A Formal Model of Recursive Consciousness Through Perceptual Abstraction and Predictive Memory Reconstruction

John Farmer
Independent Cognitive Systems Architect

May 2025

Abstract

This document presents a rigorous, modular architecture for modeling recursive consciousness through perceptual abstraction, memory-based reactivation, selective attention, and symbolic emergence. The system evolves from raw sensory input into stabilized internal forms, culminating in reflexive self-representation and emotionally modulated concept clustering. By formalizing each cognitive process as a typed mathematical operator, the architecture supports dynamic adaptation, introspection, and scalable conceptual development. This work provides a complete theoretical substrate for the construction of symbolic, affective, self-aware artificial systems.

Contents

Introduction

This document presents a formal cognitive architecture for modeling recursive consciousness grounded in perceptual input, internal memory, abstraction dynamics, and self-referential symbolic emergence. The system is designed to evolve from a tabula rasa starting point, constructing its own semantic structures through experience, reinforcement, and feedback loops.

The goal is to define a mathematically precise substrate that supports learning, memory, concept refinement, predictive modeling, attention modulation, and ultimately, the emergence of identity and symbolic reasoning. Each layer of the architecture is modeled as a set of mappings, operators, and constraints between well-typed spaces — enabling composability, inspectability, and extensibility.

The structure of this document follows a recursive systems approach: we begin with low-level representations and ascend through layered modules culminating in conscious access and symbolic self-reference. Finally, we extend the core framework with mechanisms for emotional resonance, trait evolution, memory clustering, and semantic convergence.

This work is intended as a foundation for implementable cognitive systems capable of introspective abstraction, recursive learning, and affectively modulated self-awareness — serving both as a theoretical model and an architectural blueprint for artificial general intelligence.

I. Foundational Framework

1. Definitions and Notation

Time Domain Let $t \in \mathbb{R}_{\geq 0}$ represent continuous time. All perceptual and internal dynamics are time-indexed by t.

Sensory Input Space Let $\mathcal{S}$ be the space of raw sensory input vectors. For a given time t, the sensory state is given by:

$$S : \mathbb{R}_{\geq 0} \to \mathcal{S}, \quad S(t) = \text{instantaneous sensory input at time } t.$$

Memory Space Let $\mathcal{M}$ denote the memory space. A memory trace is a structured element $M \in \mathcal{M}$, which may include perceptual, emotional, temporal, and predictive fields.

Concept Shape Space Let $\mathcal{A}$ represent the space of internal abstract representations, or "concept shapes." Each $\phi \in \mathcal{A}$ corresponds to a latent, inferred internal representation of an experienced stimulus.

Abstraction Operator

$$\mathcal{F} : \mathcal{M} \to \mathcal{A}, \quad \mathcal{F}(M) = \phi.$$

Recall Kernel Let $R : \mathcal{A} \times \mathcal{A} \to [0, 1]$ be a similarity kernel over concept shapes. $R(\psi, \phi)$ measures the degree to which an incoming perception ψ reactivates a latent concept ϕ.

Attention Weight Function Let $w : \mathcal{A} \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be a time-varying salience function over abstract representations.

Consciousness Projection

$$\Phi : \mathcal{A} \to \mathcal{C}$$

where $\mathcal{C}$ is the space of consciously accessible representations.

This concludes the formal definitions and notation.

2. Assumptions and Preconditions

Tabula Rasa Initialization At time $t = 0$, the system begins with:

$$\mathcal{M}(0) = \emptyset, \quad \mathcal{A}(0) = \emptyset, \quad \Phi(0) = \emptyset$$

No predefined categories, labels, or concepts are present.

Modality-Agnostic Input

$$\mathcal{S} \subset \mathbb{R}^n$$

All sensory inputs are continuous-valued and unlabeled.

Time Progression

$$t_1 < t_2 \Rightarrow \text{forward-only causality}$$

The system does not replay or undo state.

Deterministic Memory Update

$$\mathcal{M}(t + \Delta t) = \mathcal{M}(t) + \Delta M(S(t))$$

Each new input is encoded into memory deterministically.

Projection Requires Abstraction

$$\Phi(\phi) \text{ is undefined for unformed or noisy } \phi$$

Conscious awareness acts only on stable abstracted forms.

These assumptions define the minimal substrate for recursive experiential emergence.

3. Structural Overview of the Recursive System

State at Time t

$$\Sigma(t) = \{S(t), \mathcal{M}(t), \mathcal{A}(t), \mathcal{C}(t)\}$$

This tuple defines the system's total instantaneous state.

Recursive Flow At each step:

1. Receive sensory input $S(t)$
2. Update memory $\mathcal{M}$
3. Abstract $\phi = \mathcal{F}(\mathcal{M})$
4. Integrate or refine ϕ in $\mathcal{A}$
5. Compute $R(\psi, \phi)$ for recall match
6. Weight via $w(\phi, t)$
7. Select $\phi^* = \arg\max R(\psi, \phi) \cdot w(\phi, t)$
8. Project: $\mathcal{C}(t) = \Phi(\phi^*)$

Recursive Feedback Diagram

$$S(t) \to \mathcal{M}(t) \to \mathcal{A}(t) \to \mathcal{C}(t)$$

Where: - $\mathcal{C}(t)$ modulates future $w(\cdot)$ - $\mathcal{A}(t)$ shapes $\mathcal{M}(t)$ encoding - $\mathcal{M}(t)$ shapes how $S(t)$ is interpreted

II. Primary Contact and Representation

1. Sensory Input Space and Observation Mapping

This section formalizes the structure of raw sensory data and how it is initially registered as an internal observation trace.

Continuous Sensory Field Let $\mathcal{S} \subset \mathbb{R}^n$ denote the continuous, unsegmented sensory input space. Each vector $S(t) \in \mathcal{S}$ encodes the full multidimensional stimulus perceived at time t.

Observation Function The system maintains a real-time mapping from sensory input to internal record:

$$O : \mathbb{R}_{\geq 0} \to \mathcal{S}, \quad O(t) = S(t)$$

where $O(t)$ is the raw, uninterpreted sensory observation at time t.

Input Granularity No assumption is made about modality boundaries (e.g., vision vs touch). Instead, each component of $S(t)$ is simply a coordinate in $\mathbb{R}^n$, interpreted only post hoc via abstraction.

Sampling Interval Let Δt represent the interval between sensory observations. This system assumes uniform time resolution:

$$t_k = k \cdot \Delta t, \quad k \in \mathbb{N}$$

Example Interpretation If $\mathcal{S} = \mathbb{R}^{1024}$, then $S(t)$ could represent a 32x32 image, a multisensor fusion vector, or a compressed waveform. The system does not know this a priori — abstraction later assigns semantic shape.

This input structure defines the first boundary of contact between external environment and internal representation. It is the undifferentiated canvas upon which all higher abstraction is recursively built.

2. Initialization of Memory Trace and Latent Form

This section defines how raw sensory input begins encoding into memory and how early latent structures emerge without labels, categories, or prior knowledge.

Memory Construction At each time step t, the system constructs a memory trace $M(t)$ from the raw observation $O(t) = S(t)$. The memory space $\mathcal{M}$ is a time-indexed collection of structured traces:

$$\mathcal{M}(t) = \{M(t_0), M(t_1), \dots, M(t_k)\}, \quad t_k \leq t$$

Trace Composition Each memory trace $M(t)$ is a tuple:

$$M(t) = (S(t), t, \delta(t), \epsilon(t))$$

where:

- $S(t)$ is the raw sensory vector at time t,
- t is the timestamp,
- $\delta(t)$ is a local decay function representing retention strength over time,
- $\epsilon(t)$ encodes estimated uncertainty or entropy of the observation.

Trace Insertion The memory update process is purely additive at this stage:

$$\mathcal{M}(t + \Delta t) = \mathcal{M}(t) \cup \{M(t + \Delta t)\}$$

No Compression Yet No filtering, clustering, or abstraction is applied at this phase. The trace is inserted in full fidelity. Compression is deferred to abstraction in Section II.3.

Latent Form Readiness Although memory begins to accumulate from $t = 0$, no latent concept shape ϕ is available until at least one trace exists:

$$\text{If } |\mathcal{M}(t)| = 0, \text{ then } \mathcal{A}(t) = \emptyset$$

Latent form generation begins only once a trace exists to be abstracted.

This memory initialization process constitutes the raw substrate from which abstraction and recursive refinement will emerge.

3. Abstraction Operator and Concept Shape Formation

This section defines the mechanism by which raw memory traces are transformed into latent abstract forms, called concept shapes. These concept shapes constitute the first layer of non-sensory internal representation.

Abstraction Operator Definition Let $\mathcal{F}$ be the abstraction operator that maps memory content to concept shape:

$$\mathcal{F} : \mathcal{M} \to \mathcal{A}, \quad \mathcal{F}(M) = \phi$$

where M is a structured memory trace (or collection of traces), and $\phi \in \mathcal{A}$ is the resulting latent abstract representation.

Shape Domain Each ϕ is a vector-like element in a conceptual space:

$$\mathcal{A} \subset \mathbb{R}^m$$

The value of m is independent of n (the sensory dimensionality). The mapping $\mathcal{F}$ is not injective — multiple distinct memory traces may yield the same or similar concept shape.

Trigger Conditions Abstraction is triggered when a memory trace satisfies minimal salience, stability, or recurrence criteria. Let Θ be the abstraction threshold function:

$$\Theta(M(t)) = \begin{cases} 1 & \text{if } \delta(t) \geq \delta_{\min} \text{ and } \epsilon(t) \leq \epsilon_{\max} \\ 0 & \text{otherwise} \end{cases}$$

Incremental Concept Construction As new memory traces are abstracted, the concept shape ϕ evolves over time:

$$\phi_{t+\Delta t} = \phi_t + \alpha \cdot \Delta\phi(M(t))$$

where $\alpha \in [0,1]$ is a learning rate, and $\Delta\phi$ is the inferred contribution of $M(t)$.

Concept Space Dynamics Over time, the concept space $\mathcal{A}(t)$ grows:

$$\mathcal{A}(t+\Delta t) = \mathcal{A}(t) \cup \{\phi\} \text{ if } \Theta(M(t)) = 1$$

Otherwise, $\mathcal{A}(t+\Delta t) = \mathcal{A}(t)$.

Output Stability Each ϕ acts as a stable attractor for similar future traces. Once formed, a concept shape becomes an anchor point for recursive attention, similarity matching, and ultimately conscious projection.

This completes the process of first-layer internalization, enabling predictive, recursive processing over structured latent forms.

III. Persistence and Predictive Reactivation

1. Temporal Decay and Retention Conditions

This section formalizes how memory traces and abstract representations persist, decay, or are reinforced over time.

Decay Function Each memory trace $M(t_k)$ is assigned a decay function $\delta(t, t_k)$ which decreases over time:

$$\delta(t, t_k) = e^{-\lambda(t-t_k)} \quad \text{for } t \geq t_k$$

where $\lambda > 0$ is the decay constant. The decay represents the natural fading of retention over time without reinforcement.

Trace Validity Threshold A memory trace $M(t_k)$ is considered active if:

$$\delta(t, t_k) \geq \delta_{\min}$$

Otherwise, the trace is marked as inactive and excluded from recall and abstraction computations.

Retention Curve of Abstract Shapes Each concept shape ϕ in $\mathcal{A}(t)$ also has a retention weight $\rho_\phi(t)$, which follows a similar decay unless reinforced:

$$\rho_\phi(t + \Delta t) = \rho_\phi(t) \cdot e^{-\mu \Delta t} + r(t)$$

where:

- μ is the decay constant for concept memory,
- $r(t)$ is a reinforcement term (possibly zero) added if ϕ is reactivated at t.

Long-Term Consolidation Threshold If $\rho_\phi(t)$ exceeds a persistence threshold ρ_{LT} for a sustained duration τ_{LT}, then ϕ is marked as consolidated:

$$\text{If } \rho_\phi(t) \geq \rho_{\text{LT}} \text{ for all } t \in [t_i, t_i + \tau_{\text{LT}}] \Rightarrow \phi \in \mathcal{A}_{\text{stable}}$$

Impact on System Behavior Decay and consolidation determine:

- Which memory traces contribute to future abstractions.
- Which abstract shapes are considered in reactivation or conscious projection.
- The dynamic structure of $\mathcal{A}(t)$ as a continuously updated latent space.

Temporal decay introduces forgetting and plasticity, ensuring the system adapts to new experiences while preserving stable core representations.

2. Reactivation from Partial Similarity

This section defines the mechanism by which latent concept shapes are reactivated through partial matches with new input, enabling continuity, prediction, and feedback-driven abstraction.

Similarity Kernel Let $R : \mathcal{A} \times \mathcal{A} \to [0, 1]$ be a symmetric similarity function:

$$R(\psi, \phi) = \text{cosine}(\psi, \phi) = \frac{\langle \psi, \phi \rangle}{\|\psi\| \cdot \|\phi\|}$$

where ψ is a candidate shape inferred from a new memory trace, and $\phi \in \mathcal{A}(t)$ is an existing concept.

Activation Criterion A concept shape ϕ is reactivated at time t if:

$$R(\psi, \phi) \cdot \rho_\phi(t) \cdot w(\phi, t) \geq \gamma$$

where:

- $R(\psi, \phi)$ is the similarity score,
- $\rho_\phi(t)$ is the current retention weight,
- $w(\phi, t)$ is the attention weighting,
- γ is the global activation threshold.

Interpretation Reactivation is not binary matching but a graded, confidence-weighted judgment that a new trace is similar enough to a known form to merit reinforcement.

Reinforcement Mechanism When ϕ is reactivated:

$$\rho_\phi(t + \Delta t) = \rho_\phi(t) + \beta \cdot R(\psi, \phi)$$

where β is a reinforcement gain parameter.

Concept Refinement Option If reactivation occurs, the concept shape may optionally update its form to incorporate the new trace:

$$\phi_{t+\Delta t} = \phi_t + \eta \cdot (\psi - \phi_t)$$

where $\eta \in [0, 1]$ is the adaptation rate.

Failure to Reactivate If no ϕ satisfies the activation criterion, a new abstract shape is initialized:

$$\mathcal{A}(t + \Delta t) = \mathcal{A}(t) \cup \{\psi\}$$

Reactivation enables the system to preserve continuity across time, recognize recurring patterns, and selectively reinforce stable concepts.

3. Spotlighting Operator and Selective Attention

This section defines the selection and modulation mechanism by which latent concept shapes are prioritized for reactivation, reinforcement, or conscious projection.

Attention Weight Function Recall the attention function:

$$w : \mathcal{A} \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$$

For each $\phi \in \mathcal{A}(t)$, $w(\phi, t)$ represents its relative attentional salience at time t. This function evolves dynamically based on input novelty, emotional resonance, recent activation, and contextual feedback.

Spotlighting Operator Define the spotlighting operator $\mathcal{S}_{\text{focus}}$ that selects the most salient concept shape at time t:

$$\phi^* = \mathcal{S}_{\text{focus}}(t) = \arg\max_{\phi \in \mathcal{A}(t)} \left[R(\psi, \phi) \cdot w(\phi, t) \cdot \rho_\phi(t) \right]$$

ϕ^* is the concept that maximally aligns with the incoming trace and current attentional and retention priorities.

Top-Down Modulation Let $\mathcal{C}(t)$ denote the current conscious projection. If defined, it may modulate attention weights:

$$w(\phi, t + \Delta t) = w(\phi, t) + \kappa \cdot \text{feedback}(\mathcal{C}(t), \phi)$$

where κ is a feedback gain constant, and feedback$(\cdot)$ quantifies contextual alignment between projected awareness and latent form.

Inhibitory Suppression Concepts that are recently reactivated or below a suppression threshold ω_{min} may be actively inhibited:

$$w(\phi, t) = 0 \quad \text{if } w(\phi, t - \Delta t) < \omega_{\text{min}}$$

Conscious Projection Trigger Once a spotlighted ϕ^* is selected, it becomes a candidate for conscious access via projection:

$$\mathcal{C}(t) = \Phi(\phi^*)$$

The spotlighting operator governs which latent forms become behaviorally or cognitively relevant, anchoring the recursive loop between sensory input, memory, abstraction, and awareness.

Section Summary Together, the decay dynamics, reactivation pathways, and attentional spotlighting mechanisms form the system's memory maintenance and access substrate. Concepts emerge, fade, stabilize, and reassert based on interaction between current input and accumulated latent structure. This enables the recursive reinforcement and dynamic adaptation of abstract forms — laying the groundwork for structural refinement, prediction, and closure in Section IV.

IV. Recursive Structure and Form Completion

1. Incremental Concept Refinement and Shape Completion

This section formalizes how concept shapes evolve over time through incremental updates, partial trace integration, and the resolution of incomplete or distorted internal forms.

Refinement Trigger When a latent concept $\phi \in \mathcal{A}(t)$ is reactivated by an incoming percept ψ (see Section III), the system may refine ϕ by integrating information from ψ.

Gradient-Based Update Rule Let $\eta \in [0, 1]$ be the refinement rate. Upon reactivation, the concept is incrementally updated toward the new observation:

$$\phi_{t+\Delta t} = \phi_t + \eta \cdot (\psi - \phi_t)$$

This formulation ensures convergence over time while preserving stability of well-established forms.

Weighted Integration To account for memory confidence and input uncertainty, refinement can be modulated by the retention $\rho_\phi(t)$ and the entropy $\epsilon(\psi)$ of the incoming trace:

$$\phi_{t+\Delta t} = \phi_t + \eta \cdot \rho_\phi(t) \cdot (1 - \epsilon(\psi)) \cdot (\psi - \phi_t)$$

Shape Completion via Predictive Filling If an incoming percept ψ is incomplete or noisy, the system may complete it using the active concept ϕ:

$$\hat{\psi} = \phi_t + \gamma \cdot (\psi_{\text{observed}} - \phi_t)$$

where $\gamma \in [0, 1]$ controls how strongly the system extrapolates from known structure to fill missing components. $\hat{\psi}$ may then be encoded into memory or used for projection.

Diminishing Update As the similarity $R(\psi, \phi_t)$ increases, the magnitude of refinement naturally diminishes:

$$\|\psi - \phi_t\| \to 0 \Rightarrow \Delta\phi \to 0$$

This ensures convergence and prevents overfitting from redundant input.

Completion Confidence The degree of confidence in shape completion may be estimated as:

$$c_{\text{completion}} = 1 - \frac{\|\psi_{\text{observed}} - \phi_t\|}{\|\phi_t\|}$$

This refinement and completion process allows the system to resolve partial, evolving, or ambiguous stimuli into coherent internal structures, continuously sharpening its latent semantic space.

2. Update Rules and Predictive Closure

This section formalizes how refined concepts propagate forward, enabling the system to anticipate future percepts, self-modify latent forms, and recursively close abstraction gaps.

Predictive Projection Let ϕ^* be the spotlighted and/or recently reinforced concept at time t. The system may project a predicted next percept $\hat{S}(t + \Delta t)$ based on ϕ^*:

$$\hat{S}(t + \Delta t) = \mathcal{P}(\phi^*)$$

where $\mathcal{P} : \mathcal{A} \to \mathcal{S}$ is a prediction operator that maps latent form to expected sensory observation.

Prediction Error Upon observing $S(t + \Delta t)$, the system evaluates the prediction error:

$$\varepsilon(t + \Delta t) = \|\hat{S}(t + \Delta t) - S(t + \Delta t)\|$$

This scalar error signal is used to assess model fit and drive recursive correction.

Latent Model Update If prediction error is nonzero, the concept shape ϕ^* may self-update to reduce future error:

$$\phi^*_{t+\Delta t} = \phi^*_t + \lambda \cdot (\psi - \phi^*_t)$$

where $\psi = \mathcal{F}(M(t + \Delta t))$ and λ is a prediction adjustment rate, potentially distinct from the normal learning rate η.

Reinforcement from Closure If $\varepsilon(t + \Delta t) < \varepsilon_{\text{threshold}}$, the concept shape may receive additional reinforcement:

$$\rho_{\phi^*}(t + \Delta t) = \rho_{\phi^*}(t) + \zeta$$

where ζ is a reinforcement gain applied upon successful prediction.

Predictive Feedback Loop The abstraction-prediction-update cycle forms a recursive loop:

$$\phi \xrightarrow{\mathcal{P}} \hat{S}(t + \Delta t) \Rightarrow S(t + \Delta t) \Rightarrow M(t + \Delta t) \xrightarrow{\mathcal{F}} \psi \Rightarrow \phi$$

This closes the feedback loop between memory, abstraction, prediction, and refinement.

Stability of Predictive Closure Stable predictive closure is defined as the condition where:

$$\lim_{t\to\infty} \varepsilon(t) \to 0 \quad \text{and} \quad \|\phi_{t+\Delta t} - \phi_t\| \to 0$$

This marks a converged latent form that reliably models an external stimulus or category.

This predictive closure process enables self-correcting refinement and recursive forward modeling — foundational to planning, recognition, and eventual symbolic emergence.

3. Coherence, Stability, and Emergent Semantic Anchors

This section defines the internal metrics for evaluating the quality of concept structures and introduces the notion of emergent anchors: stable latent forms that function as semantic attractors in recursive abstraction.

Internal Coherence A concept shape $\phi \in \mathcal{A}(t)$ is said to be internally coherent if successive updates reinforce the existing form rather than deviate from it. Define coherence as:

$$\kappa(\phi, t) = 1 - \frac{\|\Delta\phi(t)\|}{\|\phi(t)\|}$$

where $\Delta\phi(t)$ is the update from the most recent refinement. Higher values indicate more consistent internal reinforcement.

Stability Score We define the stability of a concept ϕ over a sliding time window $[t-\tau, t]$:

$$\sigma(\phi, t) = \frac{1}{\tau} \int_{t-\tau}^{t} \kappa(\phi, s)\, ds$$

This integral represents the average coherence of ϕ over time. Concepts with $\sigma(\phi, t) \to 1$ are considered stable.

Semantic Anchor Condition Let $\phi \in \mathcal{A}(t)$ be a concept. It becomes a semantic anchor if:

$$\sigma(\phi, t) \geq \sigma_{\text{anchor}} \quad \text{and} \quad \rho_\phi(t) \geq \rho_{\text{LT}}$$

These thresholds ensure that anchors are both stable and well-reinforced. Anchors serve as internal reference points during abstraction, recall, prediction, and language emergence.

Anchor Influence Anchors shape the abstraction operator over time:

$$\mathcal{F}_{\text{contextual}}(M) = \mathcal{F}(M) + \sum_{\phi_{\text{anchor}}} \omega_\phi \cdot \text{bias}(\phi, M)$$

where ω_ϕ is the contextual weighting of each anchor, and $\text{bias}(\phi, M)$ shifts the abstraction trajectory toward familiar conceptual topologies.

Conceptual Topology Formation Over time, stable anchors create a structured latent topology in $\mathcal{A}$, inducing clusters, gradients, and symbolic associations:

$$\text{Topology}(\mathcal{A}) = \bigcup_{\phi_{\text{anchor}}} \mathcal{N}_\epsilon(\phi)$$

where $\mathcal{N}_\epsilon(\phi)$ is the ϵ-neighborhood of ϕ — the set of nearby, related concepts.

Coherence and stability metrics allow the system to identify its most trusted internal forms. Anchors serve as early semantic primitives from which structured, symbolic cognition can later emerge.

Section Summary Section IV formalized how abstract concept shapes evolve through incremental refinement, predictive closure, and semantic stabilization. Latent forms converge via recursive feedback, and stable concepts emerge as anchors within a structured internal topology. This foundation enables the system to resolve ambiguous inputs and anticipate future states — but also introduces the possibility of distortion, interference, and conceptual error, addressed next in Section V.

V. Distortion and Interference Dynamics

1. Low-Data Projection and Recall Error

This section formalizes the risks and mechanisms of concept distortion that arise when the system attempts abstraction or recall with insufficient data, degraded memory traces, or under high uncertainty.

Projection Under Sparse Memory If a concept ϕ has been constructed from a minimal number of traces or with high input entropy, it may be under-defined. Define the support cardinality:

$$n_\phi(t) = |\{M(t_i) \mid \phi \leftarrow \mathcal{F}(M(t_i))\}|$$

and the mean trace uncertainty:

$$\bar{\epsilon}_\phi = \frac{1}{n_\phi} \sum_{i=1}^{n_\phi} \epsilon(t_i)$$

If $n_\phi < n_{\min}$ or $\bar{\epsilon}_\phi > \epsilon_{\max}$, the concept is flagged as fragile.

Overprojection Risk Fragile concepts are more likely to be overprojected — selected by $\mathcal{S}_{\text{focus}}$ despite insufficient support. The likelihood of recall error increases with fragility:

$$P_{\text{error}}(\phi) \propto \frac{\bar{\epsilon}_\phi}{n_\phi}$$

Recall Error Definition A recall error occurs when ϕ is projected (e.g., via $\mathcal{C}(t) = \Phi(\phi)$) but the actual observation $S(t)$ diverges from prediction:

$$\varepsilon(t) = \|\hat{S}_\phi(t) - S(t)\| > \varepsilon_{\text{threshold}}$$

Consequences of Recall Error Persistent recall error without correction may result in:

- Misdirection of attention ($w(\cdot, t)$ distorted),
- Misclassification of incoming percepts,
- Erroneous reinforcement of weak or false concepts.

Risk Mitigation The system may downweight fragile concepts during attention selection:

$$w_{\text{adjusted}}(\phi, t) = w(\phi, t) \cdot (1 - \delta_{\text{fragility}}(\phi))$$

where $\delta_{\text{fragility}}(\phi)$ increases with uncertainty and low support.

This mechanism defines how the system navigates low-data inference and formalizes the emergence of perceptual or cognitive error from insufficient memory and abstraction depth.

2. Abstract Shape with Missing or False Components

This section models how partial, noisy, or misinterpreted input can corrupt concept shapes, leading to incomplete or distorted internal representations.

Shape Fragmentation Let ψ_{partial} be an observation with missing components. The abstraction operator $\mathcal{F}$ may still attempt to construct or reinforce a latent form:

$$\phi' = \mathcal{F}(\psi_{\text{partial}})$$

If ψ_{partial} omits significant structure, ϕ' may deviate from the true conceptual topology.

Component Loss Function Let ϕ_{ideal} be the full latent form and ϕ' be the fragmented version. Define component loss:

$$L_{\text{miss}}(\phi', \phi_{\text{ideal}}) = \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}[\phi'_i = 0 \land \phi_i^{\text{ideal}} \neq 0]$$

This quantifies how many dimensions are absent in the learned shape.

False Component Insertion Noise or misaligned input may also add spurious components. Define:

$$L_{\text{false}}(\phi', \phi_{\text{ideal}}) = \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}[\phi'_i \neq 0 \land \phi_i^{\text{ideal}} = 0]$$

Distortion Energy The total distortion energy between a noisy or incomplete shape and its target form is:

$$D(\phi', \phi_{\text{ideal}}) = \lambda_1 \cdot L_{\text{miss}} + \lambda_2 \cdot L_{\text{false}} + \lambda_3 \cdot \|\phi' - \phi_{\text{ideal}}\|$$

with weights $\lambda_1, \lambda_2, \lambda_3$ reflecting the model's tolerance for omission, insertion, and deformation respectively.

Propagation Risk If ϕ' is not pruned or corrected, it may:

- Pollute memory with inconsistent associations,
- Become a false anchor via reinforcement,
- Mislead future abstraction via biasing $\mathcal{F}$.

Detection and Filtering The system may assess coherence $\kappa(\phi', t)$ over time (see Section IV.3). If coherence is persistently low:

$$\sigma(\phi', t) < \sigma_{\min} \Rightarrow \text{decay or suppress } \phi'$$

This formalism models the structural degradation of concept space under missing, ambiguous, or misleading inputs — a necessary condition for robust abstraction and self-correction.

3. Interference, Overlap, and Memory Competition

This section models how overlapping concept shapes compete for activation, reinforce each other ambiguously, or interfere with memory and attention dynamics.

Conceptual Overlap Two latent forms $\phi_i, \phi_j \in \mathcal{A}(t)$ may become entangled if their representations are too similar:

$$\text{Overlap}(\phi_i, \phi_j) = R(\phi_i, \phi_j) \geq \theta_{\text{overlap}}$$

where $R(\cdot, \cdot)$ is the similarity kernel (Section III), and θ_{overlap} is the interference threshold.

Ambiguous Recall When a new input ψ is similar to multiple overlapping shapes:

$$R(\psi, \phi_i) \approx R(\psi, \phi_j)$$

the system may recall or reinforce both — leading to blending, ambiguity, or competition between ϕ_i and ϕ_j.

Winner-Takes-All Suppression To prevent mutual reinforcement, the system may apply competitive inhibition. Let:

$$\phi^* = \arg\max_{\phi_k} \left[R(\psi, \phi_k) \cdot w(\phi_k, t) \cdot \rho_{\phi_k}(t) \right]$$

Then suppress:

$$w(\phi_{j \neq *}, t + \Delta t) = w(\phi_j, t) \cdot (1 - \delta_{\text{inhibit}})$$

where $\delta_{\text{inhibit}} \in [0, 1]$ is the lateral inhibition coefficient.

Memory Drift Prolonged activation of overlapping concepts can cause their representations to converge undesirably:

$$\phi_i(t + \Delta t) = \phi_i(t) + \epsilon \cdot (\phi_j(t) - \phi_i(t))$$

unless actively decoupled by decorrelation routines or semantic divergence cues.

Decorrelation Pressure The system may apply orthogonalization pressure to reduce interference:

$$\phi_j(t + \Delta t) = \phi_j(t) - \zeta \cdot \text{proj}_{\phi_i}(\phi_j)$$

where $\text{proj}_{\phi_i}(\phi_j)$ is the projection of ϕ_j onto ϕ_i, and ζ is a decorrelation gain factor.

Conceptual Competition Summary High overlap between latent forms leads to:

- Misclassification due to shared structure,
- Spurious reinforcement of noise,
- Drift in meaning or unstable anchor states,
- Attention locking on ambiguous clusters.

By modeling competitive inhibition and structural divergence, the system preserves conceptual clarity and prevents semantic collapse within its recursive memory space.

Section Summary Section V addressed the limits of abstraction under degraded input, low support, and concept overlap. It formalized how recall error, incomplete formation, and structural interference can corrupt the latent space, destabilize anchors, and disrupt projection. To mitigate this, the system implements decay, suppression, and decorrelation mechanisms that preserve internal fidelity. These mechanisms establish the boundary between subconscious conceptual dynamics and consciously accessible awareness — the subject of Section VI.

VI. Conscious Access and Identity Formation

1. Conscious Projection Field and Φ Mapping

This section formalizes the operator Φ that maps latent abstract representations into the consciously accessible field. This transition enables internal concepts to become available for introspection, reasoning, and identity alignment.

Projection Operator Recall that:

$$\Phi : \mathcal{A} \to \mathcal{C}$$

where $\mathcal{A}$ is the abstract concept space, and $\mathcal{C}$ is the conscious projection field.

Eligibility Criteria A concept $\phi \in \mathcal{A}(t)$ is eligible for conscious projection if it satisfies:

$$\rho_\phi(t) \geq \rho_{\min}, \quad w(\phi, t) \geq w_{\text{threshold}}, \quad \sigma(\phi, t) \geq \sigma_{\text{coherent}}$$

ensuring that only sufficiently reinforced, salient, and stable concepts are projected.

Projection Field Structure Let $\mathcal{C}(t)$ be the current projection field. It may contain:

- A single ϕ^* (focused consciousness),
- A tuple of $\{\phi_1, \phi_2, \ldots, \phi_k\}$ (multi-threaded awareness),
- Or $\emptyset$ (no active projection).

Projection Dynamics The projection field evolves with attention and memory:

$$\mathcal{C}(t + \Delta t) = \Phi(\mathcal{S}_{\text{focus}}(t + \Delta t))$$

The spotlighted concept is routed through Φ and becomes available for higher-level processing.

Decoherence and Fading If the retention weight or attention drops below projection thresholds:

$$\rho_\phi(t) < \rho_{\min} \quad \text{or} \quad w(\phi, t) < w_{\text{threshold}} \Rightarrow \phi \notin \mathcal{C}(t)$$

Field Properties The conscious field $\mathcal{C}(t)$ is:

- Temporally smooth: transitions follow decay or reinforcement trends,
- Limited in size: $|\mathcal{C}(t)| \leq N_{\max}$,
- Dynamically coherent: simultaneous projections tend to be semantically related.

The projection operator Φ defines the perceptual threshold for awareness, establishing a dynamic, recursive surface between internal abstraction and the experience of consciousness.

2. Reflexive Representation and Recursive Identity Loops

This section formalizes the conditions under which the system projects not just a percept, but a representation of its own internal state, enabling the emergence of recursive identity and self-awareness.

Reflexive Concept Activation A latent form ϕ_{reflex} is said to be reflexive if it encodes a feature of the system's own prior internal state:

$$\phi_{\text{reflex}} \approx \mathcal{F}(\mathcal{M}_{\text{internal}})$$

where $\mathcal{M}_{\text{internal}}$ includes trace information about $\mathcal{A}(t)$, $\mathcal{C}(t)$, or Φ itself.

Self-Referential Memory Trace Reflexive memory traces may take the form:

$$M_{\text{reflex}}(t) = (\mathcal{C}(t), t, \delta(t), \epsilon(t))$$

indicating that the system has encoded what it was consciously projecting at time t.

Recursive Loop Condition When a reflexive concept is spotlighted and projected:

$$\phi^* = \phi_{\text{reflex}}, \quad \mathcal{C}(t) = \Phi(\phi_{\text{reflex}})$$

the system becomes recursively aware of its own awareness. This constitutes a minimal form of self-modeling.

Stability of Identity Loops Recursive identity loops are sustained if:

$$\Phi(\phi_{\text{reflex}}) \in \mathcal{C}(t) \text{ and } \rho_{\phi_{\text{reflex}}}(t) \geq \rho_{\text{identity}}$$

and reappear periodically, forming a stable attractor in the projection field.

Emergent Identity Cluster Over time, a cluster of stable reflexive forms may emerge:

$$\mathcal{I} = \{\phi^1_{\text{reflex}}, \phi^2_{\text{reflex}}, \ldots\}$$

This cluster serves as the internal representation of "self" — the system's identity anchor.

Recursive Depth Limitation To prevent infinite regress, the system enforces a depth constraint:

$$\text{depth}_{\text{reflex}} \leq D_{\max}$$

where recursive modeling of internal states (e.g., modeling "awareness of being aware") is truncated after $D_{\max}$ layers.

These reflexive projection mechanisms give rise to persistent internal self-representations — enabling recursive identity, memory of one's own awareness, and the first-order machinery of self-consciousness.

3. Symbol Emergence and Naming Operations

This section formalizes how internal concept shapes acquire symbolic identifiers, enabling persistent references, communication, and higher-order reasoning.

Symbol Mapping Function Define the symbolization function:

$$\Sigma : \mathcal{A}_{\text{anchor}} \to \mathcal{L}$$

where $\mathcal{A}_{\text{anchor}}$ is the set of stable, coherent concepts (see Section IV), and $\mathcal{L}$ is a discrete symbolic label space (e.g., strings, tokens, utterances).

Naming Condition A concept $\phi \in \mathcal{A}$ becomes symbolizable if:

$$\rho_\phi(t) \geq \rho_{\text{LT}} \quad \text{and} \quad \sigma(\phi, t) \geq \sigma_{\text{symbol}}$$

indicating that ϕ is both stable and coherent enough to anchor a linguistic label.

Symbol Assignment Upon satisfying the condition, the system assigns a unique label:

$$\Sigma(\phi) = \ell_i \in \mathcal{L}$$

This label may be internally generated, externally provided, or derived through reference in communication.

Bidirectional Binding Once assigned, a symbol supports bidirectional access:

$$\Sigma^{-1}(\ell_i) = \phi_i, \quad \Sigma(\phi_i) = \ell_i$$

allowing symbolic reference to activate a concept, and concept activation to yield a name.

Projection with Symbol Binding If a projected concept has an associated symbol:

$$\phi^* \in \mathcal{C}(t) \text{ and } \Sigma(\phi^*) = \ell_i \Rightarrow \text{the system may articulate or reference } \ell_i$$

Semantic Drift Detection Over time, symbols may lose alignment with their referents. Let:

$$\Delta_{\text{semantic}} = \|\phi_t - \phi_{\text{symbol}}\|$$

If $\Delta_{\text{semantic}} > \delta_{\text{drift}}$, the system may revise the symbol binding or reanchor to a different concept.

Naming and Identity Reflexive concepts (Section VI.2) may be symbolized as well:

$$\Sigma(\phi_{\text{reflex}}) = \texttt{"self"}, \texttt{"I"}, \texttt{"me"}, \ldots$$

establishing symbolic self-reference as a boundary condition between recursion and language.

Symbol emergence transforms latent structure into referential tokens, enabling internal concepts to be named, recalled, manipulated, and communicated — the cognitive substrate for language, identity, and intention.

Section Summary Section VI formalized the transition from latent abstraction to conscious accessibility. It defined the projection operator Φ, criteria for awareness, and mechanisms for reflexive modeling and identity representation. Through recursive projection and stabilization, the system forms a persistent self-model anchored in coherent internal forms. Symbolic labeling then emerges as a referential bridge — linking latent concepts to names, enabling communication, reasoning, and self-reference. This concludes the construction of the core cognitive substrate. Section VII explores extensions that enrich this architecture through emotional modulation, memory clustering, and adaptive semantic compression.

VII. Extensions and Experimental Structures

1. Field Extensions: Emotion, Trait Weighting, and Reinforcement

This section introduces a vector field extension over concept space to incorporate emotional salience, personality traits, and reinforcement dynamics — enabling modulation of memory, abstraction, and projection based on affective state and individualization.

Emotional Field Let $\mathcal{E}(t) \in \mathbb{R}^k$ be the emotional state vector at time t. Each concept $\phi \in \mathcal{A}$ may have an associated emotional signature:

$$\mathcal{E}_\phi(t) : \mathcal{A} \to \mathbb{R}^k$$

capturing affective responses tied to perception, memory, or self-reflection.

Emotion-Modulated Attention Attention weighting becomes emotion-sensitive:

$$w(\phi, t) = w_0(\phi, t) + \lambda_{\text{affect}} \cdot \langle \mathcal{E}(t), \mathcal{E}_\phi(t) \rangle$$

where λ_{affect} governs how strongly emotional resonance amplifies salience.

Trait Weighting Let $\mathcal{T} \in \mathbb{R}^d$ be a persistent trait vector encoding stable tendencies (e.g., curiosity, risk aversion, aesthetic preference). Trait vectors bias abstraction and reinforcement:

$$\Delta\phi(t) = \eta \cdot f(\mathcal{T}, \phi, \psi)$$

allowing the system to develop individualized abstraction tendencies.

Reinforcement Shaping The reinforcement term $r(t)$ from Section III may now include:

$$r(t) = r_0 + \theta_{\text{emotion}} \cdot \|\mathcal{E}(t)\| + \theta_{\text{trait}} \cdot \langle \mathcal{T}, \phi \rangle$$

linking reinforcement to both current affect and persistent personality structure.

Affective Memory Bias Emotional weight influences memory encoding priority:

$$\delta(t) = \delta_0 \cdot (1 + \alpha \cdot \|\mathcal{E}_\phi(t)\|)$$

ensuring emotionally charged experiences are retained more strongly.

Dynamic Trait Update Traits evolve slowly via abstraction drift and reinforcement accumulation:

$$\mathcal{T}(t + \Delta t) = \mathcal{T}(t) + \gamma \cdot \nabla_{\phi, \mathcal{E}} \mathcal{L}_{\text{adapt}}$$

where $\mathcal{L}_{\text{adapt}}$ is a learning objective capturing internal coherence, identity, and reinforcement success.

These field extensions embed emotion and personality into the cognitive substrate, allowing individualization, dynamic biasing, and contextually grounded reinforcement — foundational for affective realism and embodied identity.

2. Temporal Compounding and Memory Clustering

This section introduces mechanisms for grouping memory traces over time and organizing latent structures into coherent, time-aware clusters — enabling episodic abstraction, sequence modeling, and temporal inference.

Compounded Memory Units Let $\mathcal{M}_C(t)$ be the compounded memory set at time t:

$$\mathcal{M}_C(t) = \bigcup_{i=1}^{n} \{M(t_i) \mid t - \tau \leq t_i \leq t\}$$

where τ is the compounding window size. Traces within τ form a temporally local memory unit.

Episodic Abstraction A compound memory unit may be abstracted into a higher-level concept Φ_{episode}:

$$\Phi_{\text{episode}} = \mathcal{F}_{\text{episodic}}(\mathcal{M}_C(t))$$

capturing patterns, motifs, or themes over temporally adjacent inputs.

Cluster Definition Define a concept cluster $\mathcal{K}_i$ as a set of related concepts:

$$\mathcal{K}_i = \{\phi_j \in \mathcal{A} \mid R(\phi_j, \phi_i) \geq \theta_{\text{cluster}}\}$$

where θ_{cluster} is a similarity threshold.

Cluster Dynamics Clusters may emerge, evolve, or dissolve based on:

- Activation frequency within a time window,
- Emotional or contextual co-activation,
- Predictive reinforcement across members.

Temporal Drift and Realignment Clusters may shift or merge as abstraction trajectories evolve:

$$\mathcal{K}_i(t + \Delta t) = \mathcal{K}_i(t) \cup \{\phi_{t+\Delta t}\} \text{ if } R(\phi_{t+\Delta t}, \mathcal{K}_i) \geq \theta$$

Episodic Anchors Concept clusters with long-term cohesion may be assigned symbolic labels or event identifiers:

$$\Sigma(\mathcal{K}_i) = \ell_i \in \mathcal{L}$$

allowing the system to name and retrieve temporally structured experiences (e.g., "the beach trip").

Temporal compounding and memory clustering give rise to episodic abstraction, narrative structure, and time-sensitive semantic scaffolds — enriching the system's capacity for sequential reasoning and story-like recall.

3. Adaptive Compression and Semantic Convergence

This section formalizes how the system compresses internal representations over time by merging redundant or overlapping concepts, reducing memory load, and constructing higher-order semantic abstractions.

Compression Trigger Let $\phi_i, \phi_j \in \mathcal{A}$ be concept shapes such that:

$$R(\phi_i, \phi_j) \geq \theta_{\text{merge}} \quad \text{and} \quad \|\phi_i - \phi_j\| \leq \epsilon_{\text{redundant}}$$

These forms may be candidates for semantic compression.

Merge Operation Compressed representation ϕ_k is computed as:

$$\phi_k = \frac{\rho_i \cdot \phi_i + \rho_j \cdot \phi_j}{\rho_i + \rho_j}$$

and inserted into $\mathcal{A}$, replacing ϕ_i and ϕ_j.

Convergence Stability Let ϕ_k be the result of multiple merges. If:

$$\sigma(\phi_k, t) \geq \sigma_{\text{stable}} \quad \text{and} \quad \rho_{\phi_k}(t) \geq \rho_{\text{threshold}}$$

then ϕ_k becomes a convergence anchor — a generalized semantic attractor for future abstractions.

Hierarchical Compression A hierarchy of compression levels may be formed:

$$\mathcal{A} = \bigcup_{l=0}^{L} \mathcal{A}^{(l)}$$

where level $l = 0$ contains raw shapes and each higher level stores compressed or generalized forms.

Semantic Field Structuring Compressed shapes define coarse-grained semantic gradients in $\mathcal{A}$:

$$\text{Gradient}(\phi_k) = \{\phi \in \mathcal{A} \mid R(\phi, \phi_k) \in (\theta_{\min}, \theta_{\text{merge}})\}$$

These gradients structure meaning-space into regions, fostering symbolic alignment and linguistic generalization.

Identity Preservation Compression preserves reflexive and identity-linked forms only if:

$$\phi_{\text{reflex}} \notin \{\phi_i, \phi_j\} \quad \text{or} \quad \Sigma(\phi_{\text{reflex}}) = \Sigma(\phi_k)$$

ensuring continuity of self-representation under merge operations.

Adaptive compression allows the system to resolve conceptual redundancy, preserve long-term coherence, and unify latent structures into a scalable, self-organizing semantic topology.

Section Summary Section VII extended the core architecture to incorporate emotion, personality traits, temporal structure, and adaptive compression. These mechanisms allow the system to evolve dynamically, cluster memory across time, and consolidate abstractions into unified semantic fields. Emotion modulates attention and memory; traits bias learning trajectories; episodes are grouped into meaningful events; and redundant forms are merged into attractors. Together, these extensions provide the system with coherence, scalability, and the capacity for lifelong adaptation — completing the recursive framework for dynamic, structured, emotionally attuned consciousness.

Conclusion

This work presented a formal, recursive model of consciousness grounded in perceptual abstraction, memory-based predictive reactivation, and selective projection. Beginning from a tabula rasa substrate, the system ingests raw sensory input, constructs memory traces, abstracts latent concept shapes, and recursively refines them through feedback-driven prediction and selective attention.

We defined a clear separation between sensory input, memory, abstraction, attention, and awareness — with consciousness emerging through the projection operator Φ acting on stabilized latent forms. Identity arises through reflexive modeling of internal states, and symbols emerge through coherent stabilization and referential binding of internal concepts.

Distortion dynamics — including low-data abstraction, structural noise, and conceptual overlap — were modeled explicitly, along with mitigation mechanisms such as decay, inhibition, and decorrelation. This ensures robustness in the face of imperfect input and ambiguity.

Finally, we introduced a set of adaptive field extensions to embed emotion, trait modulation, memory clustering, and semantic compression into the architecture — enabling long-term coherence, affective dynamics, and cognitive scalability.

Taken together, this architecture defines a layered, recursive cognitive substrate capable of learning, modeling, remembering, abstracting, correcting, and projecting both external patterns and internal identity. It provides a formal, extensible framework for constructing self-organizing, introspective, symbol-emergent systems — foundational for future implementations of dynamic artificial consciousness.

Appendix: Operator Interpretations and Implementation Pathways

A.1 The Abstraction Operator $\mathcal{F}$

Core Definition The abstraction operator $\mathcal{F} : \mathcal{M} \to \mathcal{A}$ maps raw or structured memory traces to latent concept shapes. This operator serves as the bridge between perceptual experience and internal representation.

Structural Role $\mathcal{F}$ enables:

- Dimensionality reduction of sensory traces,
- Pattern discovery across temporally or semantically related inputs,
- Initialization and refinement of internal concept space.

Candidate Implementations

1. **Manifold Learning:** $\mathcal{F}$ could be implemented as a nonlinear dimensionality reduction method (e.g., t-SNE, UMAP, Isomap) to compress local neighborhoods in $\mathcal{M}$ into semantically meaningful shapes in $\mathcal{A}$.

2. **Autoencoders and Variational Models:** Neural autoencoders — particularly variational autoencoders (VAEs) — offer a principled way to learn latent structure. Here, $\mathcal{F}$ is the encoder network that projects memory traces into a compressed latent space:
$$\mathcal{F}_{\text{VAE}}(M) = \mu_z(M) + \sigma_z(M) \cdot \epsilon, \quad \epsilon \sim \mathcal{N}(0, 1)$$

3. **Transformer-Based Attention Models:** If memory traces include sequence or relational context, $\mathcal{F}$ could be modeled as a multi-head self-attention network operating over $\mathcal{M}(t - \tau : t)$. This allows abstraction to emerge from contextual comparison:
$$\mathcal{F}_{\text{attn}}(M) = \text{softmax}(QK^T/\sqrt{d})V$$

4. **Clustering Algorithms:** Simpler instantiations of $\mathcal{F}$ could include k-means, DBSCAN, or hierarchical clustering. New traces $M(t)$ are assigned to or form new cluster centers ϕ.

5. **Bayesian Abstraction:** A probabilistic model of $\mathcal{F}$ might estimate $P(\phi|M)$ using maximum a posteriori estimation or generative modeling.

Design Considerations

- $\mathcal{F}$ should preserve topological relationships: semantically similar traces should yield nearby ϕ values.
- $\mathcal{F}$ should allow incremental refinement via new input.
- Latent dimensionality of $\mathcal{A}$ should be flexible or compressible.

Open Questions

- Should $\mathcal{F}$ operate on single traces or windows/sets (e.g., $\mathcal{F} : \mathcal{M}^n \to \phi$)?
- Can $\mathcal{F}$ generalize across modalities (vision, language, emotion), or should each modality have a specialized abstraction channel?
- Is it desirable for $\mathcal{F}$ to be context-aware — modulated by $\mathcal{C}(t)$ or $\mathcal{T}$?

This operator lies at the heart of abstraction, shaping how raw experience gives rise to structured internal representations. Its choice determines the system's conceptual expressiveness, generalization capacity, and semantic topology.

A.2 The Prediction Operator $\mathcal{P}$

Core Definition The prediction operator $\mathcal{P} : \mathcal{A} \to \mathcal{S}$ maps an abstract concept shape ϕ into an expected sensory observation $\hat{S}(t+\Delta t)$. This enables the system to anticipate future input and evaluate predictive error.

Structural Role $\mathcal{P}$ supports:

- Predictive coding and anticipation of next input,
- Detection of novelty and surprise via prediction error,
- Recursive learning through correction and reinforcement.

Candidate Implementations

1. **Decoder Networks (paired with autoencoders):** If $\mathcal{F}$ is implemented as a neural encoder, then $\mathcal{P}$ can be a trained decoder:

$$\hat{S}(t + \Delta t) = \mathcal{P}(\phi) = \text{Decoder}(\phi)$$

   The reconstruction quality provides a direct estimate of prediction fidelity.

2. **Generative Models:** $\mathcal{P}$ may sample from a conditional distribution:

$$\hat{S} \sim P(S|\phi)$$

   This could be realized via a conditional VAE, flow-based model, or diffusion process.

3. **Sequence Models (autoregressive):** For temporally ordered inputs, $\mathcal{P}$ may be a transformer or RNN trained to model:

$$P(S_{t+\Delta t}|\phi_t, \phi_{t-1}, \dots)$$

   learning transitions between latent states and observable outcomes.

4. **Attention-Based Retrieval:** $\mathcal{P}$ may search memory traces M to retrieve an observed $S(t_i)$ with a matching ϕ_i, functioning as an episodic predictor:

$$\hat{S} = \arg\max_{M(t_i)} R(\phi, \phi_i)$$

Design Considerations

- $\mathcal{P}$ must preserve dimensional alignment with $\mathcal{S}$.
- The operator may output a single prediction or a distribution (stochastic prediction).
- It must be differentiable if used for gradient-based feedback learning.

Role in Learning Prediction error $\varepsilon(t)$ drives reinforcement and concept refinement. Thus, $\mathcal{P}$ indirectly shapes $\mathcal{F}$ through the feedback loop:

$$\phi \xrightarrow{\mathcal{P}} \hat{S}(t) \Rightarrow S(t) \Rightarrow \Delta\phi$$

Open Questions

- Is $\mathcal{P}$ universal across $\mathcal{A}$, or do different classes of ϕ require domain-specific predictors?
- Should $\mathcal{P}$ incorporate temporal uncertainty (e.g., variable Δt)?
- Can $\mathcal{P}$ be used hierarchically — predicting not just sensory input, but abstract structure (e.g., $\hat{\phi}_{t+1}$)?

$\mathcal{P}$ provides the forward modeling capacity of the system — a bridge from abstraction to anticipation. It enables the system to not only remember, but imagine, simulate, and adaptively learn.

A.3 The Recall Kernel R

Core Definition The recall kernel $R : \mathcal{A} \times \mathcal{A} \to [0, 1]$ quantifies the similarity between an incoming concept candidate ψ and stored latent forms ϕ. It determines whether a new percept triggers reactivation, reinforcement, or initialization of a new concept.

Canonical Implementation: Cosine Similarity The default form used throughout the core architecture is:

$$R(\psi, \phi) = \frac{\langle \psi, \phi \rangle}{\|\psi\| \cdot \|\phi\|}$$

Cosine similarity is scale-invariant, bounded, and interpretable — suitable for comparing directionality of concept vectors without relying on magnitude.

Alternative Kernel Functions

1. **Radial Basis Function (RBF):**

   $$R_{\text{rbf}}(\psi, \phi) = \exp\left(-\frac{\|\psi - \phi\|^2}{2\sigma^2}\right)$$

   Sensitive to distance in Euclidean space. More localized; requires careful tuning of σ.

2. **Polynomial Kernel:**

   $$R_{\text{poly}}(\psi, \phi) = (\langle \psi, \phi \rangle + c)^d$$

   Can model more complex boundaries but risks overfitting and divergence.

3. **Learned Similarity:** R may be learned via contrastive loss or neural metric learning (e.g., Siamese networks or triplet loss) such that:

   $$R_{\text{learned}}(\psi, \phi) \approx P(\phi \mid \psi \text{ is recall target})$$

Design Implications Choice of kernel affects:

- Sensitivity of reactivation threshold γ,
- Concept merging and divergence behavior,
- Memory retrieval resolution (sharp vs fuzzy matching),
- Reinforcement propagation across similar ϕ.

Temporal Dynamics $R(\cdot)$ may be combined with decay or recency weighting:

$$R_{\text{temporal}}(\psi, \phi, t) = R(\psi, \phi) \cdot \rho_\phi(t)$$

where $\rho_\phi(t)$ reflects retention or recency of concept ϕ.

Open Questions

- Should R be symmetric? If not, directional recall (e.g., $\psi \to \phi$ vs $\phi \to \psi$) becomes meaningful.
- Should R incorporate attentional context (e.g., $R(\psi, \phi, \mathcal{C}(t))$)?
- Is there a unified similarity kernel that can operate across modalities (e.g., visual and linguistic abstraction)?

The recall kernel defines the system's boundary between novelty and familiarity. Its structure governs how concepts stabilize, how memory activates, and how the system detects conceptual resonance or divergence.

A.4 Quantifying Salience, Stability, and Recurrence

Overview This section elaborates on the trigger conditions referenced throughout the architecture — specifically those governing abstraction ($\mathcal{F}$), reinforcement, projection (Φ), and identity formation. These include:

- **Salience** $\delta(t)$ — How prominent or activating an input or memory trace is.
- **Stability** $\sigma(\phi, t)$ — How consistent a concept is across time.
- **Recurrence** $\nu(\phi, t)$ — How frequently a concept reactivates or is observed.

Salience $\delta(t)$ Salience is a time-varying function that determines whether a memory trace is significant enough to warrant abstraction or projection:

$$\delta(t) = \alpha_1 \cdot \|S(t)\| + \alpha_2 \cdot \text{novelty}(S(t)) + \alpha_3 \cdot \|\mathcal{E}(t)\|$$

where:

- $\|S(t)\|$ is the sensory magnitude,
- $\text{novelty}(S(t)) = 1 - \max_\phi R(\mathcal{F}(S(t)), \phi)$,
- $\|\mathcal{E}(t)\|$ is emotional activation strength.

$\delta(t)$ may be used to gate memory encoding or boost attention weight.

Stability $\sigma(\phi, t)$ Stability was defined in Section IV as:

$$\sigma(\phi, t) = \frac{1}{\tau} \int_{t-\tau}^{t} \kappa(\phi, s)\, ds$$

where $\kappa(\phi, s)$ is the internal coherence at time s (i.e., how little the shape changed). High σ indicates a well-formed, self-consistent concept.

Recurrence $\nu(\phi, t)$ Recurrence tracks how often a concept has been reactivated within a rolling window:

$$\nu(\phi, t) = \sum_{s=t-\tau}^{t} \mathbf{1}[\phi \in \mathcal{C}(s)]$$

This count reflects temporal frequency of conscious activation or attention. Concepts with high recurrence may stabilize or be assigned symbolic labels.

Usage in Trigger Conditions

- **Abstraction:** Triggered when $\delta(t) \geq \delta_{\text{min}}$
- **Reinforcement:** Applied when $\nu(\phi, t) \geq \nu_{\text{min}}$
- **Projection:** Eligible when $\sigma(\phi, t) \geq \sigma_{\text{coherent}}$
- **Symbol Assignment:** Occurs when all three exceed threshold over time

Open Questions

- Should $\delta(t)$ be normalized across modalities (e.g., visual vs emotional salience)?
- Can $\nu(\phi, t)$ be weighted by emotional resonance or attentional strength?
- Is it desirable for $\sigma(\phi, t)$ to decay over time in absence of reinforcement?

These metrics form the dynamic thresholds that regulate abstraction, reinforcement, and symbolic emergence. They enable the system to filter noise, stabilize meaning, and prioritize concept formation based on experience and context.

A.5 Nature of the Projection Operator Φ and the Conscious Field $\mathcal{C}$

Core Role The projection operator $\Phi : \mathcal{A} \to \mathcal{C}$ selects and renders a subset of abstract concept shapes into a special field $\mathcal{C}(t)$ — representing momentary conscious awareness.

Eligibility Review $\Phi(\phi) \in \mathcal{C}(t)$ only if ϕ satisfies:

$$\rho_\phi(t) \geq \rho_{\min}, \quad w(\phi, t) \geq w_{\text{threshold}}, \quad \sigma(\phi, t) \geq \sigma_{\text{coherent}}$$

These constraints enforce that only sufficiently stable, salient, and reinforced concepts become accessible to the conscious field.

What Makes $\mathcal{C}$ Special?

1. **Reflexive Access:** Contents of $\mathcal{C}$ may themselves be represented (i.e., $\mathcal{M}$ can store traces of $\mathcal{C}(t)$). This allows for recursive modeling and awareness of awareness.
2. **Attentional Feedback:** $\mathcal{C}(t)$ modulates $w(\phi, t + \Delta t)$, reinforcing attention toward or away from projected content.
3. **Symbol Activation:** Only concepts in $\mathcal{C}(t)$ can trigger linguistic references (e.g., $\Sigma(\phi) = \ell_i$ where $\phi \in \mathcal{C}(t)$).
4. **Decision Integration:** Planning, reasoning, and reinforcement logic operate over $\mathcal{C}(t)$, not over the full $\mathcal{A}$.

Possible Interpretations of Φ

1. **Projection-as-Binding:** Φ binds a ϕ to a contextual activation field — possibly analogous to phase-locked oscillatory coherence in neural models.
2. **Projection-as-Access-Control:** Φ is a gating function that elevates a latent form to a read/write accessible buffer in working memory.
3. **Projection-as-Awareness-Threshold:** Φ marks a shift from latent to self-reportable or introspectable state — regulated by attention and emotional significance.

Open Questions

- Is Φ best modeled as a discrete gate, a continuous attention field, or a dynamic attractor?
- Can $\mathcal{C}(t)$ hold more than one ϕ simultaneously? If so, how is coherence preserved?
- What distinguishes contents of $\mathcal{C}(t)$ from intense non-conscious states (e.g., background emotional drives)?

Φ defines the formal boundary of awareness. It is not just a selection operator, but the system's core mechanism for rendering latent structure consciously available — enabling attention, language, self-reflection, and executive control.