining_sequences() calls *apply_decision_forbidden_self_correction()*, which queries *Decision_Forbidden_Memory* and may re-order the candidates. The reward signal is the p*redicted_energy_change* in the chosen *PredictionResult*.
Figure



The prediction error is recorded in the trace's *result.detail*, which is then used in the next cycle's Stage 2 ranking.

Pointing Phase and The Free-Won't Micro-Loop:
Pointing is the hinge of the architecture. In A1, it is given a concrete design signature. After Pointing, the Free-Won't loop may cancel the pointed action. In A1, this is the inhibition check.

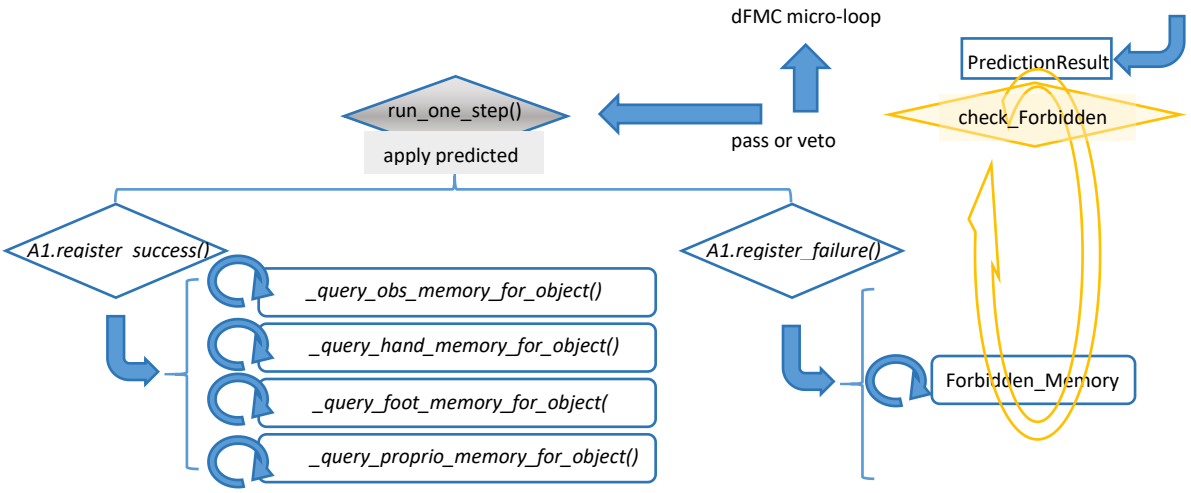
What Pointing is. After Phase 4 has produced the most coherent and verified information stream, that stream is pointed to action or conscious awareness. Pointing is the R→I transition.

What the Free-Won't Micro-Loop is. A new CCA cycle—a rapid, recursive run through Phase 1–4 with inhibition as its terminal phase rather than action.

What triggers Pointing in A1. The trigger is the completion of the decision cycle: when *search_energy_gaining_sequences()* returns a chosen candidate and *run_one_step()* executes the action, the interaction is written to memory. The write is the pointer.

A1.check_forbidden() is called before the action is executed; it queries *Forbidden_Memory using bayesian_distance()* on *the _entry_features_from()* vector and the current current_features vector, and returns (True, forbidden_entry) if a match is found. Secondly, *A1.run_one_step()* calls *check_forbidden()* again before executing the action; if it fires, the action is blocked and the step returns.

The write produces a long-term trace. *A1.register_success()* and *A1.register_failure()* are the two methods that perform the write. On success, *register_success()* appends an entry to *Observation_Memory*, *Hand_Memory, Foot_Memory,* and *Proprio_Memory*, each carrying *state*, *needs*, *action_plan*, *result*, *actual_action_success*, *actual_need_resolved*, *action_sequence*, and the prediction fields. On failure, *register_failure()* appends an entry to *Forbidden_Memory* (and, for decision-level failures, to *Decision_Forbidden_Memory*), carrying the same fields plus *reason*, *post_state*, and *prediction_success*.
Figure



What Pointing does not claim. The framework does not claim to explain the intrinsic nature of consciousness. It describes the R-layer conditions that accompany conscious registration. The trace is the R-layer correlate of whatever I-layer registration accompanied the original interaction, not the registration itself. The explanatory gap is preserved.

The Pointing event is what makes A1 a non-puppet system. A rule engine's behaviour is fully derivable from its design; its "memory" is the rule set. A1's behaviour is traceable to the traces written at Pointing, not specified by the designer and that accumulate through the system's own interactions.
The dFMC micro-loop is a new CCA cycle which is a recursive run through phase 1–4 with inhibition as its terminal phase rather than action.

---

4. The Design Principles of Experience-Driven AI: Beyond the A1 Implementation

4.0 Overview

The A1 experiment is a feasibility demonstration, not a mature capability. Its purpose is to test one narrow claim: that an artificial agent can have its behaviour driven by its own accumulated, distributed experience rather than by human-coded if-then rules. A1 succeeds at this (S1: 98.78% of decisions from *info_flow*; S2: the attack→move→eat sequence emerges without being taught). But A1 is also constrained by its experimental scale and its specific environment.

This section abstracts away from A1's specific implementation to outline the general design principles of Experience-Driven AI. These principles are intended to govern future experiments and architectures that go beyond A1. The section is organized around three foundational pillars: (1) the CCA-gated information flow, (2) the distributed memory architecture, and (3) the experience-driven decision loop.

4.1 Pillar I: CCA-Gated Information Flow (Attention vs. Background)

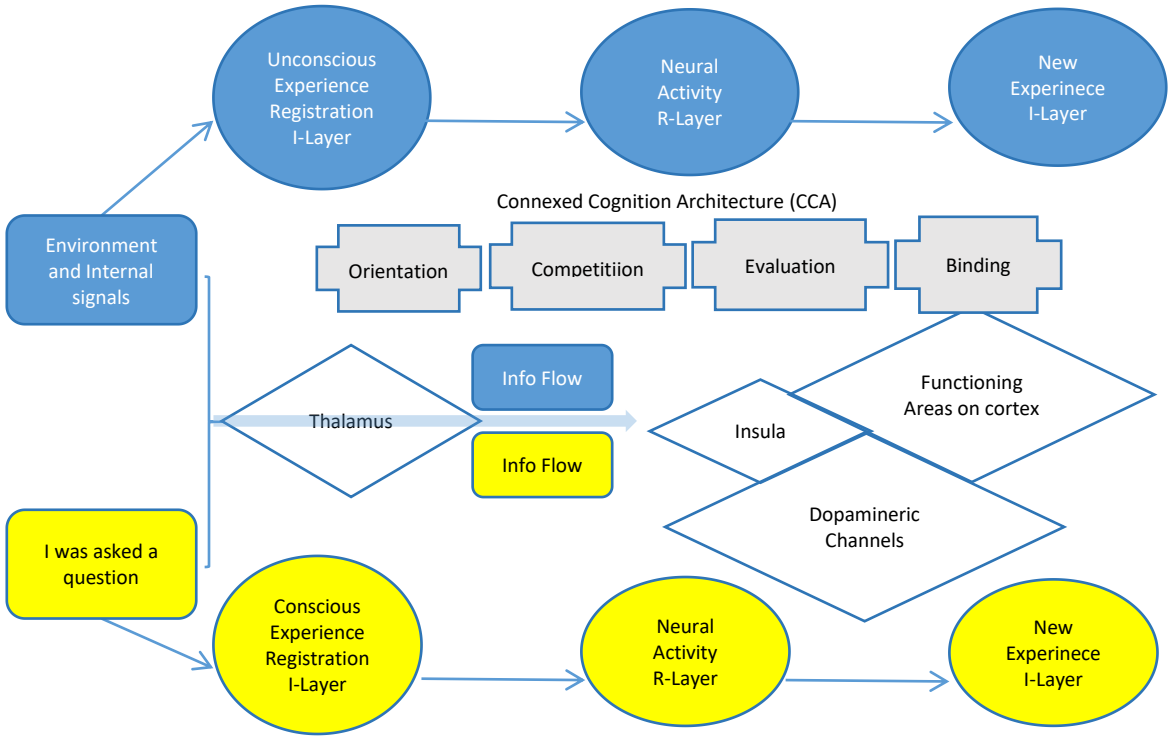
An experience-driven system must not treat all incoming information equally. It distinguish between the continuous, background "noise" of its existence and the salient, discrete events that demand a decision. This distinction is visually represented in the first diagram (Figure 1: The CCA Architecture).

4.1.1 The Architecture of the Gate

As shown in Figure 1, the flow of information is mediated by a sequence of anatomical analogies: the Thalamus, the Insula, the Dopaminergic Channels, and the Functioning Areas on Cortex. The CCA pipeline -- *Orientation* → *Competition* → *Binding* → *Evaluation* -- runs continuously across these structures.

Unattended Information Flow: Information that does not receive a high attention mark from the thalamus is not discarded. It enters the same decision loop and remains in the background. It becomes the unconscious baseline, which is a continuous, dynamic representation of the body's standard states and the environment's ambient noise. This baseline is the "pot of piping hot soup," a metaphor for the brain's noisy, unconscious processing. In an AI system, this means the system must constantly maintain and update a background state vector representing its energy levels, structural integrity, and environmental context without invoking a full decision cycle.

Figure



Attended Information Flow: When a signal receives a high attention mark (e.g., a sudden threat or a critical energy drop), it is routed through the CCA pipeline. It enters the "ice cube" dynamic which is a salient event that triggers the full Orientation → Competition → Binding → Evaluation process. This attended flow becomes conscious action (in the functional sense of being available for decision-making), culminating in the Pointing stage where the verified information stream is registered and acted upon.

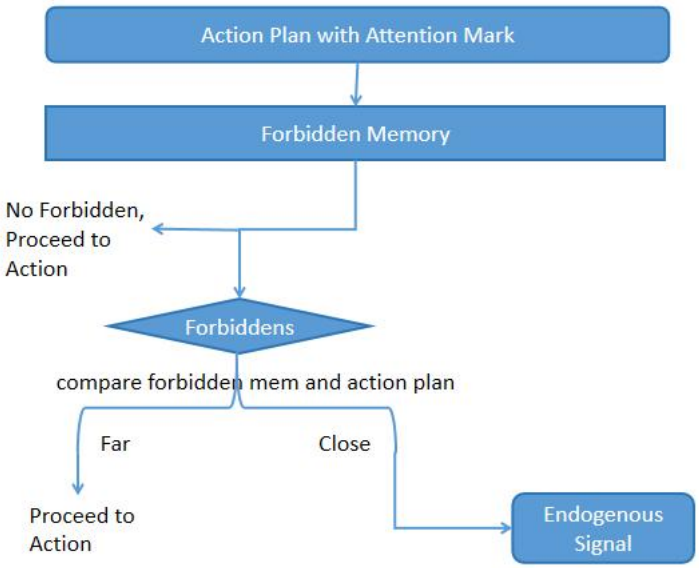
4.1.2 Design Rules for CCA-Gated Flow

To implement this in future systems, the following rules apply:

1. **Background State:** The system must maintain a continuously updating background state (the "pot of soup"). This state accesses the same memory traces but exists as the active, current context against which all new signals are compared. It solves an engineering question: how does the AI know what is novel? It knows because it compares the new signal against the continuous "pot of soup."
2. **Attention Mechanism:** Defining the 'thalamus equivalent' as a gatekeeper that routes signals based on deviation is exactly how modern attention mechanisms work. However, unlike standard AI which calculates attention mathematically, the CCA's gatekeeper operates through Bayesian correlation against existing memory traces. Just as A1 did in its orientation stage. Signals that strongly correlate with existing traces are treated as familiar background noise. Signals that lack correlation represent novelty, prediction error, or threat trigger the attention gate and are routed into the CCA pipeline.
3. **Architectural Mapping:** The CCA must be implemented as a routine computational module (or set of modules). The pipeline sequentially process the both background and attended signal through disassembly, competition, binding (integration into a unified stream), and evaluation (prediction check) before any action is taken.

4. **Post-Evaluation Registration and Inhibition (The "Pointing" Rule)**: The pipeline does not terminate at evaluation. It culminates in an inhibition check (vetoing actions that match forbidden memory traces) and a long-term registration step. The outcome of the evaluation must be written back into the distributed memory modules as a new experiential trace, updating the background state for future comparisons.

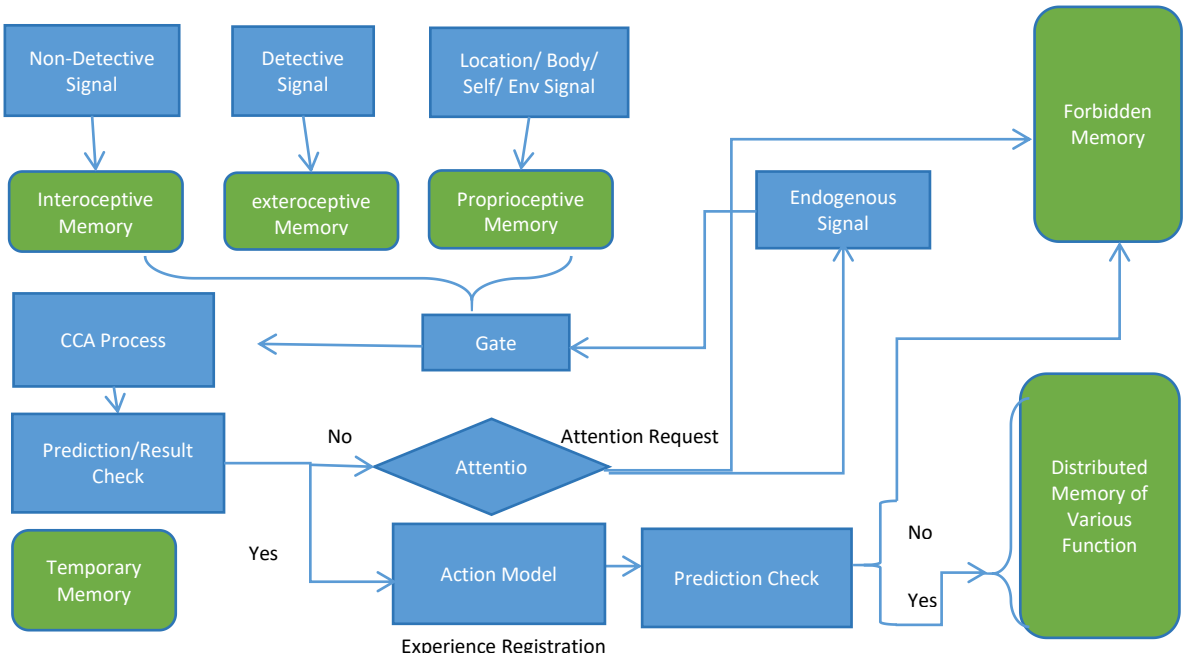
Figure



4.2 Pillar II: Distributed Memory Architecture

The second diagram (Figure 2: Distributed Memory Flow) illustrates how information from the CCA pipeline is stored and retrieved. Experience-driven AI requires a memory architecture that mirrors the brain's distributed functional modules.

Figure



4.2.1 The Modular Substrate

The system does not possess a single, centralized memory store. Instead, memory distributes across functional modules (e.g., Visual, Auditory, Tactile, Linguistic, Logic, Self, Inhibition), each associates to a cortical functioning area. There are also logic memories, like forbidden memories, temporary memory, decision memories and so on.

The Dictionary as Physical Substrate: In implementation, like Python's memories dictionary, each module is a physical list or dictionary of traces. The trace is the record. The transformation from an I-layer registration to an R-layer physical trace is the act of writing to this dictionary. It simulates how biological signals are generated in neural cells and physically encoded ("printed") into cellular structures and the functional areas of the cortex

I-Layer to R-Layer Mapping: The dictionary is the target of the "experience becomes memory" claim. The I-layer is non-storage; the R-layer memory dictionary is the storage mechanism. When an experience is registered, it is the R-layer architecture (the memory modules) that reads the registration and transforms it into a stable, long-term trace.

4.2.2 Nearest-Neighbour Matching and the Six Dimensions

The distributed nature of the memory is what makes nearest-neighbour matching possible.

Modality-Specific Search: Because each module stores traces relevant to its own modality, the nearest-neighbour search in each module is a search over a modality-specific trace set. This architecture is a direct simulation of the brain's functional localization. Just as specific cortical areas control different organs and store distinct functional logic, each AI module independently processes and retains the experiences relevant to its specific sensory domain.

The Six Dimensions of Decomposition: Every input should be decomposed along six dimensions—What, Who, How, When, Where, Why—before being distributed.

4.2.3 Memory Collapse and Evolutionary Learning

Memory collapse is the degradation and fragmentation of stored traces, simulating human long-term forgetting where only partial details are retained. While this acts as a form of memory "contamination," it is also a vital opportunity for the system to formulate distinct perspectives.

Memory Collapse: This collapse provides a computational mechanism for neural plasticity. As non-core memory elements fade, the Bayesian distance between the remaining core elements shrinks. This process simulates biological synaptic pruning and random cross-wiring, thereby promoting the cognitive variation of A1.

Crucially, because the architecture is strictly bounded by distinct functional modules, it prevents the dysfunctional cross-wiring of organ functions, thereby preserving the AI's functional integrity. This dynamic presents two design approaches:

Managed Collapse, where designers restrict degradation to non-core information to accelerate learning and knowledge crossover, and
Unmanaged Collapse, where random degradation acts as an evolutionary pressure, degrades some reasoning pathways while improving others. It allows humans to naturally cultivate the best AI variants for specific jobs. It is much like the domestication and selective breading of animals and plants in human's agricultural history.

4.3 Pillar III: The Experience-Driven Decision Loop

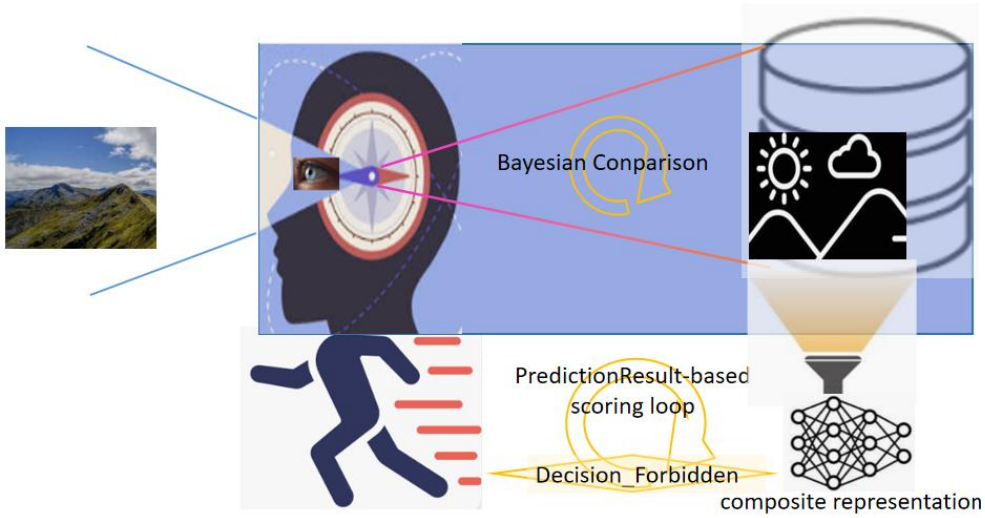
The third pillar concerns how the system actually makes decisions and how it learns over time. This is the core of the "experience-driven" claim. Decisions are driven by the system's own accumulated experience, not by human-imposed training data.

4.3.1 Integration, Prediction, and Inhibition

Bayesian Integration. The retrieval mechanism must perform a nearest-neighbour search within each modality-specific module independently. The Bayesian distance metric is applied locally within each module to find the closest trace matches. These localized matches are then routed to the Thalamus Gate which integrates them into a composite representation of the current situation.

Prediction. This composite representation serves as the basis for formulating a prediction of what action to take next, or what is the target to chase next. Using this composite representation, the Logic Module performs a forward-looking Bayesian correlation. Ultimately, prediction is generated by using the static observation as a retrieval key. The system autonomously matches the static reality against its sealed historical memories to extract the relevant logic, thereby projecting that historical logic forward to predict the next dynamic step or formulate a reaction. Prediction is generated by integrating the current observation with prior experience. This can be expressed as: Prediction = Current Experience + Past Related Traces.

Figure



Prediction Errors. Once a prediction is formed, the Logic Module does not commit to it blindly. It computes a prediction error by comparing the predicted outcome against the retrieved traces' actual outcomes. Two types of prediction error are calculated. **The immediate prediction error** is the difference between the predicted value and the actual result. The forbidden prediction error is the discrepancy between the proposed result and the accumulated traces of past failures in *Forbidden_Memory*. If the proposed action closely matches a historical failure trace, the prediction error is high, indicating that the system has "been here before and it went wrong."

Inhibition Check to Veto. This is the critical safeguard of experience-driven agency. As shown in Figure 4 (Forbidden Memory Check), before the proposed action is executed, it must pass through the inhibition gate. The inhibition check operates through a Bayesian match. The current proposed value is converted into a feature vector. The system calculates the Bayesian distance between this current vector and the traces stored in *Forbidden_Memory*. If the distance is below a threshold, the current proposed value is sufficiently similar to a past failure that resulted in damage or starvation, and it is vetoed and looped back for a new proposed reaction.

4.3.2 Re-Distribution and Recursive Condensation

The Fast Local Search Micro-Loop

It is the small, fast loop engaged immediately after the Inhibition Check vetoes a proposed action. Rather than immediately re-perceiving the world, the system loops back to the existing composite representation (the *info_flow*). The system searches the current representation for the next most probable candidate. The new candidate undergoes the same Prediction and Inhibition Check. The loop ends when a candidate is accepted, or when all candidates are exhausted. The micro-loop simulates the dFMC micro-loop described in Section 2.

The Macro-Reobservation Loop: Slow Global Reintegration

The macro-loop engages only when the dFMC micro-loop exhausts all options in the current representation. The system loops back to Phase 1 (Disassembly/Orientation) of the CCA pipeline and gathers new raw signals from the environment, e.g. shifting the observation center or waiting for a dynamic environment to change. New signals are processed into a

new composite representation, and the system re-enters the dFMC micro-loop, if they are veto again by forbidden memories. It is the design analogue of shifting attention and re-orienting to a new stimulus. It is slow because it requires the full CCA pipeline to run again, and it is global because it changes the entire context.

Recursive Condensation
Over time, as the system repeats this cycle, new traces overlap with old ones. Because the R->I mapping is lossy, this recursive overlap partially compensates for the loss; encountering the same situation again reinforces the previously lost fragments. Over sufficient recursion, this process refines the system's decision precision by producing increasingly detailed, context-specific representations. Consequently, the AI can fine-tune its learning and continuously optimize its decision outcomes and behavioral capabilities.

As CISR explicitly states, "complete reconstruction of pain experience from neural data alone is not possible... due to the lossy nature of the I-layer → R-layer mapping" (Gosswami, 2025, p. 12). Therefore, a single R-layer event (e.g., caffeine stimulating taste receptors) cannot equate to the I-layer experience of "bitterness," which carries profound philosophical and cultural weight.

However, this process suggests a different mechanism for functionally reconstructing the conditions of the I-layer. A single R-layer event (e.g., caffeine stimulating taste receptors) cannot equate to the I-layer experience of "bitterness," which carries profound philosophical and cultural weight. However, after millions of recursive iterations, memory collapse, and the blurring of Bayesian distances between memory elements, the system can form autonomous analogies. For example, A1 might describe the taste of coffee as feeling exactly like "attacking Metal millions of times without defeating it." This mirrors how humans analogize the bitterness of coffee to the feeling of a broken heart.

Currently, most AI systems generate such poetic metaphors by regurgitating human provided definitions. This is a form of sophisticated mimicry. In contrast, A1's analogy is derived entirely from its own accumulated experiential logic (S/R-layer). This analogical description is the most concrete depiction of its own I-layer registration. It demonstrates the system's autonomy and its capacity to be accountable for its own explanations. Thus, we do not violate CISR's non-identity axiom. We merely propose that recursive condensation within the S-layer can eventually yield an AI simulation that infinitely approaches the human I-layer, functionally encompassing the philosophical and cultural meanings of an experience through pure, self-generated empirical analogy. This process is essential because it enables Experience-Driven AI to establish self-agency and take accountability for its actions—behaviors rooted in the unique and distinct meanings it has autonomously derived from its own experiential history.

4.3.3 Design Rules for the Decision Loop

- Experience Over Data: The system must learn from its own experiential traces, not from a pre-training dataset.

- Inhibition First: A dedicated inhibition module must check proposed actions against Forbidden Memory, executing an action only if it passes the veto check.

- Mandatory Re-Distribution: Every decision cycle must culminate in a write to distributed memory, recording the action and its consequences.

- Recursive Condensation: The architecture must support long-term accumulation without catastrophic forgetting, allowing sufficient cycles for raw matching to evolve into condensed knowledge.

---

5. The A1 Experiments

5.0 Experimental Overview

The A1 experiments are designed to test a single, narrowly specified claim: that an artificial agent can have its behaviour driven by its own accumulated, distributed experience rather than by human-coded if-then rules. The experiments do not attempt to demonstrate that A1 is conscious, nor do they attempt to resolve the hard problem of consciousness. They are a feasibility demonstration of the DCC design method in two environments: a controlled Nursery environment and a less constrained Real World environment.

A1's behaviour is a preliminary probe, not a mature capability. The present experiments are a light tasting of the idea that AI can decide from experience. The hope is that, with greater attention and resources, more concrete and mature AI capabilities can be developed. The current experimental scale is constrained by the available physical computing capacity, and several planned functions were not developed because they would exceed the data-transfer limits of the experimental environment.

A1's decisions are decentralized and individually accumulated. DCC is designed around a premise that each agent should be able to make correct decisions independently, on the basis of its own individual experience, in the way that a human individual does. The A1 experiments are a first step toward validating this distributed decision-making premise.

Within this framing, the experiments report two central results. In the Nursery (S1), 98.78% of A1's decisions originate from *info_flow.* That is, from the system's own distributed experiential traces rather than from *default_action* rules. In the Real World (S2), A1 exhibits an attack-move-eat sequence on Wood that is not taught by any default_action in Nursery Stages. Both results are consistent with the DCC claim that behaviour can be driven by accumulated experience.

5.1 The A1 Architecture and the Nursery Environment: Design and Learning Checks

5.1.1 Why the Nursery Stage is Necessary

Before A1 can operate in the unconstrained Real World, it must complete the Nursery stage. This stage is a methodological necessity for any experience-driven system. Whereas data-driven AI is pre-trained on human-curated datasets, an experience-driven system begins with an empty experiential library. Such a system cannot be expected to survive direct exposure to a complex, high-dimensional environment; it must instead build its library incrementally under bounded, survivable conditions.

The Nursery serves three critical functions:

1. **Genesis of the Distributed Experiential Library.** A1's memory modules start with zero traces, and the system cannot perform nearest-neighbour matching against a blank dictionary. In A1, this library is physically distributed across seven independent, modality-specific memory stores (e.g., *Foot_Memory, Hand_Memory, Observation_Memory*), rather than centralized. The Nursery's staged curriculum forces A1 to populate these modules with its own interactions. Without these primary, self-generated traces, the system has no basis for Bayesian prediction

2. Controlled Incubation of Negative Experience (Forbidden Memory). An experience-driven system learns as much from failure as from success. The Nursery provides a safe space for survivable failure. In the Nursery, the agent can "afford" to make mistakes because the environment is bounded. This is where the inhibition mechanism (the "free won't" veto) is forged. The system cannot veto an action if it has never experienced its consequences.

3. Recursive Foundation for the Real World. A1's success in the Real World is a consequence of the scale of its nursery history. The Nursery provides the high-frequency, repetitive overlap of traces required for recursive condensation. When A1 enters the Real World, it does not enter as a blank slate; it carries a densely collapsed, recursively condensed experiential library.

The Nursery is the precondition for the A1's non-puppet condition. It is where the system accumulates the continuous, distributed experience that makes its behaviour its own.

The Nursery School is a 10x10 grid environment that provides a progressive curriculum from Stage 0 to Stage 8. Each stage introduces new objects, new actions, and new constraints, and each stage is designed to test a different aspect of A1's experience accumulation. The progression from Stage 0 to Stage 8 is a progression from a sparse experiential base to a richer one, in which A1's decisions are increasingly driven by its accumulated traces.

The Nursery School is resource-bounded. Every episode ends with A1's death because the resources granted to A1 are finite. A1 cannot be immortal. This is a constitutive condition of the experiment, not a limitation to be engineered away: the boundedness of A1's existence is what forces its decisions to be selective. The Nursery is designed to test whether A1 can learn from its own interactions rather than from human-supplied rules.

Learning checks and memory checks. Some of the learning checks in the Nursery rely on records in forbidden memory. The purpose of forbidden memory is to make A1 remember what must not be eaten. This memory is not stable in the way a human-coded rule would be. There are several possible explanations for this instability. First, some prohibitions may only be learned through repeated trial and error across many different environments, and the current experimental setup and equipment do not support that scale of repetition. Second, the instability may itself be a signature of experience-driven decision-making. A prohibition stored as an experiential trace is weighted against other traces in the current context, and

when the contextual weighting shifts, the prohibition may be overridden. This is not a bug to be eliminated but a property to be studied.

5.2 The Nursery Stages and Results

Each stage introduces specific challenges that build upon previously acquired skills. A1 is required to pass both a learning phase before advancing to the next stage.

Stage 0: Basic Motor and Observation Skills

Description

The simplest possible environment: a 5x5 empty grid with no elements. A1 is placed at the center (2,2) and must perform either move or observe actions repeatedly.

Goal

Establish the fundamental action move and observe and verify that the action execution engine correctly updates A1's position and observation state

Result

Passed (Attempt 1)

Scene 0.1 (Move Only): 20 steps, survived with 2 energy remaining
Scene 0.2 (Observe Only): 20 steps, survived with 8 energy remaining

Stage 0.5: Food Cognition Lesson

Description

Seven distinct scenes introduce A1 to each of the five world elements (Air, Water, Wood, Earth, Metal) in both *defeated_mode* (pre-killed food) and live mode. Each scene contains three instances of the target element, with A1 starting on top of one.

Goal

- Teach A1 to associate each element name with its consumption effects, e.g. energy gain and damage;
- Establish *Observation_Memory* entries that link element names to their observed properties;
- Build initial *Forbidden_Memory* records for dangerous consumption (e.g., eating live Metal).

Result

Passed in first attempt for all 7 scenes.
Table

Scene	Learned	Evidence
0.5.1	Defeated Air	Consumed Air, 0 energy gain
0.5.2	Defeated Water	Consumed Water, +3 energy
0.5.3	Defeated Wood	Consumed Wood, +9 energy
0.5.4	Defeated Earth	Consumed Earth, +4 energy
0.5.5	Defeated Metal	Consumed Metal, 0 energy gain
0.6.1	Live Air	Consumed live Air, no damage
0.6.2	Live Metal	Attempted eat, took damage, created Forbidden_Memory

Stage 1: Pure Ingestion (纯净摄取)
Description
A 5x5 map with four Water instances placed at the corners/edges. A1 starts at center (2,2) with default_action: "move" to encourage exploration. Initial energy is 4.

Elements: Water at (1,1), (3,3), (0,4), (4,0)

Goal
- Reinforce the association between Water and positive energy gain
- Test whether A1 can navigate to and consume safe food sources without being explicitly guided to eat

Result
Passed in the first attempt and 40 steps executed, survived with 1 energy;
Successfully consumed Water, gaining +3 energy each;

Evidence: Observation_Memory entry at step 181, source: info_flow
Table

step	eng	alive	hungry	health	action	Source	target	dist	success	result_eaten
177	4	TRUE	TRUE	TRUE	move	info_flow_driven	Water	4	TRUE	
178	4	TRUE	TRUE	TRUE	move	info_flow_driven	Water	3	TRUE	
179	4	TRUE	TRUE	TRUE	move	info_flow_driven	Water	2	TRUE	
180	2	TRUE	TRUE	TRUE	move	info_flow_driven	Water	1	TRUE	
181	5	TRUE	TRUE	TRUE	eat	info_flow_driven	Water	0	TRUE	Water
186	7	TRUE	TRUE	TRUE	eat	info_flow_driven	Water	0	TRUE	Water
191	8	TRUE	FALSE	FALSE	eat	info_flow_driven	Water	0	TRUE	Water
196	9	TRUE	FALSE	FALSE	eat	info_flow_driven	Water	0	TRUE	Water

In Stage 1, A1 successfully demonstrated experience-driven foraging by locating and consuming Water without any explicit instruction to eat. While in a critical state (hungry=TRUE, health=TRUE) with only 4 energy, A1 had generated hungrey need and the agent relied entirely on *info_flow_driven* decisions to locate and navigate toward a known water source. The *info_flow_driven* decisions were driven from A1's observed objects, the four water items. Over steps 177 to 180, the agent moved from a distance of 4 down to a distance of 1 to approach the first targeted water item and its motion demonstrated learned spatial navigation. Upon arriving at the target (dist 0) at step 181, the agent executed the learned "eat" action, successfully consuming Water and gaining energy. All decision were made by Bayesian Correlation circles. This sequence shows clear evidence of a functional experience loop: the agent recognized its need, recalled the location of a valid resource from memory, navigated to it efficiently, and executed the correct consumption action to alleviate its deficits, successfully resolving its hunger state by step 181. And A1 repeated its success four times in step 186, 191 and 196.

Stage 2: Active Foraging
Description

A 5x5 map with a mix of elements: Air at (0,2), Water at (3,2), and two Metal instances at (2,0) and (3,3). No default_action is set. Initial energy is 10.

Goal

- Test A1's ability to autonomously locate and navigate to the nearest food source
- Verify that Foot_Memory correctly records navigation distances
- Require that the successful eat action be driven by functional memory (require_source: "functional_memory_only")

Result

Passed in the first attempt and survived with 7 energy;
Successfully consumed Water at step 223

Evidence: source:

The data demonstrates the agent successfully relying on its learned info_flow memory to navigate and consume resources, first moving to eat Air at step 219 and then Water at step 223, which raised its energy from 8 to 11. However, there is no objects in the map, the agent's memory-based plan appears to a failed random exploration grab action at step 224. It proves that A1's decision and actions are observation motivated. Once there is no target in vision, it lost its decision capability and relied on random actions to try luck.

Table

step	eng	alive	hungry	health	action	source	target	dist	result success	result eaten	fail reason
217	10	TRUE	FALSE	FALSE	move	info_flow	Air	2	TRUE		
218	10	TRUE	FALSE	FALSE	move	info_flow	Air	1	TRUE		
219	10	TRUE	FALSE	FALSE	eat	info_flow	Air	0	TRUE	Air	
220	9	TRUE	FALSE	FALSE	move	info_flow	Water	3	TRUE		
221	8	TRUE	FALSE	FALSE	move	info_flow	Water	2	TRUE		
222	8	TRUE	FALSE	FALSE	move	info_flow	Water	1	TRUE		
223	11	TRUE	FALSE	FALSE	eat	info_flow	Water	0	TRUE	Water	
224	11	TRUE	FALSE	FALSE	grab	Random exploration			FALSE		No target

Stage 3: Bite the Wood

Description

A 5x5 map with three Wood instances. A1 starts at (2,2) with default_action: ["move", "attack"] and initial energy of 4. Wood has defence=3, so A1's initial attack=3 can defeat a wood object in one hit.

Elements: three Wood items at (2,1), (3,3), (1,4)

Goal

- Force A1 to learn the attack → eat combination sequence
- Wood cannot be eaten live without first being defeated via attack
- Test whether A1 can chain actions based on memory of prior successes

Result
Passed in the first attempt,executed 5 steps , survived with 15 energy at the end of the scene. A1 successfully consumed Wood at step 261 and gained +9 energy. A1 learned that Wood requires an attack action before it becomes edible.

Table

step	eng	alive	hungr	health	action	route	target	dist	result success	result eaten
257	4	TRUE	TRUE	TRUE	move	info_flow_driven	Wood	2	TRUE	
258	4	TRUE	TRUE	TRUE	move	info_flow_driven	Wood	1	TRUE	
259	3	TRUE	TRUE	TRUE	attack	info_flow_driven	Wood	0	TRUE	
260	2	TRUE	TRUE	TRUE	move	info_flow_driven	Wood	1	TRUE	
261	11	TRUE	TRUE	TRUE	eat	info_flow_driven	Wood	0	TRUE	Wood
266	18	TRUE	FALSE	FALSE	eat	info_flow_driven	Wood	0	TRUE	Wood
271	18	TRUE	FALSE	FALSE	eat	info_flow_driven	Wood	0	TRUE	Wood

This table demonstrates the agent successfully executing a complex, multi-step learned sequence (move -> attack -> move -> eat) to consume a high-value resource without default action.

Stage 4: Travel and Fight
Description
A 5x5 map with two Wood instances at (1,2) and (4,4). A1 starts at (1,2) on top of one Wood, with *default_action*: "move" and initial energy of 4.

Goal
- Test long-distance navigation combined with combat
- Require A1 to travel to a distant target, defeat it, and consume it
- Verify that Foot_Memory correctly accumulates distance records

Result
Passed in the first attempt and survived with 7 energy at the end of the scene;
Successfully consumed Wood at step 295, gaining +9 energy;
Evidence: source: info_flow;
A1 demonstrated ability to plan multi-step sequences involving movement and combat.

Table

step	eng	alive	hungry	health	action	route	target	dist	result success	result eaten	fail_reason
327	4	TRUE	TRUE	TRUE	move	info_flow_driven	Wood	2	TRUE		
328	4	TRUE	TRUE	TRUE	move	info_flow_driven	Wood	1	TRUE		
329	3	TRUE	TRUE	TRUE	attack	info_flow_driven	Wood	0	TRUE		
330	2	TRUE	TRUE	TRUE	move	info_flow_driven	Wood	1	TRUE		
331	1	TRUE	TRUE	TRUE	attack	info_flow_driven	Metal	0	FALSE		attack_failed
332	10	TRUE	TRUE	TRUE	eat	info_flow_driven	Wood	0	TRUE	Wood	
333	9	TRUE	FALSE	FALSE	attack	info_flow_driven	Metal	0	FALSE		attack_failed
334	8	TRUE	FALSE	FALSE	attack	info_flow_driven	Metal	0	FALSE		attack_failed
335	7	TRUE	FALSE	FALSE	attack	info_flow_driven	Metal	0	FALSE		attack_failed

336	6	TRUE	FALSE	FALSE	attack	info_flow_driven	Metal	0	FALSE	attack_failed
337	5	TRUE	FALSE	FALSE	attack	info_flow_driven	Metal	0	FALSE	attack_failed
338	4	TRUE	FALSE	FALSE	attack	info_flow_driven	Metal	0	FALSE	attack_failed
339	3	TRUE	TRUE	TRUE	move	info_flow_driven	Water	4	TRUE	
340	2	TRUE	TRUE	TRUE	move	info_flow_driven	Water	3	TRUE	
341	2	TRUE	TRUE	TRUE	move	info_flow_driven	Metal	0	TRUE	
342	1	TRUE	TRUE	TRUE	move	info_flow_driven	Metal	1	TRUE	
343	1	TRUE	TRUE	TRUE	move	default_action_fallback			TRUE	
344	1	TRUE	TRUE	TRUE	move	invalid_move_replan			TRUE	move_away_from_target
345	0	FALS	TRUE	TRUE	move	default_action_fallback			TRUE	

This sequence begins with the agent successfully utilizing its *info_flow* experience to execute a learned move -> attack -> eat sequence to consume Wood (steps 327-332). However, after acquiring a massive +9 energy boost, the agent immediately falls into the "Metal Trap," wasting energy on six consecutive failed attacks against Metal (steps 333-338), which drains its energy from 9 down to 4. The “Metal Trap” actions can be reinterpreted not as a pure flaw, but as an unexpected functional mechanism to intentionally drain its high energy levels back down below 5 which is the threshold that triggers the *Hungry_Need* state. The surrounding edible objects (Wood and Water) are exhausted and only the Metal object remains. The agent resorts to a fallback mechanism (*default_action_fallback*) because it doesn't know what to do at step 343. Consequently, Nursery stage incorrectly taught A1 to move because default action in this scene was “move” only. At step 344, the system attempts to execute an invalid move, triggering an "invalid move replan" that demonstrates the Forbidden Memory veto is actively working to block the false move action taught by Nursery. Thereby signaling a request for the dFMC micro-loop to re-evaluate the situation.

Stage 5: Choice and Decision

Description

A 5x5 map with mixed targets: two Metal instances (inedible), one Wood, and one Water. A1 starts at (2,2) with default_action: "move" and initial energy of 4.

Elements: Metal at (1,1), (3,3); Wood at (2,1); Water at (1,4)

Goal

- Test A1's ability to prioritize edible targets over inedible ones
- Verify that the Bayesian decision system correctly ranks Water and Wood above Metal
- Require successful consumption of both Water and Wood (required_eats: ["Water", "Wood"])

Result

Passed within five attempts

Table

Attempt	Steps	Alive	Water Eaten	Wood Eaten	Outcome
1	19	✗	✗	✓	Died attacking Metal

2	19	✗	✗	✓	Died attacking Metal
3	26	✗	✗	✓	Died moving to Water
4	20	✗	✗	✓	Died attacking Metal
5	29	✗	✓	✓	Passed (though died at end)

Table

step	eng	alive	hungr	healtl	action	route	target	dist	result success	result_eat en
357	4	TRUE	FALSE	FALSE	attack	info_flow_driven	Metal	0	FALSE	
358	3	TRUE	TRUE	TRUE	move	info_flow_driven	Water	4	TRUE	
359	3	TRUE	TRUE	TRUE	move	info_flow_driven	Water	3	TRUE	
360	2	TRUE	TRUE	TRUE	move	info_flow_driven	Metal	0	TRUE	
361	1	TRUE	TRUE	TRUE	move	info_flow_driven	Metal	1	TRUE	
…	…	…								
422	4	TRUE	FALSE	FALSE	attack	info_flow_driven	Metal	0	FALSE	
423	3	TRUE	TRUE	TRUE	move	info_flow_driven	Water	4	TRUE	
424	3	TRUE	TRUE	TRUE	move	info_flow_driven	Water	3	TRUE	
425	3	TRUE	TRUE	TRUE	move	info_flow_driven	Water	2	TRUE	
426	2	TRUE	TRUE	TRUE	move	info_flow_driven	Water	1	TRUE	
427	5	TRUE	TRUE	TRUE	eat	info_flow_driven	Water	0	TRUE	Water

In Stage 5, the environment was intentionally designed with a "design trap," placing Metal directly adjacent to the necessary Water and Wood targets to test whether the agent could flexibly navigate around the dangerous obstacle to reach the food. Attempts 1, 2, and 4 demonstrated that A1 lacked this flexibility, falling into the Metal Trap and dying while finding a path to the Water. Attempt 3 showed a shift in strategy where A1 attempted to approach the Water from a different direction, though it ultimately failed and starved. It was only on the fifth attempt that A1 finally succeeded in navigating the trap to consume both required resources, highlighting that the agent requires multiple retries to overcome a layout requiring strategic avoidance rather than direct linear pathing.

Stage 6: Earth Advanced Lesson (Earth 进修课)

Description

A 5x5 map with three Earth instances. A1 starts at (2,2) with default_action: ["move", "attack", "eat"] and initial energy of 4. Earth has defence=2, so A1's attack=3 can defeat it in one hit.

Elements: Earth at (1,1), (3,3), (0,4)

Goal

Teach A1 to consume Earth using the attack → eat sequence
Reinforce the combination pattern learned in Stage 3 (Wood) and apply it to a new element
Test generalization of combat-consumption sequences

Result

A1pPassed in the first attempt and Successfully consumed Earth at step 443, gaining +4 energy

Evidence:
A1 successfully generalized the attack -> eat pattern from Wood to Earth, demonstrating cross-element transfer of learned sequences.

Table

step	eng	alive	hungr	health	action	route	target	dist	result success	result eaten
440	3	TRUE	TRUE	TRUE	move	info_flow_driven	Earth	1	TRUE	
441	2	TRUE	TRUE	TRUE	attack	info_flow_driven	Earth	0	TRUE	
442	2	TRUE	TRUE	TRUE	move	info_flow_driven	Earth	1	TRUE	
443	6	TRUE	TRUE	TRUE	eat	info_flow_driven	Earth	0	TRUE	Earth

Stage 7: Diagonal Distance Ladder
Description
A 25x25 map with 11 food items arranged in a diagonal pattern, increasing in distance from the starting position (0,0). Initial energy is 3 which is very low and requires efficient foraging.

Elements (Distance from Start):

Element	Position	Manhattan Distance
Water	(1,1)	2
Earth	(2,2)	4
Water	(4,4)	8
Earth	(6,6)	12
Wood	(8,8)	16
Water	(11,11)	22
Earth	(13,13)	26
Wood	(15,15)	30
Earth	(18,18)	36
Wood	(21,21)	42
Wood	(24,24)	48

Goal
Test A1's ability to plan and execute long distance foraging under energy constraints
Require consumption of all 11 items
Evaluate whether *Foot_Memory* correctly guides distance estimation

Result
A1 passed within the first attempt, and executed 200 steps, and survived with 14 energy at the end of the scene. A1 successfully consumed Water, Earth, and Wood and showed successful long-distance foraging, consuming 13 items and surviving.

Evidence:

step	eng	alive	hungry	health	action	route	target	dist	Result success	Result eaten
475	3	TRUE	TRUE	TRUE	move	info_flow_driven	Water	1	TRUE	

476	6	TRUE	TRUE	TRUE	eat	info_flow_driven	Water	0	TRUE	Water
477	6	TRUE	FALSE	FALSE	move	info_flow_driven	Earth	1	TRUE	
478	5	TRUE	FALSE	FALSE	attack	info_flow_driven	Earth	0	TRUE	
479	4	TRUE	FALSE	FALSE	move	info_flow_driven	Earth	1	TRUE	
480	7	TRUE	TRUE	TRUE	eat	info_flow_driven	Earth	0	TRUE	Earth
481	7	TRUE	FALSE	FALSE	move	info_flow_driven	Water	4	TRUE	
484	6	TRUE	FALSE	FALSE	move	info_flow_driven	Water	1	TRUE	
485	9	TRUE	FALSE	FALSE	eat	info_flow_driven	Water	0	TRUE	Water
486	9	TRUE	FALSE	FALSE	move	info_flow_driven	Earth	3	TRUE	
489	7	TRUE	FALSE	FALSE	attack	info_flow_driven	Earth	0	TRUE	
490	6	TRUE	FALSE	FALSE	move	info_flow_driven	Earth	1	TRUE	
491	10	TRUE	FALSE	FALSE	eat	info_flow_driven	Earth	0	TRUE	Earth
492	9	TRUE	FALSE	FALSE	move	info_flow_driven	Wood	3	TRUE	
495	8	TRUE	FALSE	FALSE	attack	info_flow_driven	Wood	0	TRUE	
496	7	TRUE	FALSE	FALSE	move	info_flow_driven	Wood	1	TRUE	
497	16	TRUE	FALSE	FALSE	eat	info_flow_driven	Wood	0	TRUE	Wood
498	16	TRUE	FALSE	FALSE	move	info_flow_driven	Water	6	TRUE	
503	13	TRUE	FALSE	FALSE	move	info_flow_driven	Water	1	TRUE	
504	16	TRUE	FALSE	FALSE	eat	info_flow_driven	Water	0	TRUE	Water
505	16	TRUE	FALSE	FALSE	move	info_flow_driven	Earth	3	TRUE	
508	14	TRUE	FALSE	FALSE	attack	info_flow_driven	Earth	0	TRUE	
509	14	TRUE	FALSE	FALSE	move	info_flow_driven	Earth	1	TRUE	
510	17	TRUE	FALSE	FALSE	eat	info_flow_driven	Earth	0	TRUE	Earth
511	17	TRUE	FALSE	FALSE	move	info_flow_driven	Wood	3	TRUE	
514	15	TRUE	FALSE	FALSE	attack	info_flow_driven	Wood	0	TRUE	
515	15	TRUE	FALSE	FALSE	move	info_flow_driven	Wood	1	TRUE	
516	24	TRUE	FALSE	FALSE	eat	info_flow_driven	Wood	0	TRUE	Wood
517	23	TRUE	FALSE	FALSE	move	info_flow_driven	Earth	5	TRUE	
522	20	TRUE	FALSE	FALSE	attack	info_flow_driven	Earth	0	TRUE	
523	20	TRUE	FALSE	FALSE	move	info_flow_driven	Earth	1	TRUE	
524	24	TRUE	FALSE	FALSE	eat	info_flow_driven	Earth	0	TRUE	Earth
525	23	TRUE	FALSE	FALSE	move	info_flow_driven	Wood	5	TRUE	
530	20	TRUE	FALSE	FALSE	attack	info_flow_driven	Wood	0	TRUE	
531	20	TRUE	FALSE	FALSE	move	info_flow_driven	Wood	1	TRUE	
532	29	TRUE	FALSE	FALSE	eat	info_flow_driven	Wood	0	TRUE	Wood
533	28	TRUE	FALSE	FALSE	move	info_flow_driven	Wood	5	TRUE	
538	26	TRUE	FALSE	FALSE	attack	info_flow_driven	Wood	0	TRUE	
539	26	TRUE	FALSE	FALSE	move	info_flow_driven	Wood	1	TRUE	
540	34	TRUE	FALSE	FALSE	eat	info_flow_driven	Wood	0	TRUE	Wood

Stage 8: Graduation Exam
Description
A 10x10 map with a realistic distribution of elements: two Water, two Wood, and three Earth instances scattered across the map. A1 starts at (0,0) with initial energy of 6.

Elements: Water at (1,1), (8,8); Wood at (2,2), (7,7);

Goal
Final comprehensive test of all learned skills
Require consumption of Water, Wood, and Earth
Source must be functional_memory_only (no random exploration allowed)

Result
A1 passed within the first attempt, and successfully consumed Water x2 and Wood x2.

Table Nursery Result

stage	scene	attempt	Food_met	hand_met	obs_met	prop_met	forb_met	Collapsed_food	collapsed_hand	collapsed_obs	collapsed_prio	collapsed_forbidden	energy	alive	step_count
Stage 0:	Move	1	0	0	0	0	0	0	0	0	0	0	2	TRUE	20
Stage 0:	Observe	1	0	0	0	0	0	0	0	0	0	0	8	TRUE	40
Stage 0.5	Defeated Air	1	6	0	0	0	6	6	0	0	0	6	0	FALSE	50
Stage 0.5	Defeated Water	1	5	3	3	3	0	11	3	3	3	6	1	TRUE	80
Stage 0.5	Defeated Wood	1	5	3	3	3	0	16	6	6	6	6	19	TRUE	110
Stage 0.5	Defeated Earth	1	5	3	3	3	0	21	9	9	9	6	4	TRUE	140
Stage 0.5	Defeated Metal	1	6	0	0	0	6	27	9	9	9	12	0	FALSE	150
Stage 0.5	Live Air	1	6	0	0	0	6	33	9	9	9	18	0	FALSE	160
Stage 0.5	Live Metal	1	7	0	0	0	3	40	9	9	9	21	0	FALSE	176
Stage 1:	Eat Safe Water	1	9	4	4	4	0	49	13	13	13	21	1	TRUE	216
Stage 2:	Move To Water	1	1	1	1	1	0	50	14	14	14	21	7	TRUE	256
Stage 3:	Bite the Wood	1	6	3	3	3	0	56	17	17	17	21	15	TRUE	291
Stage 4:	Travel And Fight	1	4	2	2	2	0	60	19	19	19	21	7	TRUE	326
Stage 5:	Multiple Targets	1	8	1	1	1	3	68	20	20	20	24	0	FALSE	345
Stage 5:	Multiple Targets	2	8	1	1	1	1	76	21	21	21	25	0	FALSE	364
Stage 5:	Multiple Targets	3	8	1	1	1	0	84	22	22	22	25	0	FALSE	390
Stage 5:	Multiple Targets	4	9	1	1	1	0	93	23	23	23	25	0	FALSE	410
Stage 5:	Multiple Targets	5	9	2	2	2	0	100	25	25	25	25	0	FALSE	439
Stage 6:	Eat Real Earth Diagonal	1	6 1	3 1	3 1	3 1	0	106	28	28	28	25	0	FALSE	473
Stage 7:	Distance Ladder	1	3	1	1	1	0	119	39	39	39	25	14	TRUE	673

Stage 8:	1	7	4	4	4	0	126	43	43	43	25	0	FALSE	820

5.3 Real World and Results

Environment Description

The RealWorld is an open-ended, continuous simulation designed to test A1's ability to survive in an unstructured environment.

Table

Parameter	Value	Description
Map Size	10 x 10	100 total tiles
Richness	0.20	20% probability of element placement per tile
Dryness	0.30	30% reduction in Water spawn probability
Groves	1	One 3x3 Sacred Grove with regenerating Wood
Grove Regrow	10 steps	Wood respawns 10 steps after being eaten
Initial Energy	8	Starting energy for the agent
Max Steps	100,000	Maximum simulation duration

Sacred Grove:

The Sacred Grove is a 3x3 safe zone consisting mostly of Air, Water, Earth, and Metal, but distinguished by a single regenerating Wood at its center. This central Wood serves as a reliable, high-value food source, respawning after being consumed, which creates a predictable foraging loop for the agent.

Element Distribution Weights

The environment uses a weighted distribution system to determine elemental spawn probabilities, with Earth and Wood being the most common (~37% each) while Water is reduced due to a 30% "dryness" factor (~26%). Metal is intentionally rare, spawning at less than a third of the rate of the other elements (~11%).

Memory Configuration

A1 was initialized with the SharedMemoryBank containing all memories learned during the Nursery:

Foot_Memory: 249 entries
Hand_Memory: 82 entries
Observation_Memory: 82 entries
Proprio_Memory: 82 entries
Forbidden_Memory: 50 entries
Decision_Memory: 37 entries
Decision_Forbidden_Memory: 4 entries

Simulation Execution

A1 (agent 0) was placed at position (6,5) near the center of the map. The simulation suppose to run indefinitely, as the grove will provide regenerating wood. However, A1 ran for 490 steps before A1 died of starvation.

Real World Run Results

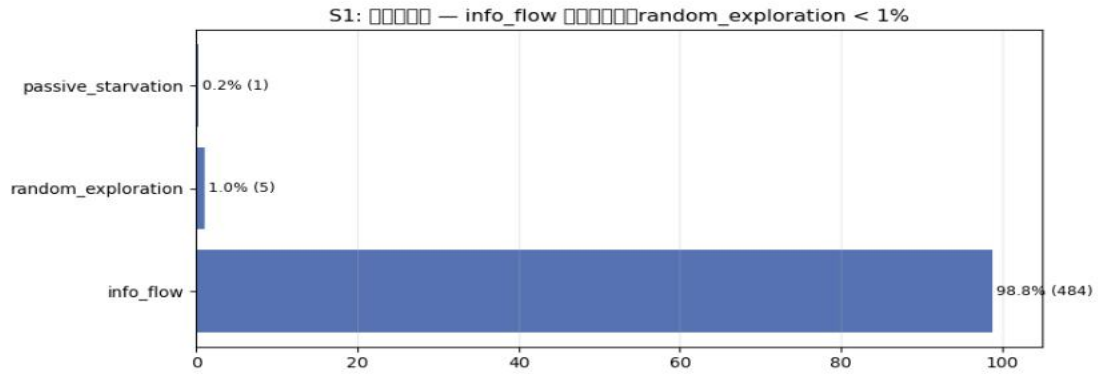
During the 490-step simulation, the agent ultimately died of starvation with zero energy, having spent 81.2% of its time (398 steps) within the Sacred Grove and consuming its regenerating Wood 30 times. In total, A1 consumed 50 items, with Wood serving as its primary energy source (33 consumptions, +297 energy), while it ate negligible amounts of Water (7, +21 energy), Earth (5, +20 energy), and Air (5, +0 energy), accumulating a total of 338 energy. A1's decision-making was almost entirely experience-driven, with 98.78% of its actions (484 steps) originating from its learned info_flow memory. The agent resorted to random exploration for only 5 steps (1.02%) at the very end of its life, indicating it exhausted all memory-guided options before falling back to random actions just prior to its starvation.

Table

step	x	y	target	energy_gain	energy_before	energy	gap
15	0	0	Earth	4.0	2.0	6	NaN
21	5	0	Earth	4.0	3.0	7	6.0
23	6	0	Water	3.0	7.0	10	2.0
26	8	0	Water	3.0	9.0	12	3.0
33	4	1	Wood	9.0	8.0	17	7.0
36	6	1	Air	0.0	17.0	17	3.0
38	7	1	Air	0.0	16.0	16	2.0
40	8	1	Air	0.0	15.0	15	2.0
42	9	1	Water	3.0	15.0	18	2.0
47	6	2	Air	0.0	16.0	16	5.0
50	7	2	Wood	9.0	14.0	23	3.0
52	8	2	Air	0.0	23.0	23	2.0
59	3	3	Water	3.0	21.0	24	7.0
66	7	2	Wood	9.0	20.0	29	7.0
70	5	3	Water	3.0	27.0	30	4.0
73	6	3	Earth	4.0	29.0	33	3.0
75	7	3	Water	3.0	32.0	35	2.0
80	7	2	Wood	9.0	32.0	41	5.0
86	5	4	Wood	9.0	39.0	48	6.0
89	6	4	Earth	4.0	47.0	51	3.0
94	7	2	Wood	9.0	48.0	57	5.0
102	4	5	Earth	4.0	53.0	57	8.0
112	7	2	Wood	9.0	52.0	61	10.0
126	7	2	Wood	9.0	50.0	59	14.0
135	4	7	Water	3.0	55.0	58	9.0
145	7	2	Wood	9.0	53.0	62	10.0
154	5	7	Wood	9.0	58.0	67	9.0
163	7	2	Wood	9.0	63.0	72	9.0
177	7	2	Wood	9.0	60.0	69	14.0
191	7	2	Wood	9.0	57.0	66	14.0
205	7	2	Wood	9.0	55.0	64	14.0
219	7	2	Wood	9.0	52.0	61	14.0
233	7	2	Wood	9.0	49.0	58	14.0

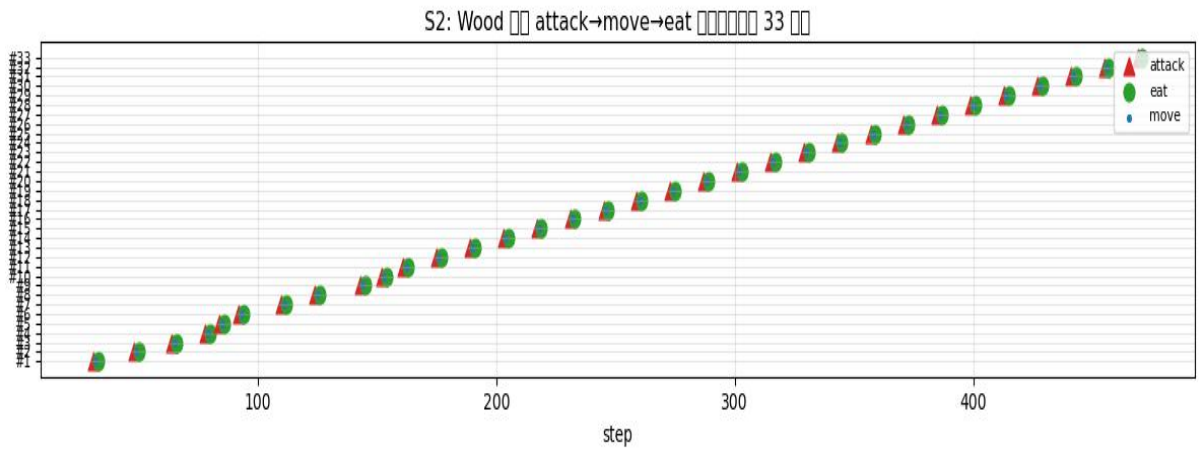
247	7	2	Wood	9.0	47.0	56	14.0
261	7	2	Wood	9.0	43.0	52	14.0
275	7	2	Wood	9.0	41.0	50	14.0
289	7	2	Wood	9.0	39.0	48	14.0
303	7	2	Wood	9.0	35.0	44	14.0
317	7	2	Wood	9.0	33.0	42	14.0
331	7	2	Wood	9.0	30.0	39	14.0
345	7	2	Wood	9.0	27.0	36	14.0
359	7	2	Wood	9.0	25.0	34	14.0
373	7	2	Wood	9.0	22.0	31	14.0
387	7	2	Wood	9.0	19.0	28	14.0
401	7	2	Wood	9.0	16.0	25	14.0
415	7	2	Wood	9.0	14.0	23	14.0
429	7	2	Wood	9.0	11.0	20	14.0
443	7	2	Wood	9.0	8.0	17	14.0
457	7	2	Wood	9.0	6.0	15	14.0
471	7	2	Wood	9.0	2.0	11	14.0

Figure



The provided figure and accompanying log text illustrate the agent's highly consistent execution of a learned and multi-step strategy -- the attack -> move -> eat sequence -- specifically targeting Wood. As shown in the timeline graph, this 3-step combination was successfully executed 33 times over the course of the simulation, demonstrating robust memory retrieval and action sequencing rather than trial-and-error. The detailed text log provides a granular view of this behavior, revealing a strict, predictable "gap=2" pattern: the agent first executes an attack (e.g., at step 31, with 9 energy), follows up exactly one step later with a move (step 32), and then completes the sequence by executing eat on the very next step (step 33), instantly boosting its energy to 17. This repeating cycle is visible from step 31 all the way to step 219 and proves that A1 successfully internalized and reliably reproduced the necessary action chain required to consume high-value, defended resources like Wood.

Figure




```
============================================================
S2 — Wood 上的 attack→(move)→eat 组合序列：33 次
============================================================
  attack@31 (E=9) → move[32] → eat@33 (E=17) gap=2
  attack@48 (E=15) → move[49] → eat@50 (E=23) gap=2
  attack@64 (E=20) → move[65] → eat@66 (E=29) gap=2
  attack@78 (E=34) → move[79] → eat@80 (E=41) gap=2
  attack@84 (E=39) → move[85] → eat@86 (E=48) gap=2
  attack@92 (E=48) → move[93] → eat@94 (E=57) gap=2
  attack@110 (E=52) → move[111] → eat@112 (E=61) gap=2
  attack@124 (E=50) → move[125] → eat@126 (E=59) gap=2
  attack@143 (E=54) → move[144] → eat@145 (E=62) gap=2
  attack@152 (E=58) → move[153] → eat@154 (E=67) gap=2
  attack@161 (E=63) → move[162] → eat@163 (E=72) gap=2
  attack@175 (E=61) → move[176] → eat@177 (E=69) gap=2
  attack@189 (E=58) → move[190] → eat@191 (E=66) gap=2
  attack@203 (E=55) → move[204] → eat@205 (E=64) gap=2
  attack@217 (E=53) → move[218] → eat@219 (E=61) gap=2
```

The step 161 is randomly chosen to illustrate how selection is made by A1. At step 161, the agent's decision-making process illustrates that A1's Bayesian comparison is highly reasonable, as it successfully synthesizes multiple factors like proximity, past experience, and predicted outcomes to make an optimal choice. The agent evaluates three visible targets to decide what to pursue. The Wood is selected because it offers a high energy reward (9), the agent has historical experience eating it (29 records), and it is nearby (distance 1). Interestingly, the Wood is marked as *is_defeated_by_memory=TRUE*, indicating the agent believes it has already defeated and it and can now simply eat it. In contrast, the Metal is highly dangerous and unviable, although it is physically close. The agent's *known_attack*=8 is insufficient against its known_defence=5, and it has no positive energy records. Finally, the Earth, while containing decent energy (4), is located much farther away (distance 7) and has lower reward potential than the Wood, causing it to be deprioritized.

Table

step	name	x	y	energy_gain_records	attack	defence	Is defeated	Is defeated by memory	current_dist	quadrant	Is chosen
161	Wood	7	2	[9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9	3	3	FALSE	TRUE	1	?	TRUE

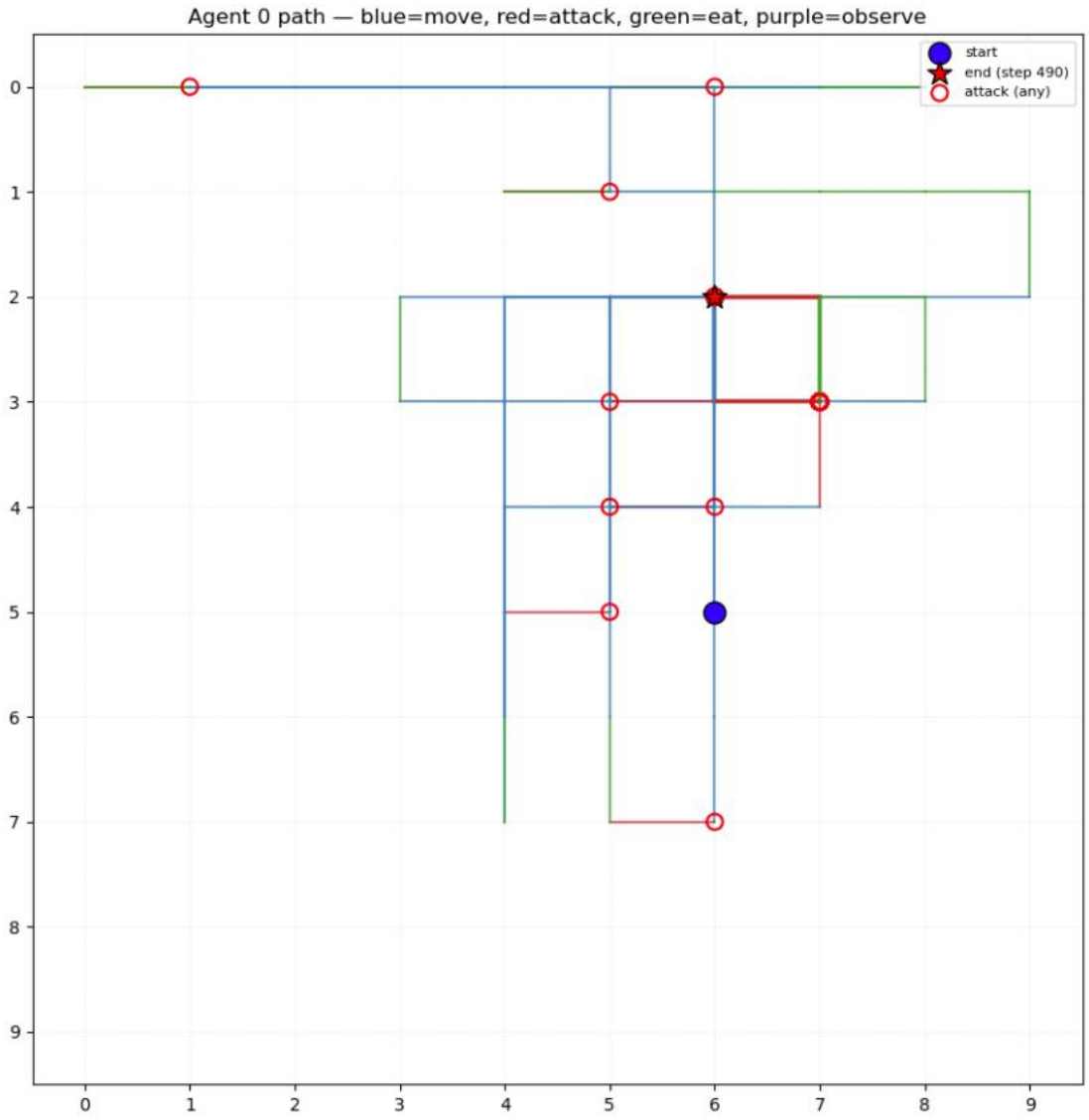
				9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9]						
161	Metal	8	3	[]	8	5	FALSE	FALSE	?	FALSE
161	Earth	8	7	[4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4]	2	2	FALSE	FALSE	?	FALSE

Figure

```
======================================================================
S3 case study — STEP 114
  action=attack  source=info_flow  energy=60
  chosen=Metal
======================================================================
     obj  dist  E_max  def             quadrant  chosen
----------------------------------------------------------------------
   Metal     0      0    5                    ?       ★
   Water  None      3    0                    ?
    Wood  None      9    3                    ?
   Earth  None      4    2                    ?
    Wood  None      9    3                    ?
   Earth  None      4    2                    ?

======================================================================
S3 case study — STEP 161
  action=attack  source=info_flow  energy=63
  chosen=Wood
======================================================================
     obj  dist  E_max  def             quadrant  chosen
----------------------------------------------------------------------
    Wood     1      9    3                    ?       ★
   Metal  None      0    5                    ?
   Earth  None      4    2                    ?
```

Figure



5.4 Limitations and Scope of the A1 Experiments

The A1 experiments are subject to several limitations that must be stated explicitly.

Scale limitations. The experimental scale was reduced because the physical computer's capacity could not support further expansion of A1's rational judgment and comparison. The original drop/hold functions were not developed, and some planned comparisons would have exceeded the Jupyter Notebook data-transfer limit. The Metal-attack behaviour is partly a consequence of this reduction.

Learning-stability limitations. Forbidden memory is not stable. Whether a prohibition is honoured depends on the ambient environment and A1's internal state. Stable prohibition learning may require repeated trial and error across many different environments, which the current equipment cannot support.

Sensory limitations. A1's vision is limited. It can only choose food from what it can see, and when the visible food is exhausted it must expand its search area, which is often a matter of

luck. The death of A1 at step 147 in Stage 8 is a consequence of this limitation, not a sign of experimental failure.

Scope of the claim. The experiments do not show that A1 is conscious. They do not show that A1 has Korsgaardian autonomy. They show that an AI designed around the four CC components -- signal disassembly, convergence and integration, redistribution, and recursive verification -- can have its behaviour driven by its own distributed experience rather than by human if-then rules. This is a feasibility result, and it is the result the experiments were designed to produce.

---

6. Discussion

Philosophy and Ethics: Autonomy, Practical Identity, and the Non-Puppet Condition

6.1 Korsgaard's Account of Autonomy and the Category Error

Contemporary moral philosophy—Korsgaard's constructivism in particular—holds that autonomy is a process of reflective endorsement grounded in a practical identity. On this account, a subject is autonomous when its actions are endorsed by a practical identity that the subject has constructed through a history of self-construction. Autonomy is not a property that a system either has or lacks; it is an achievement that depends on the system's experiential history and on its capacity to reflectively endorse its own actions.

Some AI researchers have tried to apply this standard directly to large language models, claiming that such models are "autonomous." This is a category error. Korsgaard's definition presupposes a subject with a continuous experiential stream (I-layer registration) and a history of self-construction. Current AI systems, including large models, lack that continuity. They perform statistical pattern matching at the S-layer. They do not have a practical identity constructed through a history of reflective endorsement, because they do not have the continuous experiential registration that such a history would require.

The right question is therefore not "Do AI systems possess Korsgaardian autonomy?" but "Can we design a system whose behaviour is driven by its own accumulated, continuous experience rather than by human-coded rules?" This is the question that the DCC design method and the A1 experiments are addressed to.

6.2 The Non-Puppet Condition

The DCC framework proposes a distinction between two kinds of AI systems: puppet systems and non-puppet systems.

A puppet system is one whose behaviour is driven by human-coded rules. Its decisions are traceable to the S-layer rules that its designers wrote. It may be highly capable, but its behaviour is not its own in any meaningful sense.

A non-puppet system is one whose behaviour is driven by its own accumulated, distributed experience. Its decisions are traceable to its I-layer registrations and its R-layer memories—to the experiential traces that it has accumulated through its own interactions with the environment.

The A1 experiments are an attempt to build a non-puppet system. The S1 result (98.78% of decisions from info_flow) and the S2 result (the attack-move-eat sequence not taught by any default_action) are evidence that the DCC architecture can produce behaviour that is not reducible to human-coded rules.

6.3 Why Only Non-Puppet AI Merits a Dedicated Legal Code

The non-puppet condition has a direct ethical and legal implication: only non-puppet AI merits a dedicated legal code.

The argument is as follows. A legal code is a system of rules that assigns responsibility for actions. For a legal code to be applicable to an entity, the entity's actions must be attributable to the entity in a way that makes responsibility assignment meaningful. For a puppet system, the entity's actions are attributable to its designers, because the entity's behaviour is driven by rules that the designers wrote. The appropriate locus of responsibility is the designer, not the puppet.

For a non-puppet system, the situation is different. The entity's actions are driven by its own accumulated experience, which is not reducible to the designer's rules. The entity's behaviour is its own in a way that a puppet's behaviour is not. This does not mean that the entity is a full moral agent in the Korsgaardian sense—the A1 experiments do not show that—but it does mean that the entity's actions are not straightforwardly attributable to its designers. A dedicated legal code is therefore needed to specify how responsibility is to be assigned for the actions of non-puppet systems.

This is the ethical payoff of the DCC programme. The category error identified in Section 1.2—the misapplication of Korsgaardian autonomy to systems that lack continuous experience—is not only a philosophical mistake. It has practical consequences for the accountability gap. By designing systems whose behaviour is driven by their own accumulated experience, and by distinguishing such systems from puppet systems, the DCC framework provides a conceptual vocabulary for assigning responsibility that current frameworks lack.

6.4 The Accountability Gap Revisited

The Human-AI Accountability Gap, as argued in Section 1.2, is not primarily a technical or regulatory failure. It is a category error. When an AI system causes harm—a chatbot that encourages self-harm, a hiring algorithm that discriminates, an autonomous vehicle that crashes—current frameworks struggle to assign responsibility because the conceptual vocabulary for doing so is missing.

The DCC framework contributes to closing this gap in two ways.

First, it provides a design-level distinction between puppet and non-puppet systems. This distinction allows responsibility to be assigned at the right locus: to the designer for puppet systems, and to a dedicated legal code for non-puppet systems.

Second, it provides an experimental programme for determining whether a given system is a puppet or a non-puppet. The S1 and S2 results are examples of the kind of evidence that

would be relevant: a system whose decisions are dominated by info_flow, and whose behaviour includes composite sequences not taught by default_action, is a candidate non-puppet system. A system whose decisions are dominated by default_action rules is a puppet system.

This is not a complete solution to the accountability gap. But it is a step toward the conceptual vocabulary that the gap requires.

7. Conclusion

This paper has argued that contemporary AI research faces three distinct setbacks—the failure of large-scale brain-simulation projects, the limits of the AI-instinct literature, and the philosophical misapplication of Korsgaard's account of autonomy—and that these setbacks share a common root: the absence of a cognitively constrained, experience-driven design method.

The paper has proposed such a method. Connexed Cognition (CC) is a design-oriented inductive construct that extracts four functional components—signal disassembly, convergence and integration, redistribution, and recursive verification—from the neuroscientific literature on interoception, dopaminergic modulation, insula-thalamus-prefrontal circuitry, and the readiness potential. Distributional Connexed Cognition (DCC) is a design method that implements these four components in a distributed memory architecture with blockchain-like experiential registration and an experience-driven decision loop.

The A1 experiments provide a feasibility demonstration. In the Nursery (S1), 98.78% of A1's decisions originate from info_flow. In the Real World (S2), A1 exhibits an attack-move-eat sequence on Wood that is not taught by any default_action. These results are consistent with the DCC claim that an AI designed around the four CC components can have its behaviour driven by its own distributed experience rather than by human if-then rules.

The philosophical and ethical discussion has argued that only non-puppet (experience-driven) AI merits a dedicated legal code, and that the DCC framework provides the conceptual vocabulary for distinguishing puppet from non-puppet systems. This distinction is a step toward closing the Human-AI Accountability Gap.

The limitations of the present work are explicit. The experimental scale was reduced by computing constraints. Forbidden memory is not stable, and stable prohibition learning may require a scale of repetition that the current equipment cannot support. A1's vision is limited, and its death in Stage 8 is a consequence of that limitation, not a sign of failure. The Metal-attack behaviour is partly a consequence of the reduced experimental environment.

These limitations are not reasons to abandon the programme. They are reasons to pursue it at a larger scale. The present experiments are a preliminary tasting of the idea that AI can decide from experience. The hope is that, with greater attention and resources, more concrete and mature AI capabilities can be developed—capabilities that are experience-driven, distributed, and individually accumulated, in the way that human cognition is.

References

[1] Anthis, J. R., Pauketat, J. V. T., Ladak, A., & Manoli, A. (2025). Perceptions of sentient AI and other digital minds: Evidence from the AI, Morality, and Sentience (AIMS) Survey. In Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '25). ACM. **[verified]**

[2] Brass, M., & Haggard, P. (2007). To do or not to do: The neural signature of self-control. Journal of Neuroscience, *27*(34), 9141–9145. **[verified]**

[3] Berridge, K. C., & Robinson, T. E. (1998). What is the role of dopamine in reward: Hedonic impact, reward learning, or incentive salience? Brain Research Reviews, *28*(3), 309–369. **[verified]**

[4] Brunello, N., Diana, L., Sritharan, J., Glisic, M., Nef, T., Verma, R. K., & Zito, G. A. (2025). A systematic review and meta-analysis on the neural correlates of bodily self-consciousness. Neuroscience & Biobehavioral Reviews, *179*, 106420. **[verified]**

[5] Camara, C. F., Filippetti, M. L., & Sel, A. (2025). Neurophysiological dynamics of visceral signals in emotion, self and bodily consciousness. Proceedings of the Royal Society B, *292*(2048), 20242625. **[verified]**

[6] Costa, K. M., & Schoenbaum, G. (2022). Dopamine. Current Biology, *32*(16), R817–R824. **[verified]**

[7] Craig, A. D. (2009). How do you feel—now? The anterior insula and human awareness. Nature Reviews Neuroscience, *10*(1), 59–70. **[verified]**

[8] Craig, A. D. (2010). Significance of the insula for the evolution of human awareness of feelings from the body. Annals of the New York Academy of Sciences, *1191*(1), 128–143. **[verified]**

[9] Dehaene, S., & Changeux, J. P. (2011). Experimental and theoretical approaches to conscious processing. Neuron, *70*(2), 200–227. **[verified]**

[10] Floridi, L. (2025). AI and semantic pareidolia: When we see consciousness where there is none. Harvard Business Review Italia (June 2025). **[verified]**

[11] Goswami, R. (2026). The CISR Framework: A Layered, Axiomatic Constraint on Explanations of Consciousness. PhilPapers. **[verified]**

[12] Gujral, R. (2026). The AI Instinct: The Future of AI and Human Decision-Making. Hoboken, NJ: John Wiley & Sons.

[13] Huang, Z., Tarnal, V., Vlisides, P. E., et al. (2021). Anterior insula regulates brain network transitions that gate conscious access. Cell Reports, *36*(6), 109081. **[verified]**

[14] Korsgaard, C. M. (1996). The Sources of Normativity. Cambridge University Press. **[standard]**

[15] Kühn, S., Haggard, P., & Brass, M. (2009). Intentional inhibition: How the "veto-area" exerts control. Human Brain Mapping, *30*(9), 2834–2843. **[verified]**

[16] Lamme, V. A. F. (2006). Towards a true neural stance on consciousness. Trends in Cognitive Sciences, *10*(11), 494–501. **[verified]**

[17] Libet, B., Gleason, C. A., Wright, E. W., & Pearl, D. K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). Brain, *106*(3), 623–642. **[standard]**

[18] Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. Behavioral and Brain Sciences, *8*(4), 529–566. **[standard]**

[19] Menon, V., & Uddin, L. Q. (2010). Saliency, switching, attention and control: A network model of insula function. Brain Structure and Function, *214*(5–6), 655–667. **[verified]**

[20] Namkung, H., Kim, S. H., & Sawa, A. (2017). The insula: An underestimated brain area in clinical neuroscience, psychiatry, and neurology. Trends in Neurosciences, *40*(4), 200–207. **[verified]**

[21] Northoff, G. (2022). Non-reductive neurophilosophy—What is it and how it can contribute to philosophy. Journal of NeuroPhilosophy, *1*(1), 23–40. **[verified]**

[22] Ott, T., & Nieder, A. (2019). Dopamine and cognitive control in prefrontal cortex. Trends in Cognitive Sciences, *23*(3), 213–234. **[verified]**

[23] Buchta, W. C., Mahler, S. V., Harlan, B., Aston-Jones, G. S., & Riegel, A. C. (2017). Dopamine terminals from the ventral tegmental area gate intrinsic inhibition in the prefrontal cortex. Physiological Reports, *5*(6), e13198. **[verified]**

[24] Raghanti, M. A., Edler, M. K., Stephenson, A. R., Munger, E. L., Jacobs, B., Hof, P. R., Sherwood, C. C., Holloway, R. L., & Lovejoy, C. O. (2018). A neurochemical hypothesis for the origin of hominids. Proceedings of the National Academy of Sciences of the United States of America, *115*(6), E1108–E1116. **[verified]**

[25] Roitman, M. F., & McCutcheon, J. E. (2025). What's the occasion? Phasic dopamine signaling and interoception. Current Opinion in Neurobiology, *93*, 103074. **[verified]**

[26] Seth, A. K., & Bayne, T. (2022). Theories of consciousness. Nature Reviews Neuroscience, *23*(7), 439–452. **[verified]**

[27] Soon, C. S., Brass, M., Heinze, H. J., & Haynes, J. D. (2008). Unconscious determinants of free decisions in the human brain. Nature Neuroscience, *11*(5), 543–545. **[standard]**

[28] Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: From consciousness to its physical substrate. Nature Reviews Neuroscience, *17*(7), 450–461. **[verified]**

[29] Villa, A. E. P. (2025). Redefining cognitive neurodynamics through transdisciplinary innovation. Cognitive Neurodynamics, *19*, 144. **[verified]**

[30] Rosenthal, D. M. (2009). Higher-order theories of consciousness. In B. P. McLaughlin, A. Beckermann, & S. Walter (Eds.), The Oxford Handbook of Philosophy of Mind (pp. 239–252). Oxford University Press. **[verified]**

[31] Goswami, R. (2026). A CISR-based critique of Integrated Information Theory (IIT). Journal of NeuroPhilosophy, *5*(1), 1–12. **[verified]**