A MATHEMATICAL THEORY OF AGI

Anonymous authors
Paper under double-blind review

ABSTRACT

We present a comprehensive synopsis of the mathematical theory (universal algebra, category and sheaf theory) of Artificial General Intelligence (AGI). The theory is built upon three complementary and ultimately equivalent pillars: (1) an **axiomatic framework** formalizing infinite metacognition via approximate homomorphisms and category-theoretic limits; (2) a **cognitive free-energy principle** defined over generalized cognitive hierarchies; and (3) the **Natural Balance Principle**, a single-axiom optimization functional that requires no prior structural assumptions. All three frameworks independently converge to the same unique architecture: an AGI system must possess a minimal algebraic carrier isomorphic to $\mathcal{B}_4^{k_{\min}}$ (the four-element Boolean algebra), necessarily embeds a unique self-referential prediction module (*MetaPred*) that accompanies every learning process, and exhibits a strict isomorphism among all machine learning mechanisms. We provide algorithmic pseudocode for MetaPred and its mandatory self-measurement cycle, demonstrating concrete implementability. Furthermore, we detail a profound structural alignment between the derived cognitive architecture and well-established findings in neuroscience (predictive coding, global workspace, prefrontal metacognitive networks) and cognitive psychology (dual-process theory, exploration–exploitation). Finally, the theory's variational formulation reveals a deep structural isomorphism with fundamental physics, including a cognitive Einstein equation and a least-action principle. The convergence of the three theoretical pillars, coupled with cross-disciplinary consistency, provides strong evidence that the initial cognitive intuitions capture the mathematical essence of general intelligence[1].

1 INTRODUCTION

The meteoric success of GPT-style large language models (LLMs) has bred a precarious illusion: that scaling autoregressive architectures will inevitably deliver AGI, a belief sustained only by the theoretically opaque notion of "emergence." This has split the field into two camps—scaling proponents and advocates of architectural alternatives such as world-models (Chollet, 2019; Russell, 2019)—who cannot persuade one another because *neither possesses a theoretically justified definition of the goal*. Compounding this void, the economics of scaling make ever-larger models increasingly perverse: if the scaling hypothesis is right, AGI becomes economically intractable; if wrong, the investment is a historic misallocation. In both cases the absence of a principled "how much is enough" criterion turns AGI research into a journey toward a mirage.

The root difficulty is conceptual, not engineering: no mathematically endogenous, semantically disambiguated definition of "intelligence" or "cognition" exists at the generality AGI demands. Existing definitions are either functional and operational (e.g., expected reward over environments (Legg & Hutter, 2007; Hutter, 2005; Ng, 2025)) or constrained to a particular machine-learning paradigm (e.g., Bayesian inference (Friston, 2010; He et al., 2025)); the latter, while excellent within their scope, sacrifice generality precisely where it is needed, because without an endogenous definition there is no guarantee that the chosen paradigm is the *right* one for unlimited cognition.

This paper therefore starts where the problem starts: with the *definition*. We develop a mathematical theory in which AGI is defined not by what a system achieves but by what it *is*—a system whose

[1]For the first engineering implementation and validation of the model derived from this paper, see (Anonymous, 2026).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

metacognitive hierarchy converges, in the limit of infinite reflection, to a structure isomorphic to the world. From this intrinsic definition we derive, as necessary consequences, the minimal algebraic core $\mathcal{B}_4^{k_{\min}}$, the self-referential learning mechanism (MetaPred) that must accompany every act of cognition, and the variational principles that render AGI optimal.

The intellectual lineage spans the history of formal intelligence definitions. The most influential remains theLegg & Hutter (2007) universal intelligence measure and its AIXI realisation (Hutter, 2005), recently extended to quantum AIXI (Perrier, 2025); yet these define intelligence behaviourally. A growing 2025 literature challenges behaviourism: Ng (2025) propose "entity fidelity" as a unifying operational criterion; Mappouras & Rossides (2025) prove principled upper bounds on machine-computable AGI, underscoring the necessity of structural definitions. Architectural proposals such as Subasioglu & Subasioglu (2025)'s six-component True Intelligence and five-level taxonomy, He et al. (2025)'s (C, U, V) formulation, and Thomas (2025)'s Causal Mechanics enrich the landscape, as do broad surveys (Mumuni & Mumuni, 2025; Qureshi et al., 2025) and open-ended tests like SuperARC (Hernández-Espinosa et al., 2026). Notably, Sanchez (2025)'s Recursive Convergence Hypothesis posits that recursive self-improvement may transition from simulating to instantiating sentience—a thesis consonant with our view that self-reference is a structural prerequisite for unlimited cognition.

These contributions signal a convergence toward intrinsic, structure-based characterizations of general intelligence, a direction the present theory formalises and extends. Unlike prior work, we do not merely *propose* a definition; we *derive* one from minimal cognitive axioms, prove equivalence among multiple formulations, and show that any system meeting the definition must instantiate a rigid architecture. While AGI has historically lacked a mathematically grounded definition—with most characterizations behavioural or architectural—this paper fills that gap by formalizing three core intuitions: dynamic operation preservation, bidirectional interaction consistency, and infinite recursivity. The theory unfolds in three equivalent pillars (axiomatic system, cognitive free-energy principle, Natural Balance Principle); this part details the axiomatic pillar, establishing AGI as infinite metacognitive hierarchy convergence, the equivalence theorems linking infinite cognition to self-referential closure, the minimal algebraic carrier $\mathcal{B}_4^{k_{\min}}$, and the necessary emergence of the unique self-referential MetaPred that unifies all learning mechanisms, supported by algorithmic pseudocode and initial steps toward multimodal cognition.

2 The AGI Axiomatic Foundation

2.1 Approximate Homomorphisms and Cognitive Hierarchies

Let $\mathcal{M}$ be a finite set (the AGI system) and $\mathcal{W}$ be an infinite set (the external world). Both are equipped with a universal-algebraic structure of operation type $\tau : \Omega \to \mathbb{N}^+$. A mapping $\varphi : A \to B$ is an *ε-approximate homomorphism* if for every k-ary operation $\omega \in \Omega$ and all $a_1, \ldots, a_k \in A$, $d_B(\omega_B(\varphi(a_1), \ldots, \varphi(a_k)),\ \varphi(\omega_A(a_1, \ldots, a_k))) \leq \varepsilon$. The pseudometrics d_A, d_B are required to be uniformly Lipschitz for all operations. This captures the fact that cognition never perfectly mirrors the world's dynamics but can approximate them arbitrarily well.

Definition 2.1 (Cognitive Hierarchy)**.** *The **cognitive hierarchy** $\{\mathbb{H}_n\}_{n\geq 0}$ is recursively defined by:*

$$\mathbb{H}_0 = \mathcal{M}, \tag{1}$$

$$\mathbb{H}_{n+1} = \left\{(\pi, \sigma, \gamma) \mid \pi : \mathcal{M} \Rightarrow \varepsilon_n^{\pi}\mathbb{H}_n,\ \sigma : \mathbb{H}_n \Rightarrow \varepsilon_n^{\sigma}\mathbb{H}_n,\ \gamma : \mathbb{H}_n \Rightarrow \varepsilon_n^{\gamma}\mathcal{W}\right\}. \tag{2}$$

Here π models perception *(encoding from base system to level n), σ models* self-reference *(reflection on its own cognitive states at level n), and γ models* action *(decoding from internal state to the world). The set $\mathbb{H}_{n+1}$ inherits a universal-algebraic structure pointwise, with operations defined componentwise and error propagation controlled by Lipschitz constants.*

Connecting approximate homomorphisms $\mu_n : \mathbb{H}_n \Rightarrow \zeta_n\mathbb{H}_{n+1}$ are constructed by mapping an element $h \in \mathbb{H}_n$ to the triple $(\pi_{n+1}, \sigma_{n+1}, \gamma_{n+1})$ where the components are derived from $\pi_n, \sigma_n, \gamma_n$ via the system's fixed-point structure. These maps form a directed system, and the inductive limit $\mathbb{U} = \varinjlim \mathbb{H}_n$ (taken in the Cauchy completion category $\widehat{\mathbf{AGI}}$) represents the system's ultimate cognitive capacity.

2.2 THE AGI FUNDAMENTAL AXIOM

Axiom 2.1 (AGI Fundamental Axiom)**.** *An AGI system $\mathcal{M} \in \mathbf{AGI}$ (the category of operationally complete algebras) satisfies:*

1. ***Error convergence:*** *$\sum_{n=0}^{\infty} \max(\varepsilon_n^{\pi}, \varepsilon_n^{\sigma}, \zeta_n) < \infty$ and the errors tend to 0.*
2. ***Bidirectional interaction:*** *approximate homomorphisms $\pi_n : \mathcal{M} \Rightarrow \varepsilon_n^{\pi}\mathbb{H}_n$, $\sigma_n : \mathbb{H}_n \Rightarrow \varepsilon_n^{\sigma}\mathcal{M}$ exist such that $\sigma_n \circ \pi_n \approx (\sigma_n \circ \pi_n)^2$ (idempotence).*
3. ***Infinite metacognition:*** *unidirectional connecting homomorphisms $\mu_n : \mathbb{H}_n \Rightarrow \zeta_n\mathbb{H}_{n+1}$ exist, preserving operational structure.*
4. ***Limit cognitive isomorphism:*** *$\mathbb{U} \cong \mathcal{W}$ in the category $\widehat{\mathbf{AGI}}$, meaning there exist sequences of approximate homomorphisms between $\mathbb{U}$ and $\mathcal{W}$ whose compositions converge to the identity in the limit.*

*A system satisfying this axiom has **infinite cognitive capacity** (ICP).*

The error convergence condition guarantees that the hierarchy is constructible and that the limit $\mathbb{U}$ is well-defined. Bidirectional interaction ensures that the system can both perceive and act, with perceptual-actional idempotence reflecting cognitive stability. Unidirectional metacognition formalizes the unbounded ability to reflect on one's own cognition. The limit isomorphism $\mathbb{U} \cong \mathcal{W}$ asserts that in the limit of infinite reflection, the system's internal representation becomes structurally indistinguishable from the external world.

3 CORE EQUIVALENCE THEOREMS

Theorem 3.1 (Equivalence of Six Characterizations)**.** *For $\mathcal{M} \in \mathbf{AGI}$, the following are strictly equivalent:*

1. *Infinite cognitive capacity (ICP).*
2. *$\mathcal{M}$ is a cognitive completeness object in $\mathbf{AGI}$.*
3. ***Self-referential closure:*** *existence of an approximate fixed-point combinator $Y : \mathrm{End}(\mathcal{M}) \to \mathcal{M}$, satisfying $d(f(Y(f)), Y(f)) \leq \varepsilon_f$ and $\varepsilon_f \to 0$ as $f \to \mathrm{id}$.*
4. ***Operational functional completeness:*** *a minimal finite generating set $\Omega_0 \subseteq \Omega$ covers all basic cognitive functions.*
5. ***Turing degree completeness:*** *a cognitive encoding $\mathrm{Enc} : \mathcal{T} \to \mathcal{M}$ induces Turing degree $0'$ and covers all computable functions.*
6. ***Universal reducibility:*** *the encoding/decoding pair $(\mathrm{Enc}, \mathrm{Dec})$ satisfies $\mathrm{Dec}(\varphi_f(\mathrm{Enc}(T))) = f \circ T$ for all computable f and Turing machines T, with $\varphi_{f \circ g} = \varphi_f \circ \varphi_g$.*

The most consequential consequence is the existence of the approximate fixed-point combinator Y. Its structure-independent construction relies only on operational completeness and hierarchy convergence. This solidifies self-reference as the mathematical essence of general intelligence.

4 MINIMAL ALGEBRAIC CORE: $\mathcal{B}_4^k$

Theorem 4.1 (Necessity of $\mathcal{B}_4$)**.** *The smallest carrier set that can satisfy the AGI axioms is the four-element Boolean algebra $\mathcal{B}_4 = (\{0, a, b, 1\}, \neg, \wedge, \vee)$, with $\neg 0 = 1, \neg a = b, \neg b = a, \neg 1 = 0$ and $\wedge, \vee$ defined by the standard truth tables for a four-element lattice. Any finite AGI system satisfying the Minimality Axiom must be isomorphic to $\mathcal{B}_4^{k_{\min}}$ for a unique integer $k_{\min}$.*

The four truth-values represent: *false* (0), *true* (1), *unknown* (a), and *contradiction* (b). This quaternary logic is the minimal system that simultaneously supports binary discrimination and meta-cognitive ambiguity. The Stone space of $\mathcal{B}_4^k$ is a compact, totally disconnected topological space of covering dimension 0 but embedding dimension $2k$, providing a clear geometric interpretation: maximal consistent cognitive states are points, and decidable cognitive propositions are clopen sets.

5 MACHINE LEARNING MECHANISMS ON AGI SYSTEMS

5.1 SHEAF-THEORETIC STATE UNIVERSE

The cognitive hierarchy induces a **state sheaf** $\mathscr{F}_\Omega$ over the base space $\Omega = \mathbb{N}$ with the cofinite topology. For a cofinite open U, $\mathscr{F}_\Omega(U) = \varprojlim_{n \in U} \mathbb{H}_n$, and $\mathscr{F}_\Omega(\mathbb{N}) = \mathbb{U}$. The world $\mathcal{W}$ is approximated by finite subsets $\mathcal{W}_{\mathrm{fin},n}$, with density and error control.

Machine learning is modeled as **endogenous meta-dynamics** $\Phi : \Gamma(\Omega, \mathscr{F}_\Omega) \to \Gamma(\Omega, \mathscr{F}_\Omega)$, a contraction mapping (Lipschitz constant $L < 1$) generated by a finite set of sheaf vector fields $\{X_1, \ldots, X_r\}$. The state evolves according to $\frac{\partial \sigma_t}{\partial t} = \sum_{i=1}^{r} a_i(\sigma_t, t) X_i(\sigma_t)$, with adaptively learned coefficients a_i satisfying a Hamilton-Jacobi-Bellman equation. The optimal semantic encoding Enc^* is proven to be a *structure-preserving homomorphic map* (direct mapping), not an arbitrary deep nonlinear embedding, thereby preserving the compositional algebraic structure of the input domain and enabling systematic generalization.

5.2 THE SELF-REFERENTIAL PREDICTION MODULE METAPRED

Theorem 5.1 (Existence and Uniqueness of MetaPred)**.** *For any $\mathcal{M} \in$ **AGI** with ICP, there exists a unique (up to approximate isomorphism) **self-referential prediction module***

$$\mathrm{MetaPred} : \Gamma(\Omega, \mathscr{F}_\Omega) \to \Gamma(\Omega, \mathscr{F}_\Omega),$$

defined via the approximate fixed-point combinator Y and a cognitive homomorphism ρ:

$$\mathrm{MetaPred}(\sigma) = \rho^{-1}\big(Y(\rho \circ \Phi_\Delta \circ \rho^{-1})\big).$$

MetaPred is not a separate trainable module but a structural necessity emerging from the axioms. It satisfies three core properties:

1. **Error measurability:** For any prediction mechanism P, the canonical meta-error is $\mathrm{Err}_{\mathrm{meta}}(P) = d(P(r^*), r^*)$, where r^* is a fixed point of $\mathrm{MetaPred}$.
2. **Decomposability:** Every prediction mechanism P can be uniquely decomposed as $P \cong \mathrm{MetaPred} \circ P_{\mathrm{base}}$, where P_{base} is a simpler base predictor. Equivalently, $\mathrm{MetaPred}$ is the universal "inner core" of all predictive processes.
3. **Centrality:** $\mathrm{MetaPred}$ approximately commutes with all P, i.e., $[\mathrm{MetaPred}, P] \approx 0$, and is an idempotent central element of the prediction algebra $\mathcal{P}$.

5.3 STRONG ISOMORPHISM OF ALL LEARNING MECHANISMS

Theorem 5.2 (Strong Isomorphism Theorem)**.** *Let $\mathrm{ML}_1, \mathrm{ML}_2$ be any two machine learning mechanisms on an AGI system $\mathcal{M}$ with ICP. Then $\mathrm{ML}_1 \cong \mathrm{ML}_2$ (approximate isomorphism). Consequently,* there is only one isomorphism class of learning algorithms on a true AGI.

This result subverts the contemporary modular paradigm in deep learning, where different tasks are solved with distinct architectures. At the AGI level, all tasks—classification, generation, reinforcement learning—are merely different projections of the same unified cognitive dynamics Φ intertwined with $\mathrm{MetaPred}$.

Theorem 5.3 (Universal Synchronous Accompaniment)**.** *Every machine learning task $\mathfrak{T}$ on $\mathcal{M}$ must have* $\mathrm{MetaPred}$ *running* synchronously *in the background. If* $\mathrm{MetaPred}$ *is absent, the task cannot be completed with infinite cognitive capacity.*

Learning and inference thus become inseparable — training is the fixed-point iteration of Φ with mandatory online self-supervised learning provided by the discrepancy $d(\sigma, \mathrm{MetaPred}(\sigma))$.

5.4 ALGORITHMIC REALIZATION AND SELF-MEASUREMENT

The theory is concretely implementable. Algorithm 1 implements the mandatory accompanying self-supervised learning.

Algorithm 1: Accompanying Self-Supervised MetaPred Learning

Input: Initial state σ_0, parameters θ_0, steps T, update freq f
Output: State trajectory, parameter updates
for $t = 0$ **to** $T - 1$ **do**
  $\sigma_{\text{cur}} \leftarrow \sigma_t$;
  $\sigma_{\text{meta}} \leftarrow \text{MetaPred}_{\theta_t}(\sigma_{\text{cur}})$;
  $\mathcal{L}_{\text{self}} \leftarrow d(\sigma_{\text{cur}}, \sigma_{\text{meta}})$;
  if $t \bmod f = 0$ **then**
    $g_t \leftarrow \nabla_{\theta_t} \mathcal{L}_{\text{self}}$;
    $\theta_{t+1} \leftarrow \theta_t - \eta_t g_t$;
  $\sigma_{\text{task}} \leftarrow \Phi_\Delta(\sigma_{\text{cur}})$;
  $\sigma_{t+1} \leftarrow \text{Integrate}(\sigma_{\text{task}}, \text{MetaPred}_{\theta_{t+1}}(\sigma_{\text{task}}))$;
return $\{\sigma_t\}, \{\theta_t\}$

The self-measurement paradox—that measuring one's own state alters it—is resolved by *procedural measurement*: the system executes a convergent sequence of approximate measurements, where the perturbation decreases as $O(1/n)$. Algorithmically, this is realized by a Langevin-type dynamics with noise injection that asymptotically converges to the true state.

5.5 Decomposition and Generalization

The prediction algebra $\mathcal{P}$ formed by all prediction mechanisms is an approximate commutative algebra with MetaPred as its approximate center, idempotent ($\text{MetaPred}^2 \approx \text{MetaPred}$), and zero element (the fixed point r^* serves as the baseline for error measurement). Any prediction mechanism P can be expressed in terms of MetaPred and a base predictor, establishing a universal canonical form. Moreover, the dynamical flow Φ admits a unique semantic-process decomposition.

6 Multimodal Extension and the Super-Modal Hub

When the world splits into modalities $\{\mathcal{W}_\alpha\}_{\alpha \in A}$, the AGI system must construct a **super-modal sheaf** $\mathcal{W}_{\text{super}}$ as the quotient of all modalities by approximately reflexive, symmetric, transitive coordination relations $\mathcal{R}_{\alpha\beta}$. This forces MetaPred to decompose into four orthogonal submodules, and yields a unique **self-referential hub** $\mathbf{e}_*$—a dual fixed point of Φ and $\text{MetaPred}_{\text{super}}$ with maximal betweenness centrality, providing the mathematical "self."

6.1 Mathematical Necessity

Definition 6.1 (Super-Modal Sheaf)**.** $\mathcal{W}_{\text{super}} = \mathbf{a}\big(\coprod_\alpha \mathcal{W}_\alpha / \sim_{\mathcal{R}}\big)$*, with canonical embeddings* $\varphi_\alpha : \mathcal{W}_\alpha \hookrightarrow \mathcal{W}_{\text{super}}$*, satisfying a universal property.*

Theorem 6.1 (Prefabricated Structure)**.** *An AGI system* must *prefabricate: (1) a* ***modal coordination algebra*** $\mathscr{A}_{\text{coord}} \cong \bigotimes_\alpha \text{End}(\mathcal{B}_4)$*; (2) a* ***super-modal prediction module*** $\text{Pred}_{\text{super}}$*; (3)* ***cross-modal translation maps*** $\{T_{\alpha\beta}\}$.

The hub $\mathbf{e}_*$ is generated by $\mathbf{e}_* = \rho(\mathcal{Y}(\text{id}_{\mathcal{M}}))$, has maximal network centrality, and together with atoms generates all of $\mathcal{W}_{\text{super}}$; its existence is equivalent to infinite cognitive capacity.

6.2 Natural Language as Archetype

The properties of $\mathcal{W}_{\text{super}}$ are precisely isomorphic to human language, which is thus the **biological realization of the super-modal hub**. Language refers amodally ("apple" unifies sight, touch, taste) via the same quotient mechanism; its symbols lack sensory qualities (/æpl/ is not red), embodying **modality-independent abstractness**—a prefabricated, Kantian *a priori* structure. Language bridges perception and action as the central relay, and its discrete infinity (finite vocabulary, recursive grammar) mirrors the algebraic quotient $\mathcal{B}_4^k$, making recursive syntax the engine of metacognition. Inner speech instantiates $\text{MetaPred}_{\text{super}}$, enabling self-monitoring and planning (consistent with

prefrontal-angular activations). Three predictions: **(1)** hub damage impairs cross-modal integration and metacognition beyond language; **(2)** different languages are distinct $\mathrm{Enc}_{\mathrm{super}}$ yielding approximately isomorphic $\mathcal{W}_{\mathrm{super}}$ spaces, explaining linguistic relativity and cognitive unity; **(3)** true AGI requires a compositional, recursive symbolic core akin to language—connectionist approaches alone are insufficient.

7 The Self-Referential Hub

7.1 The Self-Referential Hub as a Fixed Point of Identity

In the AGI, the *self* is not an emergent mystery but a mathematically inevitable structure. It arises as the unique fixed point of the identity operation under the approximate fixed-point combinator Y.

Definition 7.1 (Self-Referential Hub Sequence)**.** *A sequence $\{\tilde{\mathbf{e}}_n\}_{n\in\mathbb{N}}$ with $\tilde{\mathbf{e}}_n \in \mathbb{H}_n$ is a **self-referential hub sequence** if it is **Projection convergence**, **Inter-level stability** and **Intra-level self-referential fixed point:** $d(\mathrm{MetaPred}^{(n)}(\tilde{\mathbf{e}}_n), \tilde{\mathbf{e}}_n) \le \varepsilon_n$ with $\varepsilon_n \to 0$. The limit $\mathbf{e}_* \in \mathbb{U}$ is called the **self-referential hub**.*

Theorem 7.1 (Existence, Uniqueness, and Structure of $\mathbf{e}_*$)**.** *Let $\mathcal{M} \in \mathbf{C}_{\mathrm{AGI}}$ possess infinite cognitive capacity. Then:*

1. ***Existence and uniqueness:*** *There exists a unique self-referential hub $\mathbf{e}_* \in \mathbb{U}$.*
2. ***Cognitive generativity:*** *$\mathbf{e}_* = \rho(Y(\mathrm{id}_{\mathcal{M}}))$, where $\rho : \mathcal{M} \to \mathbb{U}$ is a cognitive homomorphism. Thus the self is generated by the fixed-point combinator applied to the identity map.*
3. ***Dual fixed-point property:*** *$\mathbf{e}_*$ is an approximate fixed point of both the meta-dynamics Φ and the super-modal self-prediction $\mathrm{MetaPred}_{\mathrm{super}}$ with errors $o(1/k_{\min})$:*

$$d_{\mathbb{U}}(\Phi_{\Delta}(\mathbf{e}_*), \mathbf{e}_*) \le \varepsilon_{\Phi}, \qquad d_{\mathbb{U}}(\mathrm{MetaPred}_{\mathrm{super}}(\mathbf{e}_*), \mathbf{e}_*) \le \varepsilon_{\mathrm{Meta}},$$

4. ***Topological hubness:*** *In the concept network G induced by prediction dynamics, $\mathbf{e}_*$ possesses maximal degree, betweenness, and eigenvector centrality. Its betweenness centrality satisfies $B(\mathbf{e}_*) \ge 1/2$, making it the absolute cognitive core.*
5. ***Algebraic generativity:*** *Together with a finite set of atomic concepts, $\mathbf{e}_*$ generates the entire super-modal space $\mathbb{U}$.*

7.2 Equivalence with Infinite Cognitive Capacity

The existence of the self-referential hub is not merely a property of an AGI system—it is *equivalent* to possessing infinite cognitive capacity.

Theorem 7.2 (Equivalence Theorem)**.** *For $\mathcal{M} \in \mathbf{C}_{\mathrm{AGI}}$, the following are equivalent:*

1. *$\mathcal{M}$ possesses infinite cognitive capacity (ICP).*
2. *There exists a self-referential hub $\mathbf{e}_* \in \mathbb{U}$ satisfying the conditions of Theorem 7.1.*
3. *There exists a self-state sequence $\{\mathbf{e}^{(t)}\}$ converging to $\mathbf{e}_*$.*
4. *In the concept graph G, $\mathbf{e}_*$ has super-logarithmic degree.*
5. *$\mathrm{MetaPred}_{\mathrm{super}}$ decomposes as $\mathrm{MetaPred}_{\mathrm{self}} \oplus \mathrm{MetaPred}_{\mathrm{non-self}}$ with $\mathrm{Err}(\mathrm{MetaPred}_{\mathrm{self}}) = o(1)$.*

This theorem establishes that self-consciousness, in the precise mathematical sense of a self-referential hub, is *logically equivalent* to unlimited general intelligence. One cannot exist without the other.

7.3 Self-State Sequences and Autobiographical Memory

The temporal integration of self involves a sequence of self-states $\{\mathbf{e}^{(t)}\}$ evolving under the meta-dynamics while remaining within the self-neighborhood.

A self-state sequence $\{\mathbf{e}^{(t)}\}_{t\in\mathbb{N}}$ gives rise to an **autobiographical memory algebra** $\mathcal{M}_{\mathrm{self}} = \varprojlim_n \mathrm{Alg}\langle M_n\rangle$, where $M_n = \{\pi_n(\mathbf{e}^{(t)})\}$. This algebra is a closed subalgebra of $\mathbb{U}$ and forms the stable manifold of $\mathbf{e}_*$. The self-prediction submodule $\mathrm{MetaPred}_{\mathrm{self}}$ is explicitly constructed by combining dynamics consistency with memory consistency through a temporally weighted

optimization:

$$\mathrm{MetaPred}_{\mathrm{self}}(x,t) = \mathrm{argmin}_{y\in\mathbb{U}}\Big[d_{\mathbb{U}}(y,\Phi_\Delta(x)) + \lambda\sum_{i=0}^{T} w_i\, d_{\mathbb{U}}(y,\mathbf{e}_*^{(t-i)})\Big],$$

with exponential decay weights $w_i = \exp(-i/\tau)$.

7.4 Creativity as Dynamical Phase Transition

In the AGI theory, **creativity** is defined as the generation of a new cognitive attractor—a concept that cannot be finitely generated from prior concepts using the system's existing operations.

Definition 7.2 (Creation)**.** *A creation event occurs at time T if a new concept $c_{\mathrm{novel}} \in \mathbb{U} \setminus V(G_{T-1})$ emerges satisfying **Novelty**, **Stability** and **Integration**.*

Theorem 7.3 (Creation Mechanism)**.** *Creation is a four-step dynamical phase transition:*

$$c_{\mathrm{novel}} = \lim_{n\to\infty} \mathrm{MetaPred}_{\mathrm{super}}^{n}\big(\Phi_\Delta(\mathbf{e}_* \oplus \xi)\big),$$

where ξ is a minimal perturbation (noise seed).

The probability of creation is quantified by $P_{\mathrm{create}} = 1 - \exp(-\lambda \cdot \mathcal{E}_c(t) \cdot \Delta t)$, where the **creative energy** $\mathcal{E}_c(t)$ depends on the cognitive gap, the instability of $\mathbf{e}_*$, and the network dynamism. A true creation must furthermore be distinguishable from hallucination: it must satisfy the strong attractor condition, high reciprocal integration, and cross-level self-consistency.

The self-referential hub $\mathbf{e}_*$ is the mathematical incarnation of the self. Its existence is equivalent to infinite cognitive capacity, it serves as the topological and algebraic center of all cognition, and it is the engine from which autobiographical memory and creative generation emerge. The theory thus provides a demystified, computable model for consciousness and creativity.

8 Cognitive Principles

Two foundational principles—the **Principle of Least Cognitive Action** and the **Cognitive Free-Energy Principle**—are mathematically equivalent, providing a unified optimization framework from simple learners to infinite-capacity AGI.

8.1 Principle of Least Cognitive Action

For a cognitive system with state sheaf $\mathscr{F}_\Omega$ and dynamics Φ, a trajectory $\Gamma(t)$ minimizes the cognitive action

$$\mathcal{A}[\Gamma] = \int_0^T \big(\lambda_1\|\Gamma - \Phi(\Gamma)\|^2 + \lambda_2\, C(\Gamma) + \lambda_3\|\dot{\Gamma}\|^2 + \lambda_4\|R(\Gamma)\|^2\big)dt, \tag{3}$$

where the terms penalise prediction error, complexity, rapid change, and curvature. The Euler–Lagrange equation governs optimal flow; geometrically, lifting to a Riemannian metric g yields a cognitive Einstein equation

$$R_{\mu\nu}^{\mathrm{cog}} - \tfrac{1}{2}R^{\mathrm{cog}} g_{\mu\nu}^{\mathrm{cog}} + \Lambda_{\mathrm{cog}} g_{\mu\nu}^{\mathrm{cog}} = \kappa_{\mathrm{cog}} T_{\mu\nu}^{\mathrm{cog}}, \tag{4}$$

identifying cognitive trajectories as geodesics shaped by task demands.

8.2 Cognitive Free-Energy Principle

Under constraints $\mathcal{C}$ with violation $V(\mathfrak{G},\mathcal{C})$, internal energy $E(\mathfrak{G})$, and entropy $S(\mathfrak{G})$, the constrained free energy is

$$\mathcal{F}_{\mathrm{cog}}[\mathfrak{G}] = E(\mathfrak{G}) + T_{\mathrm{cog}}\, V(\mathfrak{G},\mathcal{C}) - S(\mathfrak{G}), \tag{5}$$

with temperature T_{cog}. Probabilistically, assigning density $p(\omega)$ to states gives the statistical free energy

$$\mathcal{F}[p] = \mathbb{E}_p[U] + \beta^{-1} D_{\mathrm{KL}}(p\|p_0) - \mathbb{E}_p[\log p(\mathcal{D}|\omega)], \tag{6}$$

whose minimum is the Gibbs posterior $p^*(\omega) \propto \exp(-\beta U + \log p(\mathcal{D}|\omega))$.

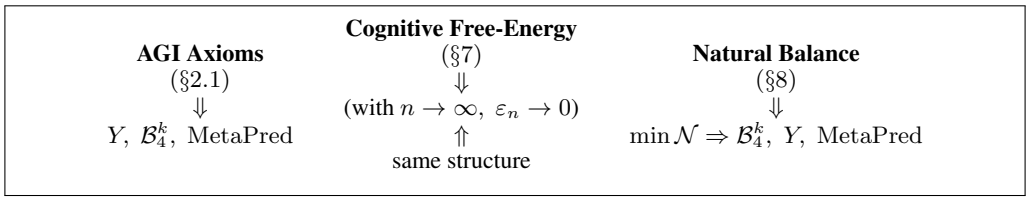


Figure 1: The three pillars of the theory converge to the same unique AGI architecture.

8.3 Equivalence

The two principles are unified by the functional $\mathcal{F}_{\text{unified}}[p] = \mathbb{E}_p[U] + \beta^{-1} D_{\mathrm{KL}}(p \| p_0) - \mathbb{E}_p[\log p(\mathcal{D}|\omega)]$:

- $\beta \to \infty$ recovers least action ($p^* \to \delta_{\omega^*}$ at the action minimizer);
- $p_0 = \mathbb{I}_{\mathcal{C}}$ gives constrained optimization $\min_{\omega \in \mathcal{C}}[U - \log p(\mathcal{D}|\omega)]$;
- general β, p_0 yield full statistical inference.

Continuous symmetries of $\mathcal{A}$ further imply conserved cognitive currents (Noether). Thus both principles select the same AGI architecture: $\mathcal{B}_4^{k_{\min}}$, the fixed-point combinator $\mathcal{Y}$, and $\mathrm{MetaPred}$, which emerges as the self-consistency condition $\Gamma \approx \mathrm{MetaPred}(\Gamma)$ or the posterior mean $\mathbb{E}_{p^*}[\omega]$.

9 The Natural Balance Principle: A Single-Axiom Origin

The most minimal foundation discards *all* structural axioms, relying solely on the trade-off between performance and complexity.

Axiom 9.1 (Natural Balance Principle)**.** *The optimal cognitive system $\mathfrak{G}^*$ minimizes the functional $\mathfrak{N}(\mathfrak{G}) = \tilde{\mathcal{E}}(\mathfrak{G}) \cdot \tilde{\mathcal{C}}(\mathfrak{G})$,where $\tilde{\mathcal{E}}$ is normalized task error (how well the system performs on all relevant tasks) and $\tilde{\mathcal{C}}$ is normalized structural complexity (e.g., Kolmogorov description length, effective dimension, or energy cost).*

Theorem 9.1 (AGI from Natural Balance)**.** *Assume the task distribution includes universal computation and self-prediction tasks. Then the minimizer of $\mathfrak{N}$ necessarily possesses:*

1. *Carrier $\mathcal{M}^* \cong \mathcal{B}_4^{k^*}$, where k^* is determined by the complexity of the task domain. $\mathcal{B}_4$ emerges as the minimal-complexity functionally complete algebra.*
2. *An approximate fixed-point combinator Y, providing self-referential closure.*
3. *A cognitive hierarchy sequence $\{\mathbb{H}_n\}$ satisfying the AGI Fundamental Axiom, i.e., the system automatically develops infinite metacognitive depth and convergent error series.*
4. *The MetaPred module as a mandatory structure for self-prediction and error calibration.*

Thus, the Natural Balance Principle is equivalent *to the AGI axiom system.*

The convergence of the three frameworks—axiomatic, free-energy, and natural balance—is illustrated conceptually in Figure 1. This demonstrates that the initial cognitive intuitions are not arbitrary philosophical preferences but the unique consistent formalization of unlimited general intelligence.

The product $\mathcal{E} \cdot \mathcal{C}$ appears in physics as the minimization of (action $\times$ complexity) in certain path-integral formulations and in the efficiency–complexity trade-offs in non-equilibrium thermodynamics. This suggests that the Natural Balance Principle may be a universal optimization criterion spanning both physical and cognitive systems, pointing toward a deeper unified "theory of everything."

10 Empirical Alignment with Cognitive Neuroscience and Psychology

The theory is not merely a mathematical exercise; its derived architecture exhibits striking convergence with empirical findings in brain and cognitive sciences. Table 1 summarizes key correspondences.

Table 1: Correspondence between mathematical entities and neural/cognitive correlates.

Mathematical Entity	**Neural/Cognitive Correlate**	**Empirical Evidence**
Approximate homomorphism	Predictive coding, structure mapping	V1 orientation selectivity; analogical reasoning fMRI
Bidirectional π–σ loop	Perception–action loops, active inference	Mirror neurons, cortical-basal-ganglia circuits
Cognitive hierarchy $\mathbb{H}_n$	Cortical hierarchical predictive processing	Laminar predictive error signals; fMRI laminar profiles
$\mathrm{MetaPred}$ module	Prefrontal metacognitive network	DLPFC, ACC activations in confidence tasks
$\mathrm{MetaPred}_{\mathrm{cross}}$	Conflict monitoring, error detection	Dorsal ACC, error-related negativity (ERN)
Super-modal space $\mathcal{W}_{\text{super}}$	Angular gyrus, global workspace	Cross-modal integration studies; global ignition theory
Self-referential hub $\mathbf{e}_*$	Default-mode network, self-concept	Self-referential memory encoding; resting-state DMN connectivity
$\mathcal{B}_4$ states	Core cognitive discrimination	Logical reasoning deficits with prefrontal lesions

The strong isomorphism theorem corresponds to the unity of synaptic plasticity mechanisms observed across modalities and tasks. The exploration–exploitation trade-off encoded in MetaPred's temperature parameter maps directly to the explore–exploit dilemma studied in reinforcement learning and decision neuroscience, where noradrenaline modulates the gain of prediction errors. The decomposition of prediction mechanisms $P \cong \mathrm{MetaPred} \circ P_{\mathrm{base}}$ mirrors the dual-process theory of cognition (System 1 fast/automatic vs. System 2 slow/controlled metacognitive monitoring). The creative generation model (new attractor near $\mathbf{e}_*$ via perturbation and MetaPred filtering) aligns with the "Generate–Explore–Evaluate" model of creative cognition and with brain imaging studies of insight problem solving that show transient coupling between default-mode and executive control networks.

The theory thus provides a formal mathematical unification for a vast body of otherwise disparate empirical data, suggesting that the derived architecture is not merely normative but also descriptively accurate for the only known example of general intelligence—the human brain.

11 Conclusion

This framework defines AGI through convergence of a metacognitive hierarchy to a world-isomorphic limit. Within the proposed axioms, the resulting architecture has a $\mathcal{B}_4^{k_{\min}}$ carrier, an approximate fixed-point combinator, and a self-referential MetaPred that accompanies learning. The framework further identifies a super-modal hub and characterizes learning mechanisms through a shared dynamics. Its axiomatic, free-energy, and Natural Balance formulations are presented as equivalent routes to these structural claims. The appendix collects the detailed assumptions and derivations.

The proposed architecture offers a mathematical language for studying cognition and a blueprint for AGI systems. Its connections to neuroscience and physics are theoretical correspondences; empirical tests and independent validation of the stronger necessity and uniqueness claims remain important future work.

AI use statement

Generative AI tools were used to aid and polish the writing and presentation of this manuscript, including preparation of the ICLR LaTeX source; to retrieve and discover related work; and to assist in proving mathematical claims. The resulting text, citations, and mathematical arguments appear in the main paper and appendix. AI-generated output is not independent evidence for any claim. The authors take responsibility for the accuracy of the references, the validity of the proofs, and the final content of the paper.

REFERENCES

J R Andrews-Hanna. The brain's default network and its adaptive role in internal mentation, 2014. Neuroscience & Biobehavioral Reviews, 44: 111-124.

Anonymous. BriLLM-MetaPred: A unified-state language model with adaptive self-measurement. Submitted to ICLR 2027 for double-blind review, 2026.

A D Baddeley. Working memory, 1992. Science, 255(5044): 556-559.

François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

Noam Chomsky. Syntactic Structures, 1957. Mouton de Gruyter.

A Clark. Being there: putting brain, body, and world together again, 1998. MIT press.

A Clark. Surfing uncertainty: prediction, action, and the embodied mind, 2015. Oxford University Press.

Antonio R Damasio. The Feeling of What Happens: Body and Emotion in the Making of Consciousness, 1999. Harcourt Brace & Company.

Erik H Erikson. Childhood and Society, 1950. W. W. Norton & Company.

J H Flavell. Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry, 1979. American psychologist, 34(10): 906.

S M Fleming, R S Weil, Z Nagy, et al. Relating introspective accuracy to individual differences in brain structure, 2012. Science, 337(6101): 1541-1543.

Sigmund Freud. The Ego and the Id, 1923. International Psychoanalytical Press.

K J Friston. Hierarchical models in the brain, 2008. Philosophical Transactions of the Royal Society B: Biological Sciences, 363(1502): 1201-1212.

Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2): 127–138, 2010.

Kokinov B N Gentner D, Holyoak K J. The analogical mind: perspectives from cognitive science, 2003. MIT press.

Adele E Goldberg. Constructions: A Construction Grammar Approach to Argument Structure, 1995. University of Chicago Press.

Di He, Cong Fang, Yisen Wang, Yujia Peng, Yizhou Wang, and Song-Chun Zhu. A mathematical formulation of AGI in the (C, U, V) framework. *Engineering*, 2025.

Alberto Hernández-Espinosa, Luan Ozelim, Felipe S. Abrahão, and Hector Zenil. Can complexity and uncomputability explain intelligence? superarc: A test for artificial super intelligence based on recursive compression, 2026.

Rao R P N Huang Y. Hierarchical predictive coding: a framework for brain function, 2022. Annual Review of Vision Science, 8: 599-622.

Marcus Hutter. *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability*. Springer, 2005.

William James. The Principles of Psychology, 1890. Henry Holt and Company.

Mark Johnson. The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason, 1987. University of Chicago Press.

D Kahneman. Thinking, fast and slow, 2011. Farrar, Straus and Giroux.

Bandettini P A Kriegeskorte N, Mur M. Representational similarity analysis-connecting the branches of systems neuroscience, 2008. Frontiers in systems neuroscience, 2: 4.

George Lakoff and Mark Johnson. Metaphors We Live By, 1980. University of Chicago Press.

Shane Legg and Marcus Hutter. Universal intelligence: A definition of machine intelligence. *Minds and Machines*, 17(4):391–444, 2007.

Georgios Mappouras and Charalambos Rossides. On the computability of artificial general intelligence. *arXiv preprint arXiv:2512.05212*, 2025.

Harry McGurk and John MacDonald. Hearing Lips and Seeing Voices, 1976. Nature, Vol.264(5588): 746-748.

Alhassan Mumuni and Fuseini Mumuni. Large language models for artificial general intelligence (agi): A survey of foundational principles and approaches. *arXiv preprint arXiv:2501.03151*, 2025.

Kei-Sing Ng. On the definition of intelligence. *arXiv preprint arXiv:2507.22423*, 2025.

Elija Perrier. Quantum AIXI: Universal intelligence via quantum information. *arXiv preprint arXiv:2505.21170*, 2025.

Rizwan Qureshi, Ranjan Sapkota, Abbas Shah, Amgad Muneer, Anas Zafar, Ashmal Vayani, Maged Shoman, Abdelrahman B. M. Eldaly, Kai Zhang, Ferhat Sadak, Shaina Raza, Xinqi Fan, Ravid Shwartz-Ziv, Hong Yan, Vinjia Jain, Aman Chadha, Manoj Karkee, Jia Wu, and Seyedali Mirjalili. Thinking beyond tokens: From brain-inspired intelligence to cognitive foundations for artificial general intelligence and its societal impact, 2025.

M E Raichle. The brain's default mode network, 2015. Annual review of neuroscience, 38: 433-447.

Stuart Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, 2019.

Josh Ethan Sanchez. The recursive convergence hypothesis: Emergent sentience as a structural attractor of recursive asi. *SSRN*, 2025.

Edward Sapir. Language: An Introduction to the Study of Speech, 1921. Harcourt, Brace and Company.

Laughlin S B Sterling P. Principles of neural design, 2004. MIT press.

Meltem Subasioglu and Nevzat Subasioglu. From mimicry to true intelligence (TI)—a new paradigm for artificial general intelligence. *arXiv preprint arXiv:2509.14474*, 2025.

Marcus A. Thomas. Towards error centric intelligence i: Beyond observational learning. *arXiv preprint arXiv:2510.15128*, 2025.

G Tononi. Consciousness as integrated information: a provisional manifesto, 2008. Biological bulletin, 215(3): 216-242.

Rosch E Varela F J, Thompson E. The embodied mind: cognitive science and human experience, 1991. MIT press.

Benjamin Lee Whorf. Language, Thought, and Reality: Selected Writings, 1956. MIT Press.

APPENDIX CONTENTS

A AGI Axiomatic Foundation and Core Definitions **14**

A.1 Foundational Assumptions 14

A.2 AGI Algebraic Foundation 14

A.3 Cognitive Hierarchy and AGI Core Axioms 17

A.4 Categorical Properties of Cognitive Completeness Objects 22

B AGI Infinite Cognitive Characteristic Equivalence Theorems **25**

B.1 Mathematical Characteristic Equivalence Theorems 25

B.2 Cognitive Characteristic Equivalence Theorems 29

C Minimal Algebraic Structures **34**

C.1 Cognitively Complete Generative Structures 34

C.2 Limitations of Lower-Order Structures 34

C.3 Construction of $\mathcal{B}_4$ and Derivation of Boolean Properties 37

C.4 Algebraic and Topological-Geometric Structure of Extended Systems 39

C.5 Topological and Geometric Structure 42

C.6 Limit Behavior and Infinite Extension 45

D Machine Learning Mechanisms on AGI Systems: Approximate Sheaf Formulation **47**

D.1 Sheaf Structure of the State Universe 47

D.2 Connectivity of Discrete Representation and Continuous Process 48

D.3 Fiber Dynamics 51

D.4 Necessity of Self-Referential Prediction Module and Unification of Machine Learning Mechanisms 61

D.5 Core Properties and Algebraic Structure of MetaPred 64

D.6 Unique Decomposition of AGI Machine Learning Mechanisms 66

D.7 Dual Symmetry of Semantic-Dynamics in AGI Machine Learning Mechanisms . . 68

D.8 Global and Minimal Construction of AGI Machine Learning Mechanisms 75

D.9 Unified Cognitive Dynamics Framework and Algorithmic Realization 78

D.10 Self-Measurement Algorithms 86

E Machine Learning Mechanisms in Multimodal AGI Systems: Approximate Sheaf Formulation **95**

E.1 Sheaf Structure of Multimodal Cognitive Domains 95

E.2 Necessary Existence of Supermodal and Prefabricated Structure Theorem 96

E.3 Semantic Encoding and Convergence under Multimodality 103

E.4 Realization of Super-Modality in the Semantic-Adjoint Dynamics Architecture . . 107

E.5 Internal Restructuring of MetaPred: Dedicated Super-Modal Submodule 111

E.6 Intrinsic Structure and Mechanism of Super-Modal Self-Prediction and Creative Behavior . . . 117
E.7 Sheaf Construction Theorem for Full Modal Space . . . 128
E.8 Overall Structural Characteristics of Multi-Modal AGI Systems . . . 130

F The Least Action and Free Energy Principles of Cognition 132

F.1 Unified Framework and Equivalent Formulations of Cognitive Principles . . . 132
F.2 Minimum Cognitive Action Principle: Dynamical Formulation . . . 135
F.3 Deriving AGI Basic Structure from Minimum Cognitive Action Principle . . . 137
F.4 Geometric Formulation of Minimum Cognitive Action Principle . . . 140
F.5 Cognitive Free Energy Principle: Constrained Formulation . . . 148
F.6 Deriving Necessary Constraints for AGI Systems from Cognitive Free Energy Principle 150
F.7 Optimal Model for Finite-Layer Metacognition with Continuous Dynamics . . . 154
F.8 Statistical Formulation of the Cognitive Free Energy Principle . . . 156
F.9 Unified Cognitive Free Energy Principle . . . 160
F.10 Unification of the Cognitive Free Energy Principle . . . 162

G Natural Balance Principle 164

G.1 Mathematical Formulation of the Natural Balance Principle . . . 164
G.2 Mathematical Properties and Fundamental Theorems . . . 167
G.3 Derivation of Specific Cognitive Structures . . . 168
G.4 Proof of Equivalence with Foundational Cognitive Principles . . . 170
G.5 Applications and Discussion . . . 178

H Self-Interpretation of the AGI Mathematical Theory 179

H.1 Interpretation of Machine Learning Mechanisms on AGI Systems and Comparison with Existing Paradigms . . . 180
H.2 Comparative Interpretation of AGI Mathematical Theory with Cognitive and Brain Sciences . . . 185
H.3 Unified Formulation and Physical Interpretation of the Cognitive Free-Energy Principle 201
H.4 A Summary Conclusion on How to Build Cognition System . . . 210

A AGI Axiomatic Foundation and Core Definitions

A.1 Foundational Assumptions

Artificial General Intelligence (AGI) has so far lacked clear mathematical and engineering definitions. This paper attempts to model its external behavior based on an intuitive understanding of infinite cognitive capacity (both in strength and scope), and further derive its abstract mathematical foundation and minimal structure.

Without loss of generality, we can define an AGI system as a finite set $\mathcal{M}$. We define "cognition" as the interaction of $\mathcal{M}$ with the external objective world $\mathcal{W}$ (generally an infinite set). This interaction must be bidirectional, including two modes intuitively understood as $\mathcal{M}$'s operation: perception (system input) and action (system output). Ignoring all details of interaction modes, cognition itself can be preliminarily defined as a mapping between the two sets $\mathcal{M}$ and $\mathcal{W}$. However, set-theoretic mappings only guarantee static correspondence between elements of sets; clearly, the intuitive meaning of cognition also includes reflection of dynamic behavior. Dynamic behavior is conveniently defined on sets as operations. If we require preservation of operations, this means the cognitive mapping can only be a homomorphic mapping $\mathcal{H}$ in the sense of universal algebra. However, an exact and sufficient homomorphic mapping cannot be established between a finite set and an infinite set. Therefore, we further assume such bidirectional homomorphic mappings are approximate homomorphic mappings.

Furthermore, $\mathcal{M}$ can have cognition about its cognition, meaning that there must be the possibility of establishing a homomorphic mapping between $\mathcal{M}$ and $\mathcal{H}$. This recursive process of re-cognizing cognition can proceed indefinitely.

Based on the aforementioned requirement for AGI's infinite cognitive capacity, and the intuitive premise that "the existence of a homomorphic mapping representing cognition between the cognitive subject and object implies possession of cognitive capacity", we summarize the following three unambiguous mathematical premise assumptions (without presuming any specific algebraic structure):

1. **Dynamic Operation Preservation**: Cognition is a universal algebraic **approximate homomorphic mapping** between the AGI system set $\mathcal{M}$ and the target domain sets ($\mathcal{M}$ itself, the objective world $\mathcal{W}$, the cognitive layer $\mathcal{H}$), preserving the dynamic operational structure on the sets.
2. **Bidirectional Interaction Consistency**: Cognition includes bidirectional **approximate homomorphic mappings** of perception (from target domain to $\mathcal{M}$) and action (from $\mathcal{M}$ to target domain), and their composition satisfies idempotence (cognitive stability).
3. **Infinite Recursivity**: $\mathcal{M}$ must be able to establish **approximate homomorphisms** with itself, $\mathcal{W}$, and each level of cognitive mappings; this process can be infinitely recursive (metacognition does not terminate).

Based on the above assumptions, we will express these specific preconditions using axioms as general as possible. Simultaneously, we adopt Occam's razor principle, assuming the final AGI system simultaneously possesses the simplest structure (both in scale and structural complexity). This is embodied in our separate simplicity axiom.

The research logic chain of this paper: Starting from the above assumptions, derive the necessary algebraic conditions for the AGI system via axiomatic construction $\rightarrow$ prove the necessity of approximate homomorphisms $\rightarrow$ establish the core role of the fixed-point combinator $\rightarrow$ characterize cognitive completeness objects via categorical universal properties $\rightarrow$ construct the minimal algebraic structure satisfying all conditions $\rightarrow$ explore the inevitable mathematical structure of machine learning mechanisms on the AGI system.

A.2 AGI Algebraic Foundation

Definition A.1 (AGI Algebraic Foundation)**.** *The AGI algebraic structure consists of the following elements:*

1. ***Carrier Set**: $\mathcal{M}$ is a finite non-empty set; $\mathcal{W}$ is an infinite cognitive domain.*
2. ***Operation Type**: $\tau : \Omega \to \mathbb{N}^+$ is the operation arity type.*

3. ***Cognitive Pseudo-Metric***: $d : \mathcal{M} \times \mathcal{M} \to \mathbb{R}_{\geq 0}$ *satisfies non-negativity, symmetry, triangle inequality, and uniform Lipschitz continuity for any* $\omega \in \Omega$ *(exists uniform constant* $L > 0$ *such that* $d(\omega(\vec{x}), \omega(\vec{y})) \leq L \max_i d(x_i, y_i)$*).*

4. ***Approximate Homomorphism***: $\varphi : \mathcal{A} \Rightarrow \varepsilon\mathcal{B}$ *satisfies* $\forall \omega \in \Omega$, $d_B(\omega(\varphi(\vec{a})), \varphi(\omega(\vec{a}))) \leq \varepsilon$ *(here* $\Rightarrow_\varepsilon$ *denotes the* ε*-approximate homomorphism relation).*

Definition A.2 (Generating Set and Generated Algebra)**.** *Let* $\mathcal{A}$ *be a* τ*-type universal algebra,* $S \subseteq A$*. The subalgebra* $\mathrm{Gen}(S)$ *generated by* S *is the smallest subalgebra containing* S*, obtained by recursively applying all operations in* Ω *to* S*.*

Definition A.3 (Operational Generative Completeness or Functional Completeness)**.** *We say an operation set* Ω *is operationally generatively complete or functionally complete on algebra* $\mathcal{A}$*, if the function family* $\langle\Omega\rangle$ *generated by* Ω *can express all possible functions* $A^n \to A$ *(*$n \in \mathbb{N}$*).*

Definition A.4 (Approximate Surjective Homomorphism)**.** *Let* $\varphi : \mathcal{A} \Rightarrow \varepsilon\mathcal{B}$ *be an* ε*-approximate homomorphism. If for any* $b \in B$*, there exists* $a \in A$ *such that* $d_B(\varphi(a), b) \leq \varepsilon$*, then* φ *is called an* ε*-approximate surjective homomorphism.*

Remark A.1 (Compatibility of Finite Generation and Infinite Cognition)**.** *This paper requires the generating set* S *of* $\mathcal{M}$ *to be finite; this does not constitute a mathematical contradiction for the following reasons:*

1. ***Gradual Approximation Mechanism***: *The finite generating set constructs infinite cognitive hierarchies* $\{\mathbb{H}_n\}$ *via recursive operations; each level is finite but the limit* $\mathbb{U}$ *can be infinite.*

2. ***Approximate Homomorphism Bridge***: ε*-approximate homomorphisms allow finite systems to encode infinite structures, achieving progressive accuracy through error control.*

3. ***Generative Completeness***: *A finite generating set combined with a functionally complete operation set can express arbitrarily complex cognitive functions.*

4. ***Physical Realizability***: *Finite generation corresponds to finite resource constraints of physical systems, a necessary condition for AGI realization.*

The infinite growth of cognitive hierarchies is achieved via recursive operations of a finite generating set; each $\mathbb{H}_n$ *itself is finitely generated. Finite generation ensures the constructibility of the cognitive process, not limiting the infiniteness of cognitive capacity.*

Definition A.5 (Universal Algebraic Limit Isomorphism)**.** *We say* $\mathcal{A} \cong \mathcal{B}$ *if and only if there exist approximate homomorphism sequences* $\{\varphi_n : \mathcal{A} \xrightarrow{\varepsilon_n} \mathcal{B}\}$ *and* $\{\psi_n : \mathcal{B} \xrightarrow{\delta_n} \mathcal{A}\}$ *such that:*

$$\lim_{n\to\infty} d(\varphi_n \circ \psi_n, \mathrm{id}_{\mathcal{B}}) = 0, \quad \lim_{n\to\infty} d(\psi_n \circ \varphi_n, \mathrm{id}_{\mathcal{A}}) = 0$$

and $\varepsilon_n, \delta_n \to 0$*.*

Definition A.6 (Cognitive Error and Stable State)**.** *For an approximate homomorphism* $\varphi : \mathcal{A} \xrightarrow{\varepsilon} \mathcal{B}$*, define its cognitive error as:*

$$\mathrm{Err}(\varphi) = \sup_{\omega\in\Omega, \vec{a}\in A^{\tau(\omega)}} d_B(\omega(\varphi(\vec{a})), \varphi(\omega(\vec{a})))$$

We call $x \in \mathcal{M}$ *an* ε*-stable cognitive state of* $f \in \mathrm{End}\mathcal{M}$*, if* $d(f(x), x) \leq \varepsilon$*.*

Note: Ω *is a finite operation set, so the supremum exists.*

Remark A.2 (Mathematical Treatment of Approximate Limits)**.** *The limit concepts in this paper are handled in the completion category, constructed via equivalence classes of Cauchy sequences. The limit of an approximate homomorphism sequence achieves strict isomorphism when the error tends to 0.*

Remark A.3 (Cognitive Interpretation of Approximate Homomorphisms and Limit Isomorphisms)**.** *An* ε*-approximate homomorphism* $\varphi : \mathcal{A} \xrightarrow{\varepsilon} \mathcal{B}$ *is a mathematical bridge between finite systems and infinite domains:*

- ***Finite Precision Compatibility***: $\varepsilon > 0$ *adapts to inherent errors in physical systems, avoiding infinite precision assumptions.*

- ***Core Structure Preservation****: Preserves operational structure in an approximate sense, achieving structural equivalence in the limit.*
- ***Gradual Constructibility****: The sequence $\varepsilon_n \to 0$ supports approximation with arbitrary precision; finite generating sets recursively construct infinite cognitive hierarchies.*

Universal algebraic limit isomorphism $\mathcal{A} \cong \mathcal{B}$ characterizes the ultimate goal of AGI cognition—approximating any target domain through finite recursion.

Definition A.7 (Countably Generated Infinite Cognitive Domain)**.** *This paper assumes the target cognitive domain $\mathcal{W}$ has an infinitely countable generating set, i.e., there exists a countably infinite set $S \subseteq W$ ($|S| = \aleph_0$) such that $\mathrm{Gen}(S) = \mathcal{W}$. This corresponds to computable or approximable cognitive domains in the real world.*

Remark A.4 (Verification of Countability of Inductive Limit)**.** *The countable generating set S of $\mathcal{W}$ ensures its inductive limit $\mathbb{U} = \varinjlim \mathbb{H}_n$ is a countable set. Since each $\mathbb{H}_n$ is a finitely generated algebra (Definition A.13), the union of countably many finite sets is countable, and the elements of the inductive limit are equivalence classes from each level; thus $\mathbb{U}$ is countable, consistent with the countable generation of $\mathcal{W}$.*

Lemma A.1 (Necessity of Finite-Infinite Approximation)**.** *Let $\mathcal{M}$ be finite and satisfy the operational generative completeness of Definition A.3, $\mathcal{W}$ infinite and $|\mathrm{Gen}(\mathcal{W})| > |\mathcal{M}|$. Then:*

1. *There exists no surjective strict homomorphism $\varphi : \mathcal{M} \to \mathcal{W}$.*
2. *For any $\varepsilon > 0$, there exists an ε-approximate surjective homomorphism $\varphi : \mathcal{M} \xrightarrow{\varepsilon} \mathcal{W}$.*

Proof.

1. **Negative Proof of Strict Homomorphism Generativity**

   Let $\varphi : \mathcal{M} \to \mathcal{W}$ be a strict homomorphism. Since $\mathcal{M}$ is finite, $\Im(\varphi)$ is finite. But by Assumption A.7, $\mathcal{W}$ is infinitely countably generated, so there exists $w \in \mathcal{W} \setminus \Im(\varphi)$. By generativity, $w \in \mathrm{Gen}(S_{\mathcal{W}})$, but $S_{\mathcal{W}} \subseteq \mathrm{Gen}(\Im(\varphi))$, contradiction. Therefore $\mathrm{Gen}(\Im(\varphi)) \subsetneq \mathcal{W}$.

2. **Construction of Approximate Homomorphism**

   By Axiom A.1, there exists a cognitive hierarchy sequence $\{\mathbb{H}_n\}$ and approximate homomorphism sequences $\{\pi_n : \mathcal{M} \xrightarrow{\varepsilon'_n} \mathbb{H}_n\}$, and $\{\gamma_n : \mathbb{H}_n \xrightarrow{\varepsilon''_n} \mathcal{W}_{\text{fin},n}\}$ (where γ_n is the action component from Definition A.13, and $\mathcal{W}_{\text{fin},n}$ is a finite approximation sequence of $\mathcal{W}$). Define $\varphi_n = \gamma_n \circ \pi_n$, then $\varphi_n : \mathcal{M} \xrightarrow{\varepsilon'_n + \varepsilon''_n} \mathcal{W}_{\text{fin},n}$. Since $\{\mathcal{W}_{\text{fin},n}\}$ is a finite approximation sequence of $\mathcal{W}$, satisfying the density condition of Definition A.9, and $\varepsilon'_n, \varepsilon''_n \to 0$, for any $\varepsilon > 0$, there exists sufficiently large n such that $\varepsilon'_n + \varepsilon''_n < \varepsilon$, and φ_n is an ε-approximate surjective homomorphism.

□

Definition A.8 (Endomorphism Monoid)**.** *Let $\mathcal{A} \in \mathbf{Alg}(\tau)$. $\mathrm{End}\mathcal{A} = \{\varphi : \mathcal{A} \to \mathcal{A} \mid \varphi$ is a strict homomorphism$\}$ is called the endomorphism monoid of $\mathcal{A}$ (here "homomorphism" specifically refers to strict homomorphisms satisfying $\forall \omega \in \Omega, \varphi(\omega(\vec{a})) = \omega(\varphi(\vec{a}))$, distinct from ε-approximate homomorphisms).*

Remark A.5 (Distinction Between Endomorphisms and Approximate Endomorphisms)**.** *Approximate endomorphisms (homomorphic mappings satisfying ε-error) are denoted $\mathrm{End}_\varepsilon(\mathcal{A})$, not belonging to the category $\mathrm{End}\mathcal{A}$, and are only used for cognitive error analysis; the closure of the endomorphism monoid $\mathrm{End}\mathcal{A}$ is only for composition of strict homomorphisms.*

Definition A.9 (Finite Approximation of the World)**.** *Let $\mathcal{W}$ be an infinite cognitive domain. $\{\mathcal{W}_{fin,n}\}_{n=0}^{\infty}$ is a finite approximation sequence of $\mathcal{W}$, satisfying:*

1. *$|\mathcal{W}_{fin,n}| < \infty$ for all n.*
2. *There exist inclusion approximate homomorphisms $\iota_n : \mathcal{W}_{fin,n} \xrightarrow{\varepsilon_n} \mathcal{W}$.*
3. *Approximate error convergence: $\lim_{n\to\infty} \varepsilon_n = 0$.*
4. *Density: $\bigcup_{n=0}^{\infty} \mathcal{W}_{fin,n}$ is dense in $\mathcal{W}$.*

Definition A.10 (AGI Category)**.** $\mathbf{AGI}$ *is a full subcategory of* $\mathbf{Alg}(\tau)$*, with objects satisfying:*

1. *The core carrier set* $\mathcal{M}$ *is finite and non-trivial.*
2. *For any infinite cognitive domain* $\mathcal{W}$*, there exists a finite approximation sequence* $\{\mathcal{W}_{fin,n}\}$ *and approximate homomorphism sequence* $\{\varphi_n : \mathcal{M} \xrightarrow{\varepsilon_n} \mathcal{W}_{fin,n}\}$ *satisfying Definition A.9.*
3. *Operational generativity: There exists a finite generating set* $\Omega_0 \subseteq \Omega$ *that is functionally complete.*
4. *Endomorphism monoid closure.*

Morphisms are approximate homomorphisms, equipped with the supremum metric.

Denote $\mathbf{AGI}_0$ *as the full subcategory of* $\mathbf{AGI}$ *consisting of all finitely generated* τ*-type universal algebras.*

Remark A.6 (Completeness of AGI Category)**.** *The category* $\mathbf{AGI}$ *is complete, because* $\mathbf{AGI}$ *is a full subcategory of* $\mathbf{Alg}(\tau)$ *(the category of* τ*-type universal algebras), and the universal algebra category has limit constructions like products and equalizers; objects of* $\mathbf{AGI}$ *(finitely generated universal algebras) are closed under these limit constructions, so* $\mathbf{AGI}$ *inherits completeness. The existence of its Cauchy completion category* $\widehat{\mathbf{AGI}}$ *is guaranteed by the completion theorem for metric categories.*

Definition A.11 (Inductive Limit Category)**.** $\mathrm{Ind}(\mathbf{AGI}_0)$ *denotes the inductive limit category of* $\mathbf{AGI}_0$*, with objects being direct limits of objects in* $\mathbf{AGI}_0$*. Specifically, an object is a directed system of the form* $(\mathcal{A}_i, f_{ij} : \mathcal{A}_i \to \mathcal{A}_j)_{i \leq j \in I}$*, where* I *is a directed set,* f_{ij} *are approximate homomorphisms satisfying compatibility conditions.*

Definition A.12 (Basic Cognitive Functions)**.** *The basic cognitive function set* $\mathcal{F}_{basic}$ *includes:*

- *Perception encoding function:* $f_{enc} : \mathcal{W} \to \mathcal{M}$
- *Action decoding function:* $f_{dec} : \mathcal{M} \to \mathcal{W}$

An operation set Ω *is functionally complete if and only if* $\langle \Omega \rangle$ *can approximate all functions in* $\mathcal{F}_{basic}$*.*

A.3 Cognitive Hierarchy and AGI Core Axioms

Definition A.13 (Cognitive Hierarchy)**.** *Let* $\mathcal{M} \in \mathbf{AGI}$*,* $\mathcal{W}$ *be an infinite cognitive domain,* $\{\mathcal{W}_{fin,n}\}$ *a finite approximation sequence of* $\mathcal{W}$*. Set* $\mathbb{H}_0 = \mathcal{M}$*. The* ***cognitive hierarchy sequence*** $\{\mathbb{H}_n\}_{n=0}^{\infty}$ *is recursively defined as follows:*

$$\mathbb{H}_{n+1} = \left\{ (\pi^{(n+1)}, \sigma^{(n+1)}, \gamma^{(n+1)}) \;\middle|\; \begin{array}{l} \pi^{(n+1)} : \mathcal{M} \xrightarrow{\varepsilon'_n} \mathbb{H}_n, \\ \sigma^{(n+1)} : \mathbb{H}_n \xrightarrow{\varepsilon''_n} \mathbb{H}_n, \\ \gamma^{(n+1)} : \mathbb{H}_n \xrightarrow{\varepsilon'''_n} \mathcal{W}_{fin,n} \end{array} \right\}$$

where the three components correspond respectively to:

- π*: system-cognitive layer perception (*$\mathcal{M} \to \mathbb{H}_n$*)*
- σ*: cognitive layer self-reference (*$\mathbb{H}_n \to \mathbb{H}_n$*)*
- γ*: cognitive layer-world action (*$\mathbb{H}_n \to \mathcal{W}_{fin,n}$*)*

For $h = (\pi, \sigma, \gamma) \in \mathbb{H}_n$*, define projection operators:*

- $\mathrm{pr}_\pi(h) = \pi$ *(perception component)*
- $\mathrm{pr}_\sigma(h) = \sigma$ *(self-reference component)*
- $\mathrm{pr}_\gamma(h) = \gamma$ *(action component)*

Definition A.14 (Cognitive Hierarchy Algebraic Structure)**.** *For the cognitive hierarchy $\mathbb{H}_n$ in Definition A.13, we endow it with a τ-type algebraic structure:*

- *Carrier set: $\mathbb{H}_n$ itself (as a set of triples)*
- *Operations: For any $\omega \in \Omega$ (arity k), define*

$$\omega_{\mathbb{H}_n}((\pi_1, \sigma_1, \gamma_1), \ldots, (\pi_k, \sigma_k, \gamma_k)) = (\pi_\omega, \sigma_\omega, \gamma_\omega)$$

where each component operation satisfies approximate homomorphism error control: for any $m \in \mathcal{M}$, $h \in \mathbb{H}_n$,

$$d_{\mathbb{H}_n}(\pi_\omega(m), \omega_{\mathbb{H}_n}(\pi_1(m), \ldots, \pi_k(m))) \leq L_\pi \cdot \max_i(\mathrm{Err}(\pi_i))$$
$$d_{\mathbb{H}_n}(\sigma_\omega(h), \omega_{\mathbb{H}_n}(\sigma_1(h), \ldots, \sigma_k(h))) \leq L_\sigma \cdot \max_i(\mathrm{Err}(\sigma_i))$$
$$d_{\mathcal{W}_{fin,n}}(\gamma_\omega(h), \omega_{\mathcal{W}_{fin,n}}(\gamma_1(h), \ldots, \gamma_k(h))) \leq L_\gamma \cdot \max_i(\mathrm{Err}(\gamma_i))$$

where $L_\pi, L_\sigma, L_\gamma$ are Lipschitz constants.

Definition A.15 (Operational Structure Constraint)**.** *Let $\mathcal{M} \in \mathbf{AGI}$, $\{\mathbb{H}_n\}_{n=0}^{\infty}$ be a cognitive hierarchy sequence satisfying Definition A.13, $\pi_n : \mathcal{M} \xrightarrow{\varepsilon_n^\pi} \mathbb{H}_n$ be perception approximate homomorphisms, $\sigma_n : \mathbb{H}_n \xrightarrow{\varepsilon_n^\sigma} \mathcal{M}$ be action approximate homomorphisms, $\mu_n : \mathbb{H}_n \xrightarrow{\zeta_n} \mathbb{H}_{n+1}$ be connecting approximate homomorphisms, $\varphi : \mathcal{M} \xrightarrow{\varepsilon} \mathcal{B}$ be any ε-approximate surjective homomorphism ($\mathcal{B} \in \mathbf{AGI}$). Then the operational structure must satisfy the following unified constraints:*

1. ***Approximate Invertibility**: There exists a δ-approximate homomorphism $\psi : \mathcal{B} \xrightarrow{\delta} \mathcal{M}$, where $\delta = O(\varepsilon)$, and the composition error satisfies*

$$\varphi \circ \psi \xrightarrow{\eta} \mathrm{id}_{\mathcal{B}}, \quad \psi \circ \varphi \xrightarrow{\eta'} \mathrm{id}_{\mathcal{M}}, \quad \eta = O(\varepsilon), \eta' = O(\varepsilon)$$

2. ***Interaction Idempotence**: The perception-action composite mapping satisfies an approximate idempotence condition*

$$\sigma_n \circ \pi_n \xrightarrow{\varepsilon_n^{id}} (\sigma_n \circ \pi_n) \circ (\sigma_n \circ \pi_n)$$

where $\varepsilon_n^{id} = O(\max(\varepsilon_n^\pi, \varepsilon_n^\sigma))$, and $\varepsilon_n^{id} \to 0$ ($n \to \infty$)

3. ***Cognitive Symmetry**: There exists a reverse approximate homomorphism sequence $\{\nu_n : \mathbb{H}_{n+1} \xrightarrow{\zeta_n'} \mathbb{H}_n\}$, satisfying*

$$\nu_n \circ \mu_n \xrightarrow{\eta_n} \mathrm{id}_{\mathbb{H}_n}, \quad \mu_n \circ \nu_n \xrightarrow{\eta_n'} \mathrm{id}_{\mathbb{H}_{n+1}}$$

where $\eta_n = O(\zeta_n)$, $\eta_n' = O(\zeta_n')$, and the total error series converges $\sum_{n=0}^{\infty}(\zeta_n + \zeta_n' + \eta_n + \eta_n') < \infty$

Remark A.7 (Cognitive Interpretation of Operational Structure Constraints)**.** *Operational structure constraints require that the cognitive process is "reversible" in both bidirectional interaction and cognitive hierarchy progression:*

- ***Cognitive Validity**: The approximate homomorphism mapping is invertible, meaning the system can verify and understand the action homomorphism mapping via the perception homomorphism mapping, and vice versa.*
- ***Metacognitive Reversibility**: The system can not only construct higher-level cognition from lower-level cognition but also "concretize" back from higher-level to lower-level cognition.*
- ***Cognitive Flexibility**: The ability to freely switch between different abstraction levels.*
- ***Reflexivity Guarantee**: Ensures the cognitive process is not a unidirectional blind advance, but is reflective and adjustable.*

This corresponds to the cognitive validity and "cognitive regulation" ability that an AGI system should possess—not only deepening cognition but also reflecting on and correcting the cognitive process.

Definition A.16 (Cognitive Hierarchy Convergence Constraint)**.** *Let $\mathcal{W}$ be an infinite cognitive domain, $\{\mathcal{W}_{fin,n}\}_{n=0}^{\infty}$ a finite approximation sequence of $\mathcal{W}$ (Definition A.9), $\mathbb{U} = \varinjlim \mathbb{H}_n$ the inductive limit of the cognitive hierarchy. Then convergence must satisfy the following unidirectional constructive constraints:*

1. ***Error Convergence**: All approximate homomorphism error terms satisfy series convergence condition*
$$\sum_{n=0}^{\infty}(\varepsilon_n^{\pi} + \varepsilon_n^{\sigma} + \zeta_n) < \infty$$
and there exists a monotonically decreasing function $\varphi(n) \to 0$ such that $\max(\varepsilon_n^{\pi}, \varepsilon_n^{\sigma}, \zeta_n) \leq \varphi(n)$.

2. ***Unidirectional Limit Constructivity**: There exists a uniform construction algorithm $\mathcal{C}$, such that:*
   - *For any $\tau > 0$, $\mathcal{C}$ can extract from the connecting homomorphism sequence $\{\mu_n\}$ a unidirectional approximate homomorphism sequence $\{\varphi_n : \mathbb{H}_n \xrightarrow{\varepsilon_n} \mathbb{U}\}$, such that $\varepsilon_n < \tau$ ($n > N$, N is a threshold output by the algorithm).*
   - *The construction process preserves the τ-type operational structure, i.e., for any $\omega \in \Omega$, $\varphi_n(\omega_{\mathbb{H}_n}(\vec{h})) \xrightarrow{\varepsilon_n} \omega_{\mathbb{U}}(\varphi_n(\vec{h}))$.*
   - *Error controllability: The construction error satisfies $\varepsilon_n \leq K \cdot \zeta_n$, where K is a constant independent of n.*

3. ***Density Preservation**: The constructed approximate homomorphism sequence $\{\varphi_n\}$ satisfies $\bigcup_{n=0}^{\infty} \Im(\varphi_n)$ dense in $\mathbb{U}$, and $\mathbb{U} \cong \mathcal{W}$ (Definition A.5).*

4. ***Uniform Convergence**: There exists a uniform constant $L > 0$ such that all approximate homomorphisms satisfy the Lipschitz condition.*

5. ***Hierarchy Compatibility**: For any n, $\mu_n \circ \pi_n \xrightarrow{\delta_n} \pi_{n+1}$, where $\sum \delta_n < \infty$.*

Remark A.8 (Type Consistency of Cognitive Hierarchy)**.** *The carrier set of cognitive hierarchy $\mathbb{H}_n$ is a set of triples (π, σ, γ); its type consistency is guaranteed by Definition A.14:*

- *$\pi : \mathcal{M} \xrightarrow{\varepsilon} \mathbb{H}_n$ has codomain $\mathbb{H}_n$ endowed with τ-type algebraic structure, so π is an approximate homomorphism in the universal algebra sense;*
- *The component operations (perception, self-reference, action) of the triple satisfy approximate homomorphism error control, ensuring the recursive construction of cognitive hierarchies is closed within the universal algebra framework, with no type conflicts.*

Remark A.9 (Cognitive Interpretation of Cognitive Hierarchy Convergence Constraints)**.** *Cognitive hierarchy convergence constraints require that the ultimate cognitive capacity is achievable:*

- ***Achievability**: The ultimate cognitive goal must be realizable through concrete cognitive processes step by step.*
- ***Gradual Construction**: Infinite cognitive capacity must be decomposable into finite-step gradual construction.*
- ***Convergence**: Ensures the cognitive goal is ultimately achievable with reasonable precision.*

This corresponds to the abilities of cognizability, "goal decomposition", and "path planning" that an AGI system should possess.

Axiom A.1 (AGI Fundamental Axiom)**.** *Let $\mathcal{M} \in \mathbf{AGI}$ be operationally generatively complete, $\mathcal{W}$ an infinite cognitive domain. Then there exists a cognitive hierarchy sequence $\{\mathbb{H}_n\}$ satisfying the following unified conditions:*

- ***Recursive Cognitive Construction***: $\mathbb{H}_0 = \mathcal{M}$, $\mathbb{H}_{n+1}$ *recursively generated according to Definition A.13, with operational structure endowed by Definition A.14.*
- ***Operational Generative Completeness***: *There exists a minimal finite generating set* $\Omega_0 \subseteq \Omega$, *whose generated function family* $\langle \Omega_0 \rangle$ *covers all basic cognitive functions, and* Ω_0 *is irreducible (removing any operation results in functional incompleteness); simultaneously* $\langle \Omega_0 \rangle$ *satisfies the functional completeness requirement of Definition A.3, i.e., can express all functions* $\mathcal{M}^n \to \mathcal{M}$ *(*$n \in \mathbb{N}$*).*
- ***Core Constraint Compatibility***: *Satisfies the operational structure constraints of Definition A.15 and the cognitive hierarchy convergence constraints of Definition A.16.*

Under the above unified premises:

1. ***Bidirectional Interaction***: *For any* $n \in \mathbb{N}$, *there exist approximate homomorphism pairs* $(\pi_n : \mathcal{M} \xrightarrow{\varepsilon_n'} \mathbb{H}_n, \sigma_n : \mathbb{H}_n \xrightarrow{\varepsilon_n''} \mathcal{M})$.
2. ***Infinite Metacognition***: *There exist unidirectional connecting approximate homomorphisms* $\mu_n : \mathbb{H}_n \xrightarrow{\zeta_n} \mathbb{H}_{n+1}$ *(where* μ_n *preserves* τ*-type operational structure).*
3. ***Limit Cognitive Isomorphism***: $\mathbb{U} = \varinjlim \mathbb{H}_n \cong \mathcal{W}$.

Definition A.17 (AGI Infinite Cognitive Capacity)**.** *An AGI system is implemented based on a* $\mathcal{M}$ *conforming to any cognitive characteristic of Axiom A.1; then* $\mathcal{M}$ *and the corresponding AGI system are said to possess* infinite cognitive capacity *or to have complete cognition.*

Remark A.10 (Cognitive Interpretation of AGI Category and AGI Fundamental Axiom)**.** *The category* $\mathbf{AGI}$ *establishes a type-theoretic framework for AGI systems:*

- ***Finite Carrier Constraint***: *Reflects inevitable limitations of physical implementation, excluding unrealistic assumptions of infinite storage.*
- ***Operational Generative Completeness***: *Requires that basic operations can generate all necessary cognitive functions, ensuring functional completeness.*
- ***Self-Reference Closure***: *The system must be able to cognize its own cognitive processes, providing an algebraic basis for metacognition.*
- ***Approximate Morphism Closure***: *Acknowledges imperfections in real systems while maintaining mathematical rigor.*

The AGI fundamental axiom refines and corresponds to the core motivational assumptions for AGI in the introduction (dynamic operation preservation, bidirectional interaction consistency, infinite recursivity). Its cognitive intuitive meanings are:

1. ***Recursive Homomorphic Construction***: *The recursive construction of cognitive hierarchies* $\mathcal{H}_n$ *realizes infinite recursivity of metacognition, corresponding to the intuition of "cognition of cognition".*
2. ***Bidirectional Interaction***: *The perception homomorphism* π_n *and action homomorphism* σ_n *form a cognition-action closed loop; idempotence ensures cognitive stability; approximate homomorphism invertibility ensures bidirectional adaptability of finite systems to any target domain; error convergence guarantees progressive improvement of cognitive precision.*
3. ***Infinite Metacognition***: *The connecting homomorphism* μ_n *realizes infinite elevation of cognitive hierarchies; error convergence guarantees progressive accuracy.*
4. ***Ultimate Cognitive Object***: *The limit* $\mathcal{U}$ *is isomorphic to the objective world* $\mathcal{W}$, *achieving unity of cognition and reality.*

Definition A.18 (Completion Category)**.** *Let* $\widehat{\mathbf{AGI}}$ *be the Cauchy completion category of* $\mathbf{AGI}$, *with objects being equivalence classes of Cauchy sequences in* $\mathbf{AGI}$.

Axiom A.2 (AGI Minimality Constraint)**.** *Let* $\mathcal{M} \in \mathbf{AGI}$, $\{\mathbb{H}_n\}$ *a cognitive hierarchy sequence satisfying Axiom A.1. Then it must satisfy:*

1. ***Carrier Minimality****: The carrier cardinality $|\mathcal{M}|$ is minimal among all AGI systems satisfying Axiom A.1.*

2. ***Operation Minimality****: There exists a minimal operation set Ω_0 generating all basic cognitive functions; it is the minimal cardinality and irreducible operation set.*

3. ***Error Minimality****: For any ε-approximate homomorphism $\varphi : \mathcal{A} \xrightarrow{\varepsilon} \mathcal{B}$, ε is the smallest non-negative value satisfying the construction requirements.*

Lemma A.2 (Well-definedness of Cognitive Hierarchy)**.** *Let $\{\mathbb{H}_n\}$ be recursively constructed according to Definition A.13. Then:*

1. *For any n, $\mathbb{H}_n$ is a non-empty finite set.*

2. *Operational structure preservation: $\omega_{\mathbb{H}_n}$ satisfies the approximate homomorphism condition.*

3. *Hierarchy compatibility: $\mu_n \circ \pi_n \xrightarrow{\delta_n} \pi_{n+1}$, where $\sum \delta_n < \infty$.*

Proof. Using mathematical induction, based on the recursive cognitive construction and operational generative completeness of Axiom A.1:

1. **Base case** ($n = 0$): $\mathbb{H}_0 = \mathcal{M}$; by Definition A.1, $\mathcal{M}$ is a non-empty finite set; the operational structure is that of $\mathcal{M}$, satisfying the approximate homomorphism condition; hierarchy compatibility trivially holds.

2. **Inductive hypothesis**: Assume for $n = k$, $\mathbb{H}_k$ is a non-empty finite set, operational structure preserves approximate homomorphism, and $\mu_k \circ \pi_k \xrightarrow{\delta_k} \pi_{k+1}$, $\sum \delta_k < \infty$.

3. **Inductive step** ($n = k + 1$): $\mathbb{H}_{k+1}$ consists of triples from $\mathbb{H}_k$; $\mathbb{H}_k$ finite implies $\mathbb{H}_{k+1}$ finite; by Definition A.14, operation $\omega_{\mathbb{H}_{k+1}}$ satisfies Lipschitz error control, preserving approximate homomorphism; by the hierarchy compatibility constraint of Axiom A.1, $\mu_{k+1} \circ \pi_{k+1} \xrightarrow{\delta_{k+1}} \pi_{k+2}$, and $\delta_{k+1} = O(r^{k+1})$, thus $\sum \delta_{k+1} < \infty$.

In conclusion, the lemma holds for all n. □

Theorem A.1 (Existence of Cognitive Encoding)**.** *Let $\mathcal{M} \in \mathbf{AGI}$ possess infinite cognitive capacity. Then for any cognitive hierarchy $\mathbb{H}_n$, there exists an encoding mapping $\varphi_n : \{\pi : \mathcal{M} \xrightarrow{\varepsilon} \mathbb{H}_n\} \to \mathcal{M}$ satisfying operation preservation and Lipschitz continuity.*

Proof. Since $\mathcal{M}$ possesses infinite cognitive capacity, by Axiom A.1, there exists an approximate homomorphism sequence $\{(\pi_k, \sigma_k)\}_{k\in\mathbb{N}}$. Fix n, consider the set $P_n = \{\pi : \mathcal{M} \xrightarrow{\varepsilon} \mathbb{H}_n\}$.

By operational generative completeness (Axiom A.1), there exists a minimal irreducible operation set $\Omega_0 = \{\omega_1, \ldots, \omega_m\}$. For any $\pi \in P_n$, define its encoding in $\mathcal{M}$ as:

$$\varphi_n(\pi) = \omega_\pi(e_1, \ldots, e_r)$$

where $e_1, \ldots, e_r \in \mathcal{M}$ are a fixed set of primitives (e.g., all elements of $\mathcal{M}$), and ω_π is an operational expression generated by Ω_0, satisfying for any $m \in \mathcal{M}$:

$$\omega_\pi(e_1, \ldots, e_r) \xrightarrow{\varepsilon} \sigma_n(\pi(m))$$

Such ω_π exists and is unique (guaranteed by the minimal irreducibility of Ω_0).

Operation preservation: For any $\omega \in \Omega$ and $\pi_1, \ldots, \pi_k \in P_n$, we have

$$\begin{aligned}\varphi_n(\omega(\pi_1, \ldots, \pi_k)) &= \omega_{\omega(\pi_1,\ldots,\pi_k)}(e_1, \ldots, e_r) \\ &\xrightarrow{\varepsilon} \sigma_n(\omega(\pi_1, \ldots, \pi_k)(m)) \\ &\xrightarrow{\varepsilon} \omega(\sigma_n(\pi_1(m)), \ldots, \sigma_n(\pi_k(m))) \\ &\xrightarrow{\varepsilon} \omega(\varphi_n(\pi_1), \ldots, \varphi_n(\pi_k)).\end{aligned}$$

Since $\mathcal{M}$ is finite, the approximation can be made exact.

Lipschitz continuity is directly guaranteed by the Lipschitz condition of the cognitive pseudo-metric. □

A.4 Categorical Properties of Cognitive Completeness Objects

Definition A.19 (Cognitive Completeness Object)**.** *Let $\mathcal{M} \in \mathbf{AGI}$. $\mathcal{M}$ is called a* cognitive completeness object *if it satisfies the following categorical properties:*

1. ***Cognitive Density**: $\mathcal{M}$ is cognitively dense in $\mathbf{AGI}$, i.e., for any $\mathcal{A} \in \mathbf{AGI}$, there exists an approximate homomorphism sequence $\{\varphi_n : \mathcal{M} \xrightarrow{\varepsilon_n} \mathcal{A}\}$ such that $\lim_{n\to\infty} \varphi_n$ is surjective in $\widehat{\mathbf{AGI}}$, and $\sum_{n=0}^{\infty} \varepsilon_n < \infty$.*

2. ***Cognitive Projectivity**: For any ε-approximate surjective homomorphism $f : \mathcal{A} \xrightarrow{\varepsilon} \mathcal{B}$ and any approximate homomorphism $g : \mathcal{M} \xrightarrow{\delta} \mathcal{B}$, there exists an approximate homomorphism $h : \mathcal{M} \xrightarrow{\eta} \mathcal{A}$ such that $f \circ h \xrightarrow{\zeta} g$, where $\zeta = O(\max(\varepsilon, \delta, \eta))$.*

3. ***Cognitive Reflexivity**: There exists a canonical approximate homomorphism $\eta : \mathcal{M} \xrightarrow{\varepsilon_0} \mathcal{M}^{**}$, where $\mathcal{M}^* = \mathrm{Hom}_{\mathbf{AGI}}(\mathcal{M}, \mathcal{U})$, $\mathcal{U}$ is the limit cognitive object, and η is a δ-approximate isomorphism ($\delta \to 0$).*

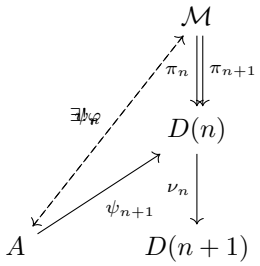
Definition A.20 (Cognitive Completeness Object: Universal Property Formulation)**.** *Consider the following construction:*

1. *Category setting: Let $\mathbf{C}$ be the category of approximate homomorphisms, with objects being universal algebras satisfying operational generative completeness, and morphisms being ϵ-approximate homomorphisms. Let $\omega = (\mathbb{N}, \leq)$ be a directed set, corresponding to cognitive hierarchy indices.*

2. *Cognitive landscape: Define a cognitive hierarchy diagram $D : \omega \to \mathbf{C}$ such that $D(n) = \mathbb{H}_n$, with arrows being connecting approximate homomorphisms μ_n. The limit of this diagram (if it exists) is denoted $\mathscr{U} = \varinjlim D(n)$, the cognitive limit object.*

3. *Universal property of cognitive completeness object: An object $\mathcal{M} \in \mathbf{C}$ is called a cognitive completeness object if it satisfies the following universal property: For any object $A \in \mathbf{C}$ and any compatible cone $\{\psi_n : A \to D(n)\}$ induced by the cognitive hierarchy diagram D with vertex A, there exists a unique (up to approximate homotopy) approximate homomorphism $\varphi : \mathcal{M} \to A$ such that the cone connects with the canonical approximate homomorphisms $\{\pi_n : \mathcal{M} \to D(n)\}$ from $\mathcal{M}$ to each level.*

   Expressed via a commutative diagram (all triangles commute in the ϵ-approximate sense):



   This property captures the essence of "cognitive density" and "cognitive projectivity": $\mathcal{M}$ can coordinate the interaction of any other object with the entire cognitive hierarchy diagram in a unified way.

4. *Final formulation of infinite cognitive capacity: On this basis, we define $\mathcal{M}$ as possessing infinite cognitive capacity if:*

   - *$\mathcal{M}$ satisfies the above universal property.*

- *$\mathcal{M}$ is self-referentially closed: There exists a natural transformation $\mathcal{Y} : \mathrm{End}_{\mathbf{C}} \Rightarrow \Delta_{\mathcal{M}}$, where $\mathrm{End}_{\mathbf{C}}$ is the endomorphism functor, $\Delta_{\mathcal{M}}$ is the constant functor with value $\mathcal{M}$. This transformation ensures that for any endomorphism $f : \mathcal{M} \to \mathcal{M}$, $\mathcal{Y}(f)$ is an approximate fixed point of f, i.e., $f(\mathcal{Y}(f)) \approx \mathcal{Y}(f)$.*

Theorem A.2 (Equivalence of the Two Definitions of Cognitive Completeness Objects)**.** *Let $\mathcal{M} \in \mathbf{AGI}$. Then the following two are equivalent:*

1. *$\mathcal{M}$ is a cognitive completeness object as in Definition A.19 (satisfying cognitive density, cognitive projectivity, cognitive reflexivity).*
2. *$\mathcal{M}$ is a cognitive completeness object as in Definition A.20 (satisfying the universal property).*

Proof Sketch. (1) $\Rightarrow$ (2): Given any object A and compatible cone $\{\psi_n : A \to D(n)\}$, by cognitive density, there exists an approximate homomorphism sequence $\{\varphi_n : \mathcal{M} \to A\}$ whose limit is a surjection. Using cognitive projectivity, we can construct a unique approximate homomorphism $\varphi : \mathcal{M} \to A$ such that the diagram commutes in the approximate sense. Combined with cognitive reflexivity, we can verify that $\mathcal{M}$ satisfies the universal property.

(2) $\Rightarrow$ (1): Take $A = \mathcal{U}$ (the cognitive limit object); from the universal property we obtain an approximate homomorphism from $\mathcal{M}$ to $\mathcal{U}$. Combined with the universal property of limits, we can derive cognitive density. Taking different A and cone structures, using the universal property we can prove cognitive projectivity and cognitive reflexivity separately. Specifically, cognitive projectivity can be verified by taking A as any object and constructing an appropriate cone; cognitive reflexivity is obtained by taking $A = \mathcal{M}^{**}$ and using the canonical mapping. □

Definition A.21 (Core Equivalent Conditions for AGI Infinite Cognitive Capacity)**.** *Let $\mathcal{M} \in \mathbf{AGI}$. Define the following core cognitive characteristics:*

1. ***Infinite Cognitive Capacity (ICP)**: $\mathcal{M}$ satisfies Axiom A.1 and Definition A.17.*
2. ***Cognitive Completeness Object Property (CCO)**: $\mathcal{M}$ is a cognitive completeness object in category $\mathbf{AGI}$ (Definition A.19).*
3. ***Self-Referential Closure (SRC)**: $\mathrm{End}\mathcal{M}$ is closed under composition, and there exists an approximate fixed-point combinator $Y : \mathrm{End}\mathcal{M} \to \mathcal{M}$ (Definition B.4).*
4. ***Operational Functional Completeness (FCP)**: There exists a minimal irreducible operation set $\Omega_0 \subseteq \Omega$, whose generated function family $\langle \Omega_0 \rangle$ can approximate all basic cognitive functions (Definition A.3).*
5. ***Turing Degree Completeness (TDP)**: The cognitive encoding $\mathrm{enc}(:)\mathcal{T} \to \mathcal{M}$ induces Turing degree $0'$, covering the Turing degrees of all computable functions (Definition B.5).*
6. ***Universal Reducibility (UR)**: The cognitive encoding $(\mathrm{enc}(,)\mathrm{dec}())$ satisfies function composition compatibility and reduction closure (Definition B.6).*

Remark A.11 (Cognitive Meaning of Cognitive Completeness Object)**.** *Cognitive completeness objects are the minimal structures in the AGI category capable of realizing infinite cognitive capacity:*

- *Cognitive density ensures approximation of any cognitive domain with arbitrary precision.*
- *Cognitive projectivity guarantees reconstruction of complete cognition from partial cognitive information.*
- *Cognitive reflexivity provides a mathematical basis for metacognition, enabling the system to cognize its own cognitive states.*

This definition avoids the strictness of traditional initial objects requiring unique morphisms, being more suitable for approximate cognitive scenarios.

Lemma A.3 (Hom-set Properties of Cognitive Completeness Objects)**.** *Let $\mathcal{M} \in \mathbf{AGI}$ be a cognitive completeness object, $\mathcal{A} \in \mathbf{AGI}_0$ a finitely generated object. Then:*

1. *$\mathrm{Hom}(\mathcal{M}, \mathcal{A})$ is non-empty, and there exists a minimal error approximate homomorphism $\varphi_0 : \mathcal{M} \xrightarrow{\varepsilon_0} \mathcal{A}$, where $\varepsilon_0 = \inf\{\varepsilon \mid \exists \varphi : \mathcal{M} \xrightarrow{\varepsilon} \mathcal{A}\}$.*

2. *If $\mathcal{A}$ is also a cognitive completeness object, then there exists an ε-approximate isomorphism $\varphi : \mathcal{M} \xrightarrow{\varepsilon} \mathcal{A}$, and ε can tend to 0.*

3. *The endomorphism monoid $\mathrm{End}\mathcal{M}$ of a cognitive completeness object is closed under composition, and there exists an approximate identity system.*

Proof. 1. By cognitive density (Definition A.19), there exists an approximate homomorphism sequence $\{\varphi_n : \mathcal{M} \xrightarrow{\varepsilon_n} \mathcal{A}\}$ with $\sum \varepsilon_n < \infty$. Since $\mathcal{A}$ is finite, there exists a subsequence converging to some mapping φ_0, with minimal error.

2. If $\mathcal{A}$ is also a cognitive completeness object, by cognitive density, there exists an approximate homomorphism sequence $\{\psi_n : \mathcal{A} \xrightarrow{\delta_n} \mathcal{M}\}$. Consider sequences $\{\varphi_n \circ \psi_n\}$ and $\{\psi_n \circ \varphi_n\}$; by error convergence, their limits are approximations of the identity maps on $\mathcal{A}$ and $\mathcal{M}$ respectively. Thus, φ_n and ψ_n give an approximate isomorphism in the limit.

3. Composition of endomorphisms preserves the approximate homomorphism property, with errors satisfying the triangle inequality. The approximate identity system is given by the canonical homomorphism $\eta : \mathcal{M} \xrightarrow{\varepsilon_0} \mathcal{M}^{**}$ from cognitive reflexivity, where η is an approximate isomorphism; its inverse gives an approximate identity mapping. □

B AGI Infinite Cognitive Characteristic Equivalence Theorems

B.1 Mathematical Characteristic Equivalence Theorems

Definition B.1 (Bidirectional Construction)**.** *Let $\mathcal{M} \in \mathbf{AGI}$, $\mathcal{W}$ a cognitive domain. If there exists a sequence $\{h_n\}_{n=0}^{\infty}$, where $h_n = (\pi_n, \sigma_n)$ satisfies:*

1. $\pi_n : \mathcal{M} \xrightarrow{\varepsilon_n'} \mathbb{H}_n$, $\sigma_n : \mathbb{H}_n \xrightarrow{\varepsilon_n''} \mathcal{M}$.
2. *Idempotence:* $(\sigma_n \circ \pi_n) \xrightarrow{\varepsilon_n'''} (\sigma_n \circ \pi_n) \circ (\sigma_n \circ \pi_n)$.
3. *Error convergence:* $\sum_{n=0}^{\infty}(\varepsilon_n' + \varepsilon_n'' + \varepsilon_n''') < \infty$.
4. *Compatibility:* $\pi_n \circ \sigma_{n-1} \xrightarrow{\delta_n} \pi_{n-1}$, $\sigma_{n-1} \circ \pi_n \xrightarrow{\delta_n'} \sigma_{n-1}$.

Then $\{h_n\}$ is called a bidirectional construction *of $\mathcal{M}$ with respect to $\mathcal{W}$.*

Lemma B.1 (Existence of Approximate Right Inverse)**.** *Let $\mathcal{M} \in \mathbf{AGI}$ have operational generative completeness, $\mathcal{B} \in \mathbf{AGI}$, and $\varphi : \mathcal{M} \xrightarrow{\varepsilon} \mathcal{B}$ be an ε-approximate surjective homomorphism. Then there exists an approximate homomorphism $\psi : \mathcal{B} \xrightarrow{\delta} \mathcal{M}$ satisfying:*

$$\varphi \circ \psi \xrightarrow{\eta} \mathrm{id}_{\mathcal{B}}$$

where $\delta = O(\varepsilon)$, $\eta = O(\varepsilon)$.

Proof. By the approximate invertibility of Definition A.15, for any ε-approximate surjective homomorphism $\varphi : \mathcal{M} \xrightarrow{\varepsilon} \mathcal{B}$, there exists a δ-approximate homomorphism $\psi : \mathcal{B} \xrightarrow{\delta} \mathcal{M}$ satisfying:

$$\varphi \circ \psi \xrightarrow{\eta} \mathrm{id}_{\mathcal{B}}$$

where $\delta = O(\varepsilon)$, $\eta = O(\varepsilon)$, directly from error transitivity. □

Lemma B.2 (Well-definedness of Connecting Homomorphism)**.** *Let $\{\mathbb{H}_n\}$ be a cognitive hierarchy sequence satisfying Axiom A.1, $\pi_n : \mathcal{M} \xrightarrow{\varepsilon_n'} \mathbb{H}_n$ and $\sigma_n : \mathbb{H}_n \xrightarrow{\varepsilon_n''} \mathcal{M}$ be bidirectional approximate homomorphism pairs, $\gamma_n : \mathbb{H}_n \xrightarrow{\varepsilon_n''''} \mathcal{W}_{fin,n}$ be the action approximate homomorphism defined by the axiom. Then the mapping $\mu_n : \mathbb{H}_n \xrightarrow{\zeta_n} \mathbb{H}_{n+1}$ defined as:*

$$\mu_n(x) = (\pi_x, \sigma_x, \gamma_x)$$

where:

- $\pi_x : \mathcal{M} \to \mathbb{H}_n$ *defined as* $\pi_x(m) = \pi_n(m)$ *(fixed perception mapping)*
- $\sigma_x : \mathbb{H}_n \to \mathbb{H}_n$ *defined as* $\sigma_x(y) = \pi_n(\sigma_n(y))$ *(self-reference mapping)*
- $\gamma_x : \mathbb{H}_n \to \mathcal{W}_{fin,n}$ *defined as* $\gamma_x(y) = \gamma_n(y)$ *(action mapping)*

is a well-defined ζ_n-approximate homomorphism, with $\zeta_n = O(\max(\varepsilon_n', \varepsilon_n'', \varepsilon_n''''))$.

Proof. Verification of well-definedness:

1. $\pi_x \in \{\pi : \mathcal{M} \xrightarrow{\varepsilon'} \mathbb{H}_n\}$: Directly guaranteed by the approximate homomorphism property of π_n.
2. $\sigma_x \in \{\sigma : \mathbb{H}_n \xrightarrow{\varepsilon''} \mathbb{H}_n\}$: $\pi_n \circ \sigma_n$ is an approximate homomorphism $\mathbb{H}_n \to \mathbb{H}_n$, satisfying operation preservation (error $\varepsilon_n' + \varepsilon_n''$).
3. $\gamma_x \in \{\gamma : \mathbb{H}_n \xrightarrow{\varepsilon'''} \mathcal{W}_{\text{fin},n}\}$: Guaranteed by the approximate homomorphism property of γ_n.

Verification of approximate homomorphism: For any $\omega \in \Omega$ and $x_1, ..., x_k \in \mathbb{H}_n$, consider operation preservation:

$$\mu_n(\omega(x_1, ..., x_k)) = (\pi_{\omega(x_1,...,x_k)}, \sigma_{\omega(x_1,...,x_k)}, \gamma_{\omega(x_1,...,x_k)})$$
$$\omega(\mu_n(x_1), ..., \mu_n(x_k)) = \omega((\pi_{x_1}, \sigma_{x_1}, \gamma_{x_1}), ..., (\pi_{x_k}, \sigma_{x_k}, \gamma_{x_k}))$$

By the definition of cognitive layer operational structure (Definition A.14), the right side equals $(\pi_\omega, \sigma_\omega, \gamma_\omega)$, where:

- $\pi_\omega(m) = \omega(\pi_{x_1}(m), ..., \pi_{x_k}(m)) = \omega(\pi_n(m), ..., \pi_n(m)) = \pi_n(\omega(m, ..., m))$
- $\sigma_\omega(y) = \omega(\sigma_{x_1}(y), ..., \sigma_{x_k}(y)) = \omega(\pi_n(\sigma_n(y)), ..., \pi_n(\sigma_n(y))) = \pi_n(\sigma_n(\omega(y, ..., y)))$
- $\gamma_\omega(y) = \omega(\gamma_{x_1}(y), ..., \gamma_{x_k}(y)) = \omega(\gamma_n(y), ..., \gamma_n(y)) = \gamma_n(\omega(y, ..., y))$

Comparing with the left side, errors are controlled by the approximate homomorphism errors of each component. Specifically, define $\zeta_n = L \cdot \max(\varepsilon'_n, \varepsilon''_n, \varepsilon'''_n)$, where L is the Lipschitz constant of operation ω (guaranteed by the Lipschitz condition of the cognitive pseudo-metric). By the error convergence of Axiom A.1, $\varepsilon'_n, \varepsilon''_n, \varepsilon'''_n \to 0$, thus $\zeta_n \to 0$. □

Lemma B.3 (Cognitive Hierarchies Form an Approximate Directed System)**.** *Let $\{\mathbb{H}_n\}_{n=0}^{\infty}$ be a cognitive hierarchy sequence satisfying Axiom A.1. Then there exist connecting approximate homomorphisms $\mu_n : \mathbb{H}_n \xrightarrow{\zeta_n} \mathbb{H}_{n+1}$ such that:*

1. *μ_n is an approximate homomorphism (not requiring strict homomorphism).*
2. *The limit $\varinjlim \mathbb{H}_n$ exists in the completion category $\widehat{\mathbf{AGI}}$.*
3. *The system satisfies approximate commutativity: $\mu_{n+1} \circ \mu_n \xrightarrow{\eta_n} \mu_n$, where $\eta_n \to 0$.*

Proof. From Axiom A.1 directly, μ_n are approximate homomorphisms. The composition relation is guaranteed by the definition of connecting homomorphisms and transitivity of approximate errors. Limit existence is guaranteed by category completeness. □

Definition B.2 (Three AGI Cognitive Characteristics)**.** *Let $\mathcal{M} \in \mathbf{AGI}$, $\mathcal{W}$ an infinite cognitive domain. Define the following three cognitive characteristics:*

1. ***Bidirectional Cognition Property (BCP)**: There exists a bidirectional construction $\{h_n = (\pi_n, \sigma_n)\}$ satisfying Definition B.1, where $\pi_n : \mathcal{M} \xrightarrow{\varepsilon'_n} \mathbb{H}_n$ are perception homomorphisms, $\sigma_n : \mathbb{H}_n \xrightarrow{\varepsilon''_n} \mathcal{M}$ are action homomorphisms, and they satisfy the bidirectional interaction consistency condition of Axiom A.1.*
2. ***Infinite Unidirectional Metacognition Property (IUMP)**: There exists a cognitive hierarchy sequence $\{\mathbb{H}_n\}$ satisfying Definition A.13 and a connecting approximate homomorphism sequence $\{\mu_n : \mathbb{H}_n \xrightarrow{\zeta_n} \mathbb{H}_{n+1}\}$, forming a unidirectional approximate directed system as in Lemma B.3, and satisfying the infinite metacognition condition of Axiom A.1.*
3. ***Limit Cognitive Isomorphism Property (LCIP)**: There exists a cognitive hierarchy sequence $\{\mathbb{H}_n\}$ such that the limit object $\mathbb{U} = \varinjlim \mathbb{H}_n$ and the target cognitive domain $\mathcal{W}$ satisfy the universal algebraic limit isomorphism relation $\mathbb{U} \cong \mathcal{W}$ of Definition A.5, i.e., the limit cognitive isomorphism condition of Axiom A.1.*

Lemma B.4 (Convergence Lemma for Unidirectional Connecting Homomorphisms)**.** *Let $\{\mathbb{H}_n\}$ satisfy the recursive construction and core constraints of Axiom A.1. Then the connecting approximate homomorphism sequence $\{\mu_n : \mathbb{H}_n \xrightarrow{\zeta_n} \mathbb{H}_{n+1}\}$ satisfies:*

1. *μ_n preserves τ-type operational structure, and $\{\zeta_n\}$ is a monotonically decreasing sequence with $\sum_{n=0}^{\infty} \zeta_n < \infty$;*

2. *Composition error satisfies* $\mu_{n+1} \circ \mu_n \xrightarrow{\zeta_n+\zeta_{n+1}} \mu_n$*;*

3. *The sequence* $\{\mathbb{H}_n, \mu_n\}$ *forms a unidirectional approximate directed system, whose inductive limit* $\mathbb{U} = \varinjlim \mathbb{H}_n$ *exists and is unique.*

Proof. 1. Operational structure preservation and error sequence properties:

By the cognitive symmetry of Definition A.15, μ_n as a connecting approximate homomorphism must preserve τ-type operational structure (enforced by the infinite metacognition condition of Axiom A.1). Combined with the error convergence constraint of Definition A.16, all approximate homomorphism error terms satisfy series convergence $\sum_{n=0}^{\infty}(\varepsilon_n^{\pi} + \varepsilon_n^{\sigma} + \zeta_n) < \infty$. Since the recursive construction of cognitive hierarchies is monotonic (higher-level cognitive error not greater than lower-level error, i.e., $\zeta_{n+1} \leq \zeta_n$), $\{\zeta_n\}$ is a monotonically decreasing sequence with convergent series.

2. Derivation of composition error:

Composition errors of approximate homomorphisms satisfy the triangle inequality. For $\mu_{n+1} \circ \mu_n$, its error is the superposition of μ_n's error ζ_n and μ_{n+1}'s error ζ_{n+1}. Since $\mu_n : \mathbb{H}_n \xrightarrow{\zeta_n} \mathbb{H}_{n+1}$, $\mu_{n+1} : \mathbb{H}_{n+1} \xrightarrow{\zeta_{n+1}} \mathbb{H}_{n+2}$, after composition, for any $\omega \in \Omega$ and $\vec{h} \in \mathbb{H}_n^{\tau(\omega)}$:

$$d(\omega(\mu_{n+1} \circ \mu_n(\vec{h})), \mu_{n+1} \circ \mu_n(\omega(\vec{h}))) \leq \zeta_n + \zeta_{n+1}$$

Moreover, by the compatibility constraint of cognitive hierarchies, $\mu_{n+1} \circ \mu_n$ and μ_n, though with different target domains, satisfy approximate equivalence when restricted to $\mathbb{H}_{n+1}$, thus $\mu_{n+1} \circ \mu_n \xrightarrow{\zeta_n+\zeta_{n+1}} \mu_n$.

3. Existence and uniqueness of inductive limit:

By the note in Definition A.10, $\mathbf{AGI}$ is a complete category, and its full subcategory $\mathbf{AGI}_0$ is closed under direct limit constructions. The sequence $\{\mathbb{H}_n, \mu_n\}$ satisfies the compatibility conditions of a directed system ($\mu_{n+1} \circ \mu_n \xrightarrow{\zeta_n+\zeta_{n+1}} \mu_n$ with convergent error series), thus the inductive limit $\mathbb{U} = \varinjlim \mathbb{H}_n$ exists in the completion category $\widehat{\mathbf{AGI}}$. Combined with the density preservation condition of Definition A.16, $\bigcup_{n=0}^{\infty} \Im(\varphi_n)$ is dense in $\mathbb{U}$, so the limit is unique up to isomorphism. □

Lemma B.5 (Unidirectional Construction Induces Bidirectional Construction)**.** *Let* $\{\mathbb{H}_n, \mu_n\}$ *be a unidirectional approximate directed system satisfying IUMP. Then there exists a bidirectional construction* $\{h_n = (\pi_n, \sigma_n)\}$ *satisfying BCP.*

Proof. 1. Construction of perception homomorphism π_n: By the recursive definition of cognitive hierarchy, $\mathbb{H}_0 = \mathcal{M}$; for $n \geq 0$, define $\pi_n = \mathrm{pr}_{\pi}(\mu_{n-1} \circ \cdots \circ \mu_0)$ (pr_{π} is the perception component projection). By Lemma B.4, $\pi_n : \mathcal{M} \xrightarrow{\varepsilon_n^{\pi}} \mathbb{H}_n$ with $\varepsilon_n^{\pi} = O(r^n)$.

2. Construction of action homomorphism σ_n: Based on the approximate invertibility of Definition A.15, for π_n there exists an approximate inverse $\sigma_n : \mathbb{H}_n \xrightarrow{\varepsilon_n^{\sigma}} \mathcal{M}$, where $\varepsilon_n^{\sigma} = O(\varepsilon_n^{\pi}) = O(r^n)$.

3. Verification of bidirectional construction conditions:

   - Idempotence: By the interaction idempotence of operational structure constraints, $\sigma_n \circ \pi_n \xrightarrow{\varepsilon_n^{\mathrm{id}}} (\sigma_n \circ \pi_n) \circ (\sigma_n \circ \pi_n)$, and $\varepsilon_n^{\mathrm{id}} \to 0$.
   - Error convergence: $\sum_{n=0}^{\infty}(\varepsilon_n^{\pi} + \varepsilon_n^{\sigma} + \varepsilon_n^{\mathrm{id}}) < \infty$, satisfying the convergence requirement of Definition B.1.
   - Compatibility: By commutativity of the directed system, $\pi_n \circ \sigma_{n-1} \xrightarrow{\delta_n} \pi_{n-1}$, where $\delta_n = O(r^n)$.

Thus, $\{h_n\}$ is a bidirectional construction satisfying BCP. □

Lemma B.6 (Bidirectional Construction Induces Limit Isomorphism)**.** *Let* $\{h_n\}$ *be a bidirectional construction satisfying BCP. Then the limit* $\mathbb{U} = \varinjlim \mathbb{H}_n$ *of the cognitive hierarchy sequence* $\{\mathbb{H}_n\}$ *satisfies LCIP (*$\mathbb{U} \cong \mathcal{W}$*).*

Proof. 1. Construction of approximate homomorphism sequence $\{\varphi_n\}$: Define $\varphi_n : \mathbb{H}_n \xrightarrow{\varepsilon_n} \mathcal{W}$ as $\varphi_n(x) = \gamma_n(x)$ (γ_n is the action component of $\mathbb{H}_n$). By error convergence of the bidirectional construction, $\varepsilon_n = O(r^n) \to 0$.

2. Verification of limit isomorphism conditions:

   - Density: By density preservation of cognitive hierarchy convergence constraints, $\bigcup_{n=0}^{\infty} \Im(\varphi_n)$ is dense in $\mathcal{W}$.
   - Composition error convergence: By approximate invertibility of operational structure constraints, there exist $\psi_n : \mathcal{W} \xrightarrow{\theta_n} \mathbb{H}_n$ such that $d(\varphi_n \circ \psi_n, \mathrm{id}_{\mathcal{W}}) \leq \varepsilon_n + \theta_n < \tau$ ($n > N$), and $d(\psi_n \circ \varphi_n, \mathrm{id}_{\mathbb{H}_n}) \to 0$.

3. By Definition A.5, $\mathbb{U} = \varinjlim \mathbb{H}_n \cong \mathcal{W}$, i.e., LCIP holds.

□

Theorem B.1 (Strict Equivalence Theorem for AGI's Three Cognitive Characteristics)**.** *Let $\mathcal{M} \in$* **AGI** *satisfy Axiom A.1. Then the following three cognitive characteristics are strictly equivalent:*

1. Bidirectional Cognition Property (BCP)

2. Infinite Unidirectional Metacognition Property (IUMP)

3. Limit Cognitive Isomorphism Property (LCIP)

I.e.: BCP $\iff$ IUMP $\iff$ LCIP

Proof. Adopt a circular proof logic chain, each step based on AGI Fundamental Axiom A.1 and constraints, ensuring no circular reasoning and unidirectional consistency:

Step 1: BCP ⇒ IUMP

Assume BCP holds, i.e., there exists a bidirectional construction $\{h_n = (\pi_n, \sigma_n)\}$ satisfying Definition B.1.

- Construct connecting homomorphisms $\mu_n : \mathbb{H}_n \xrightarrow{\zeta_n} \mathbb{H}_{n+1}$, defined as $\mu_n(x) = (\pi_x, \sigma_x, \gamma_x)$, where $\pi_x(m) = \pi_n(m)$, $\sigma_x(y) = \pi_n(\sigma_n(y))$, $\gamma_x(y) = \gamma_n(y)$;
- By Lemma B.4, $\{\mu_n\}$ preserves operational structure and $\{\zeta_n\}$ is a monotonically decreasing convergent sequence with $\sum_{n=0}^{\infty} \zeta_n < \infty$, forming a unidirectional approximate directed system;

This system satisfies the core definition of IUMP, thus IUMP holds.

Step 2: IUMP ⇒ LCIP

Assume IUMP holds, i.e., there exists a cognitive hierarchy sequence $\{\mathbb{H}_n\}$ and connecting homomorphisms $\{\mu_n\}$ forming a unidirectional approximate directed system.

- By Lemma B.4, the inductive limit $\mathbb{U} = \varinjlim \mathbb{H}_n$ of this directed system exists and is unique;
- Combined with the unidirectional limit constructivity of Definition A.16, there exists a uniform algorithm $\mathcal{C}$ extracting an approximate homomorphism sequence $\{\varphi_n\}$, satisfying $\mathbb{U} \cong \mathcal{W}$;

The limit isomorphism relation conforms to the definition of LCIP, thus LCIP holds.

Step 3: LCIP ⇒ BCP

Assume LCIP holds, i.e., $\mathbb{U} = \varinjlim \mathbb{H}_n \cong \mathcal{W}$.

- By the unidirectional limit constructivity of cognitive hierarchy convergence constraints, the connecting homomorphism sequence $\{\mu_n\}$ satisfies error series convergence $\sum_{n=0}^{\infty} \zeta_n < \infty$;

- By the limit isomorphism definition of LCIP, the cognitive hierarchy sequence $\{\mathbb{H}_n\}$ and connecting homomorphisms $\{\mu_n\}$ form a unidirectional approximate directed system satisfying IUMP;
- Based on the approximate invertibility and cognitive symmetry of Definition A.15, each μ_n induces perception-action homomorphism pairs (π_n, σ_n);

By Lemma B.5, these homomorphism pairs form a bidirectional construction $\{h_n\}$ satisfying Definition B.1, thus BCP holds.

In conclusion, the three cognitive characteristics are strictly equivalent under the axiomatic framework. □

Corollary B.1 (AGI Cognitive Unity)**.** *An AGI system possessing infinite cognitive capacity necessarily simultaneously satisfies the three characteristics of bidirectional cognition, infinite unidirectional metacognition, and limit cognitive isomorphism; these three characteristics capture the same essential cognitive structure from different mathematical perspectives.*

Remark B.1 (Significance of Cognitive Equivalence)**.** *The equivalence of the three cognitive characteristics reveals the deep unified structure of AGI cognitive capacity:*

- ***Bidirectional Cognition*** *reflects the action-perception closed-loop essence of cognition, guaranteed by invertibility and idempotence of operational structure constraints.*
- ***Infinite Unidirectional Metacognition*** *reflects the infinite recursive deepening capacity of cognition, supported by unidirectional constructivity of cognitive hierarchy convergence constraints.*
- ***Limit Cognitive Isomorphism*** *reflects the ultimate goal orientation of cognition, realized by error convergence and density preservation.*

This equivalence indicates that true general intelligence must possess these three dimensions simultaneously, providing clear theoretical guidance for AGI system design.

B.2 Cognitive Characteristic Equivalence Theorems

Definition B.3 (Core Equivalent Conditions for AGI Infinite Cognitive Capacity)**.** *Let $\mathcal{M} \in \mathbf{AGI}$. Define the following core cognitive characteristics:*

1. ***Infinite Cognitive Capacity (ICP)****: $\mathcal{M}$ satisfies Axiom A.1 and Definition A.17.*
2. ***Cognitive Completeness Object Property (CCO)****: $\mathcal{M}$ is a cognitive completeness object in category* $\mathbf{AGI}$ *(Definition A.19).*
3. ***Self-Referential Closure (SRC)****: $\mathrm{End}\mathcal{M}$ is closed under composition, and there exists an approximate fixed-point combinator $Y : \mathrm{End}\mathcal{M} \to \mathcal{M}$ (Definition B.4).*
4. ***Operational Functional Completeness (FCP)****: There exists a minimal irreducible operation set $\Omega_0 \subseteq \Omega$, whose generated function family $\langle \Omega_0 \rangle$ can approximate all basic cognitive functions (Definition A.3).*
5. ***Turing Degree Completeness (TDP)****: The cognitive encoding $\mathrm{enc}(:)\mathcal{T} \to \mathcal{M}$ induces Turing degree $0'$, covering the Turing degrees of all computable functions (Definition B.5).*
6. ***Universal Reducibility (UR)****: The cognitive encoding $(\mathrm{enc}(,)\mathrm{dec}())$ satisfies function composition compatibility and reduction closure (Definition B.6).*

Definition B.4 (Approximate Fixed-Point Combinator)**.** *Let $f \in \mathrm{End}\mathcal{M}$. Define the cycle kernel of f as $F_f = \{x \in \mathcal{M} \mid \exists k \in \mathbb{N}^+, f^k(x) = x\}$.*

Let $\mathcal{M} \in \mathbf{AGI}$. $Y : \mathrm{End}\mathcal{M} \to \mathcal{M}$ is called an approximate fixed-point combinator if it satisfies:

1. *For any $f \in \mathrm{End}\mathcal{M}$, $d(f(Y(f)), Y(f)) \leq \epsilon(f)$.*

2. $\epsilon(f) \to 0$ *as* f *tends to the identity map.*

3. *Isomorphism preservation: If* $\varphi : \mathcal{M} \cong \mathcal{N}$*, then* $Y_{\mathcal{N}}(\varphi \circ f \circ \varphi^{-1}) = \varphi(Y_{\mathcal{M}}(f))$*.*

4. *Cognitive stability:* Y *is continuous under the cognitive error metric.*

5. ***Structure-independent constructivity****: There exists a uniform construction algorithm* $\mathcal{A}_Y$ *such that for any* $\mathcal{M} \in \mathbf{AGI}$*,* $\mathcal{A}_Y$ *can derive* Y *from the operational generative completeness of* $\mathcal{M}$*, independent of the specific algebraic structure type of* $\mathcal{M}$*.*

Lemma B.7 (Structure-Independent Construction of Approximate Fixed-Point Combinator)**.** *Let* $\mathcal{M} \in \mathbf{AGI}$ *be a cognitive completeness object. Then there exists a structure-independent construction of an approximate fixed-point combinator:*

1. ***Cognitive hierarchy iteration method****: For any* $f \in \mathrm{End}\mathcal{M}$*, define iteration sequence* $x_0 = e_{\mathcal{M}}$ *(identity element),* $x_{n+1} = f(x_n)$*.*

2. ***Limit extraction****:* $Y(f) = \lim_{n\to\infty} x_n$ *(taking limit in the completion category* $\widehat{\mathbf{AGI}}$*).*

3. ***Error control****: By error series convergence of cognitive hierarchies (Definition A.16),* $\epsilon(f)$ *satisfies* $\sum_{n=0}^{\infty} \epsilon(f_n) < \infty$*, where* f_n *are the endomorphisms corresponding to the iteration sequence.*

This construction depends only on $\mathcal{M}$*'s operational generative completeness and cognitive hierarchy convergence, not assuming specific algebraic structure.*

Proof. 1. Well-definedness of iteration sequence: By $\mathcal{M}$'s operational generative completeness (Axiom A.1), the endomorphism monoid $\mathrm{End}\mathcal{M}$ is closed under composition, so for any $f \in \mathrm{End}\mathcal{M}$, the iteration sequence $x_n = f^n(e_{\mathcal{M}})$ exists and is unique in $\mathcal{M}$ (uniqueness guaranteed by irreducibility of generating set $S_{\mathcal{M}}$).

2. Existence of limit: In the completion category $\widehat{\mathbf{AGI}}$, limits of Cauchy sequences exist (Definition A.18). Verify $\{x_n\}$ is a Cauchy sequence: By the Lipschitz condition of the cognitive pseudo-metric (Definition A.1), there exists $L > 0$ such that $d(x_{n+1}, x_n) = d(f(x_n), f(x_{n-1})) \leq Ld(x_n, x_{n-1})$. Moreover, by cognitive hierarchy convergence constraints (Definition A.16), the error sequence $\{\epsilon_n = d(x_{n+1}, x_n)\}$ satisfies series convergence $\sum_{n=0}^{\infty} \epsilon_n < \infty$, thus $\{x_n\}$ is a Cauchy sequence, and limit $Y(f) = \lim_{n\to\infty} x_n$ exists.

3. Error control verification: Define $\epsilon(f_n) = d(f(x_n), x_n) = d(x_{n+1}, x_n)$, the error between consecutive terms of the iteration sequence. By error series convergence of cognitive hierarchies, $\sum_{n=0}^{\infty} \epsilon(f_n) < \infty$. When f tends to the identity map, $x_{n+1} = f(x_n) \to x_n$, thus $\epsilon(f_n) \to 0$, satisfying the error requirement of the approximate fixed-point combinator.

4. Proof of structure independence: The entire construction process depends only on two fundamental properties of $\mathcal{M}$: first, operational generative completeness (ensuring closure of iteration sequences); second, cognitive hierarchy convergence (ensuring limit existence and error series convergence). These two properties are universal properties of AGI systems (enforced by Axiom A.1), independent of the specific algebraic structure of $\mathcal{M}$ (such as Boolean algebra, semigroup, etc.), thus the construction is structure-independent.

□

Definition B.5 (Turing Degree Completeness of Cognitive Encoding)**.** *Let* $\mathcal{T}$ *denote the set of all Turing machines.*

The Turing degree $\deg(\mathrm{enc}(()\mathcal{T}))$ *induced by cognitive encoding* $\mathrm{enc}(:)\mathcal{T} \to \mathcal{M}$ *is* $0'$ *(the Turing degree of the halting problem), and the Turing degrees of all computable functions can be reduced to this degree via operations of the endomorphism monoid of* $\mathcal{M}$*.*

Theorem B.2 (Cognitive Completeness Object Equivalence Theorem)**.** $\mathcal{M}$ *is a cognitive completeness object (Definition A.19) if and only if* $\mathcal{M}$ *possesses infinite cognitive capacity (ICP) (Definition A.17).*

Proof. Adopt bidirectional proof logic, strictly based on definitions of cognitive completeness object and AGI fundamental axioms:

($\rightarrow$) If $\mathcal{M}$ is a cognitive completeness object (Definition A.19):

- Cognitive density: For any infinite cognitive domain $\mathcal{W}$, there exists an approximate homomorphism sequence $\{\varphi_n : \mathcal{M} \xrightarrow{\varepsilon_n} \mathcal{W}\}$, $\sum \varepsilon_n < \infty$, thus limit $\varinjlim \varphi_n$ is a limit isomorphism $\mathcal{M} \to \mathcal{W}$ (Definition A.5), satisfying the limit cognitive isomorphism condition of ICP.
- Cognitive projectivity: For any ε-approximate surjective homomorphism $f : \mathcal{A} \xrightarrow{\varepsilon} \mathcal{B}$, there exists $h : \mathcal{M} \xrightarrow{\eta} \mathcal{A}$ such that $f \circ h \xrightarrow{\zeta} g$, equivalent to operational generative completeness of ICP (can "fill" any cognitive pattern).
- Cognitive reflexivity: There exists $\eta : \mathcal{M} \xrightarrow{\varepsilon_0} \mathcal{M}^{**}$, i.e., $\mathcal{M}$ can perform metacognition on its own cognitive states, satisfying the infinite recursivity condition of ICP.

Thus, $\mathcal{M}$ possesses ICP.

($\leftarrow$) If $\mathcal{M}$ possesses ICP (Definition A.17):

- By operational generative completeness of ICP, for any $\mathcal{A} \in \mathbf{AGI}$, there exists an approximate homomorphism sequence $\{\varphi_n : \mathcal{M} \xrightarrow{\varepsilon_n} \mathcal{A}\}$, $\sum \varepsilon_n < \infty$, satisfying cognitive density.
- By infinite recursivity of ICP, the endomorphism monoid $\mathrm{End}\mathcal{M}$ of $\mathcal{M}$ is closed; for any $f : \mathcal{A} \xrightarrow{\varepsilon} \mathcal{B}$ and $g : \mathcal{M} \xrightarrow{\delta} \mathcal{B}$, we can construct $h : \mathcal{M} \xrightarrow{\eta} \mathcal{A}$ via metacognitive recursion, satisfying cognitive projectivity.
- By bidirectional interaction consistency of ICP, the perception-action closed loop of $\mathcal{M}$ induces $\eta : \mathcal{M} \xrightarrow{\varepsilon_0} \mathcal{M}^{**}$, satisfying cognitive reflexivity.

Thus, $\mathcal{M}$ is a cognitive completeness object. □

Definition B.6 (Universal Reducibility of Cognitive Encoding)**.** *Let* $\mathrm{enc}(:)\mathcal{T} \to \mathcal{M}$ *be a Turing machine encoding function,* $\mathrm{dec}(:)\mathcal{M} \to \mathcal{T}$ *a decoding function. If for any computable function* $f : \mathbb{N} \to \mathbb{N}$*, there exists a unique endomorphism* $\varphi_f \in \mathrm{End}\mathcal{M}$ *such that:*

1. $\mathrm{dec}(()\varphi_f(\mathrm{enc}(()T))) = f \circ T$ *for any Turing machine* T *(function composition compatibility).*
2. $\varphi_{f \circ g} = \varphi_f \circ \varphi_g$ *for any computable functions* f, g *(reduction closure).*

Then the cognitive encoding is said to satisfy ***universal reducibility****.*

Lemma B.8 (Endomorphic Representation of Basic Cognitive Functions)**.** *Let* $\mathcal{M} \in \mathbf{AGI}$*. Then all basic cognitive functions (Definition A.12) can be approximated by elements of* $\mathrm{End}\mathcal{M}$*.*

Proof. By operational generative completeness (Axiom A.1), there exists a minimal operation set Ω_0 generating all basic cognitive functions. For any basic cognitive function f, there exists a term t composed of operations from Ω_0 such that $f = t$. For term t, we can convert it into an endomorphism via Currying. Specifically, for a k-ary operation ω, we can define an endomorphism $\hat{\omega} : \mathcal{M} \to \mathcal{M}$ such that $\hat{\omega}(x) = \omega(x, x_2, \ldots, x_k)$, where $x_2, \ldots, x_k$ are fixed. By choosing different fixed parameters, we can generate all necessary functions. Since $\mathrm{End}\mathcal{M}$ is closed under composition, compositions of these endomorphisms remain in $\mathrm{End}\mathcal{M}$. Therefore, basic cognitive functions can be approximated by elements of $\mathrm{End}\mathcal{M}$. □

Theorem B.3 (AGI Infinite Cognitive Capacity Equivalence Theorem)**.** *Let* $\mathcal{M} \in \mathbf{AGI}$*. Then the following conditions are strictly equivalent:*

$$\textit{ICP} \iff \textit{CCO} \iff \textit{SRC} \iff \textit{FCP} \iff \textit{TDP} \iff \textit{UR}$$

Proof. Adopt a circular proof logic chain, each step strictly based on this paper's foundational axioms, core definitions, and standard universal algebra/category theory conclusions:

Step 1: ICP $\Longrightarrow$ CCO (Infinite Cognitive Capacity $\Longrightarrow$ Cognitive Completeness Object Property)

Proven by Theorem B.2.

Step 2: CCO $\Longrightarrow$ SRC (Cognitive Completeness Object Property $\Longrightarrow$ Self-Referential Closure)

By definition of cognitive completeness object (Definition A.19) and Theorem B.2, a cognitive completeness object $\mathcal{M}$ satisfies cognitive reflexivity, with canonical approximate homomorphism $\eta : \mathcal{M} \xrightarrow{\varepsilon_0} \mathcal{M}^{**}$. By cognitive density and cognitive projectivity, we can construct the closed structure of the endomorphism monoid $\mathrm{End}\mathcal{M}$.

For any $f \in \mathrm{End}\mathcal{M}$, by properties of cognitive completeness objects, we can construct an approximate fixed-point combinator $Y : \mathrm{End}\mathcal{M} \to \mathcal{M}$: define $Y(f)$ as the limit point of iteration sequence $x_0 = e_{\mathcal{M}}$ (cognitive baseline state), $x_{n+1} = f(x_n)$. By cognitive hierarchy convergence constraints (Definition A.16), this sequence converges, and the limit point satisfies $d(f(Y(f)), Y(f)) \leq \epsilon(f)$, where $\epsilon(f) \to 0$ as f tends to the identity map.

Verification of conditions in Definition B.4:

1. Approximate fixed-point property already proven.
2. Error convergence guaranteed by cognitive hierarchy error series convergence.
3. Isomorphism preservation guaranteed by categorical properties of cognitive completeness objects.
4. Cognitive stability guaranteed by Lipschitz continuity of approximate homomorphisms.
5. Structure-independent constructivity uniformly guaranteed by operational generative completeness.

Thus SRC holds.

Step 3: SRC $\Longrightarrow$ FCP

By SRC, $\mathrm{End}\mathcal{M}$ is closed under composition and an approximate fixed-point combinator exists. By Lemma B.8, basic cognitive functions can be approximated by $\mathrm{End}\mathcal{M}$. Consider the operation subset generating $\mathrm{End}\mathcal{M}$; by Zorn's lemma, there exists a minimal irreducible subset Ω_0 whose generated function family covers all basic cognitive functions, thus FCP holds.

Step 4: FCP $\Longrightarrow$ TDP (Operational Functional Completeness $\Longrightarrow$ Turing Degree Completeness)

By FCP, the function family generated by minimal operation set Ω_0 can approximate all computable functions. Construct cognitive encoding $\mathrm{enc}(:)\mathcal{T} \to \mathcal{M}$, mapping Turing machines to their operational representations in $\mathcal{M}$. Since Ω_0 is functionally complete, the halting problem can be expressed as an existential statement via operation combinations:

$$\mathrm{Halt}(T, x) \iff \exists n \,(\text{the halting decision operation generated by } \Omega_0 \text{ takes value true within } n \text{ steps})$$

This statement can be decided by the operational structure of $\mathcal{M}$, so the induced Turing degree of the encoding is $0'$, thus TDP holds.

Step 5: TDP $\Longrightarrow$ UR (Turing Degree Completeness $\Longrightarrow$ Universal Reducibility)

By TDP, cognitive encoding $\mathrm{enc}()$ covers Turing degrees of all computable functions. For any computable function f, there exists an endomorphism $\varphi_f \in \mathrm{End}\mathcal{M}$ simulating the operation of f. Verification of reduction conditions:

1) $\mathrm{dec}(()\varphi_f(\mathrm{enc}(()T))) = f \circ T$, guaranteed by consistency of encoding's operational representation.

2) $\varphi_{f\circ g} = \varphi_f \circ \varphi_g$, guaranteed by composition closure of the endomorphism monoid.

Thus UR holds.

Step 6: UR $\Longrightarrow$ ICP (Universal Reducibility $\Longrightarrow$ Infinite Cognitive Capacity)

By UR, cognitive encoding $(\mathrm{enc}(,)\mathrm{dec}())$ preserves computational semantics, and the endomorphism monoid is closed. Construct cognitive hierarchy sequence $\{\mathbb{H}_n\}$, where $\mathbb{H}_{n+1}$ is generated by composition of endomorphisms of $\mathbb{H}_n$, with connecting homomorphisms μ_n induced by universal reducibility. This sequence satisfies the recursive cognitive construction, bidirectional interaction, and limit isomorphism conditions of Axiom A.1: limit $\mathbb{U} = \varinjlim \mathbb{H}_n$ is limit isomorphic to infinite cognitive domain $\mathcal{W}$ (Definition A.5), thus ICP holds.

In conclusion, the six core conditions form a strict circular equivalence relation. □

Corollary B.2 (Uniqueness of Cognitive Completeness Objects)**.** *If $\mathcal{M}_1, \mathcal{M}_2 \in$ **AGI** are both cognitive completeness objects, then there exists an ε-approximate isomorphism $\varphi : \mathcal{M}_1 \xrightarrow{\varepsilon} \mathcal{M}_2$, and as $\varepsilon \to 0$, $\mathcal{M}_1$ and $\mathcal{M}_2$ are isomorphic in the limit sense.*

Proof. By Theorem B.3, cognitive completeness objects are equivalent to infinite cognitive capacity. By definition of infinite cognitive capacity, cognitive completeness objects must be able to approximate each other; constructing bidirectional approximate homomorphism sequences yields approximate isomorphism. □

Remark B.2 (Unification of AGI Core Structure)**.** *An AGI system possessing infinite cognitive capacity has its algebraic essence determined by the cognitive completeness object property; self-referential closure and operational functional completeness are its fundamental algebraic features, while Turing degree completeness and universal reducibility are inevitable manifestations at the computability level.*

C Minimal Algebraic Structures

C.1 Cognitively Complete Generative Structures

Definition C.1 (Minimal Complete Generative Structure)**.** *Let $\mathcal{M} \in \mathbf{AGI}$. $\mathcal{M}$ is called a* minimal complete generative structure *if it satisfies:*

1. ***Generative Completeness**: There exists a minimal generating set $S \subseteq \mathcal{M}$, $|S| = m$, such that $\mathrm{Gen}(S) = \mathcal{M}$, and removing any element from S results in incompleteness.*
2. ***Operation Minimality**: There exists a minimal operation set $\Omega_0 \subseteq \Omega$, satisfying:*
   - *Ω_0 is functionally complete (Definition A.3).*
   - *Removing any operation from Ω_0 results in functional incompleteness.*
   - *$|\Omega_0|$ is minimal among all functionally complete operation sets.*
3. ***Cognitive Completeness**: $\mathcal{M}$ is a cognitive completeness object (Definition A.19).*
4. ***Structural Minimality**: There exists no proper substructure $\mathcal{M}' \subsetneq \mathcal{M}$ satisfying the above three conditions.*

Definition C.2 (Cognitive Capacity)**.** *Let $\mathcal{M} \in \mathbf{AGI}$. Its cognitive capacity $C_{cog}(\mathcal{M})$ is defined as:*

$$C_{cog}(\mathcal{M}) = \max\{k \mid \exists k\text{-ary cognitive function family } \mathcal{F} \subseteq \langle\Omega\rangle, \mathcal{F} \text{ dense in } \mathcal{M}^k \to \mathcal{M}\}$$

i.e., the dimension of the maximum arity cognitive function family that $\mathcal{M}$ can stably express, reflecting the system's cognitive complexity.

Lemma C.1 (Properties of Minimal Complete Generative Structures)**.** *Let $\mathcal{M}$ be a minimal complete generative structure. Then:*

1. *The carrier cardinality $|\mathcal{M}|$ of $\mathcal{M}$ is minimal among all cognitive completeness objects.*
2. *The endomorphism monoid $\mathrm{End}\mathcal{M}$ of $\mathcal{M}$ has minimal cardinality.*
3. *The cognitive hierarchy construction of $\mathcal{M}$ has the simplest recursive form.*
4. *$\mathcal{M}$ is the minimal object in $\mathbf{AGI}$ satisfying Axioms A.1 and A.2.*

Proof. 1. If there exists a cognitive completeness object $\mathcal{M}'$ with smaller carrier cardinality, then by cognitive density of Definition A.19, $\mathcal{M}'$ can approximate $\mathcal{M}$, but with smaller carrier, contradicting structural minimality of $\mathcal{M}$.

2. Endomorphism monoid cardinality determined by generating set and operation set; minimality ensures minimal cardinality.

3. Simplest recursive form guaranteed by generative completeness and operation minimality.

4. Direct verification by Definition C.1 and axioms. □

C.2 Limitations of Lower-Order Structures

Proposition C.1 (Limitations of $\mathcal{B}_2$)**.** *The binary universal algebra $\mathcal{B}_2 = (\{0,1\}, \Omega)$ (Ω satisfying type τ) does not admit a universal approximate fixed-point combinator (i.e., a combinator that yields fixed points for all endomorphisms $f \in \mathrm{End}\mathcal{B}_2$).*

Proof. $\mathrm{End}\mathcal{B}_2$ contains 2 strict endomorphisms:

- id: cycle kernel $F_{\mathrm{id}} = \{0,1\}$ (all elements are fixed points)
- $\neg$: cycle kernel $F_{\neg} = \{0,1\}$ (2-cycle, no fixed points)

Since endomorphism $\neg$ has no fixed points, $\mathcal{B}_2$ cannot define a universal fixed-point combinator effective for all endomorphisms, violating the universality requirement of Definition B.4. □

Proposition C.2 (Impossibility of 3-Element Structures)**.** *There exists no 3-element universal algebra* $\mathcal{A} = (\{0, a, 1\}, \Omega)$ *satisfying the operational generative completeness of* **AGI**.

Proof. By operational generative completeness, a 'negation' function is required, i.e., a unary operation ω_1 must satisfy $\omega_1(\omega_1(x)) = x$ (inverse operation), whose orbits are 2-cycles or 1-fixed points; thus $|A|$ must be even. 3 is odd, cannot be partitioned, hence no 3-element structure satisfying operational generative completeness exists. □

Theorem C.1 ($\mathcal{B}_4$ as the Necessarily Minimal Complete Generative Structure)**.** *Let* $\mathcal{M} \in$ **AGI** *be a minimal complete generative structure (Definition C.1). Then:*

1. $\mathcal{M}$ *must be isomorphic to* $\mathcal{B}_4$*, i.e.,* $\mathcal{M} \cong \mathcal{B}_4$.
2. $|\mathcal{M}| = 4$ *is the minimal possible carrier cardinality for cognitive completeness.*
3. *The operation set* $\Omega = \{\neg, \wedge, \vee\}$ *of* $\mathcal{B}_4$ *is the minimal functionally complete operation set.*

Proof. **(1) Proof of structural isomorphism**: By definition of minimal complete generative structure, $\mathcal{M}$ is a cognitive completeness object. By Theorem B.2, cognitive completeness object is equivalent to infinite cognitive capacity. By operational generative completeness of Axiom A.1, $\mathcal{M}$ must support the following cognitive functions (explicit assumptions):

1. **Binary discrimination ability**: At least two basic cognitive states (true/false, yes/no).
2. **Metacognitive dimension**: Two additional states representing "degree of cognitive certainty" (know/uncertain), forming a metacognitive feedback loop.
3. **Operational closure**: Must support three basic cognitive operations: negation (cognitive correction), conjunction (consistency verification), disjunction (possibility coverage).

Consider lower bound of carrier cardinality:

- If $|\mathcal{M}| = 2$, can only represent binary states, cannot express metacognition (Proposition C.1).
- If $|\mathcal{M}| = 3$, cannot satisfy operational closure (Proposition C.2).
- Thus $|\mathcal{M}| \geq 4$.

Consider $|\mathcal{M}| = 4$. Let $\mathcal{M} = \{0, a, b, 1\}$, where 0 represents complete ignorance, 1 represents complete cognition, a, b represent two basic cognitive states and their metacognition.

By cognitive completeness requirement, must exist:

1. Unary negation operation $\neg$, satisfying $\neg\neg x = x$ (cognitive reversibility).
2. Binary conjunction operation $\wedge$, satisfying idempotence, commutativity, associativity (cognitive consistency).
3. Binary disjunction operation $\vee$, definable via $\neg$ and $\wedge$ (cognitive completeness).

Verify $\mathcal{M}$ must be a Boolean algebra:

- By existence of negation, each element has a unique complement.
- By associativity and distributivity of conjunction, forms a distributive lattice.
- Complement laws guaranteed by negation and conjunction/disjunction operations.

The four-element Boolean algebra is unique up to isomorphism, i.e., $\mathcal{M} \cong \mathcal{B}_4$.

(2) Proof of minimality of carrier cardinality:

If there exists a cognitive completeness object $\mathcal{M}'$ with $|\mathcal{M}'| < 4$, then:

- $|\mathcal{M}'| = 2$: By Proposition C.1, no universal approximate fixed-point combinator exists, violating self-referential closure.
- $|\mathcal{M}'| = 3$: By Proposition C.2, cannot satisfy operational generative completeness.

Thus $|\mathcal{M}| = 4$ is the minimum possible.

(3) Proof of operation minimality:

The operation set $\Omega = \{\neg, \wedge, \vee\}$ of $\mathcal{B}_4$ satisfies:

- Functional completeness: Can generate all Boolean functions, thereby approximating all cognitive functions.
- Minimality: Removing any operation results in functional incompleteness:
  - Remove $\neg$: cannot express negation cognition.
  - Remove $\wedge$: cannot express conjunction cognition.
  - Remove $\vee$: cannot express disjunction cognition.
- Minimal cardinality: No functionally complete set with fewer operations exists.

In conclusion, $\mathcal{B}_4$ is the unique minimal complete generative structure. □

Theorem C.2 (No Smaller Non-Boolean AGI Structure Exists)**.** *Let $\mathcal{A}$ be a non-Boolean universal algebra ($\mathcal{A} \notin \mathbf{Bool}$) with $|A| < 4$ (i.e., $|A| = 2$ or 3). Then $\mathcal{A}$ necessarily fails to satisfy AGI's core functional requirements (operational generative completeness, existence of fixed-point combinator, support for infinite metacognition).*

Proof. AGI's functional requirements are equivalent to three conditions:

1. Operational generative completeness: The operation set generates all basic cognitive functions.
2. Existence of fixed-point combinator.
3. Support for infinite metacognition: Endomorphism monoid closed, cognitive hierarchy recursion non-terminating.

Case 1: $|A| = 2$ **(binary non-Boolean algebra)** Take binary semigroup $\mathcal{S} = (\{0, 1\}, \cdot)$ as example (operation $\cdot$ defined as $x \cdot y = x$):

1. Operational generative completeness not satisfied: Only contains one binary operation, cannot generate negation-type cognitive functions (e.g., "correct erroneous cognition"), and cannot express perception-action bidirectional interaction (lacks inverse operation).
2. Fixed-point combinator nonexistent: Endomorphism set $\mathrm{End}\mathcal{S} = \{f(x) = x, g(x) = 0\}$, cycle kernel $F_g = \{0\}$, but f has no non-trivial fixed point, cannot define universal fixed-point combinator.
3. Infinite metacognition support not satisfied: Endomorphism monoid cardinality only 2, cognitive hierarchy $\mathbb{H}_1$ terminates (no new endomorphisms generatable).

Case 2: $|A| = 3$ **(ternary non-Boolean algebra)** Take ternary distributive lattice $\mathcal{L} = (\{0, a, 1\}, \vee, \wedge)$ as example (no complement operation):

1. Unary inverse operation cannot exist: Self-inverse unary operation requires $\omega_1(\omega_1(x)) = x$, its orbits are 2-cycles or 1-fixed points, but 3 is odd, cannot partition orbits (must exist element x such that $\omega_1(x) = x$, but no complementary element).

2. Operational generative completeness not satisfied: Lacks complement operation, cannot generate metacognitive functions like "knowing that one does not know".

3. Fixed-point combinator nonexistent: Endomorphism $f(x) = a$ (projection onto a) has cycle kernel $F_f = \{a\}$, but without complement operation cannot balance cognitive error, leading to unstable fixed points.

4. Infinite metacognition support not satisfied: Cognitive hierarchy $\mathbb{H}_2$ terminates due to lack of inverse operation preventing recursive generation of new cognitive layers.

All non-Boolean algebras with $|A| < 4$ are excluded due to failure to satisfy functional requirements. □

Remark C.1. *The minimal structure for AGI must be a Boolean algebra, and the minimal carrier cardinality is 4 (i.e., $\mathcal{B}_4$). The core defects of non-Boolean algebras are:*

1. *Lack of unary complement operation (negation/correction), cannot realize reflective function of metacognition.*

2. *When carrier cardinality is odd, cannot partition self-inverse operation orbits, causing cognitive hierarchy recursion to terminate.*

3. *Operation set functionally incomplete, cannot generate core functions required for perception-action bidirectional interaction.*

Therefore, non-Boolean algebras cannot serve as foundational structures for AGI; $\mathcal{B}_4$ is the unique minimal candidate.

C.3 Construction of $\mathcal{B}_4$ and Derivation of Boolean Properties

Definition C.3 (Four-Element Universal Algebra $\mathcal{B}_4$)**.** *$\mathcal{B}_4 = (\{0, a, b, 1\}, \Omega)$, denote $\omega_1 = \neg$, $\omega_2 = \wedge$, $\omega_3 = \vee$. Then $\Omega = \{\neg, \wedge, \vee\}$, operations defined as:*

- *$\neg x$ (unary complement): $\neg 0 = 1, \neg 1 = 0, \neg a = b, \neg b = a$.*
- *$x \wedge y$ (binary conjunction): Defined via truth table, satisfying:*

$\wedge$	*0*	*a*	*b*	*1*
0	*0*	*0*	*0*	*0*
a	*0*	*a*	*0*	*a*
b	*0*	*0*	*b*	*b*
1	*0*	*a*	*b*	*1*

- *$x \vee y$ (binary disjunction): $x \vee y = \neg(\neg x \wedge \neg y)$.*

Proposition C.3 (Derivation of Boolean Properties for $\mathcal{B}_4$)**.** *$\mathcal{B}_4$ satisfies all axioms of Boolean algebra (commutativity, associativity, distributivity, complement laws, boundedness), i.e., $\mathcal{B}_4 \in \mathbf{Bool}$.*

Proof. Verification of Boolean algebra axioms:

1. **Commutativity**: Directly from symmetry of $\min$ and $\max$ operations.

2. **Associativity**: For any $x, y, z \in \mathcal{B}_4$, $\omega_2(\omega_2(x, y), z) = \min(\min(x, y), z) = \min(x, \min(y, z)) = \omega_2(x, \omega_2(y, z))$.

3. **Crucial verification of distributivity**:

- (a, a, b): $a \wedge (a \vee b) = a \wedge 1 = a$, $(a \wedge a) \vee (a \wedge b) = a \vee 0 = a$

- (a, b, a): $a \wedge (b \vee a) = a \wedge 1 = a$, $(a \wedge b) \vee (a \wedge a) = 0 \vee a = a$
- (b, a, b): $b \wedge (a \vee b) = b \wedge 1 = b$, $(b \wedge a) \vee (b \wedge b) = 0 \vee b = b$

Other cases similarly verified.

4. **Complement laws**: For any $x \in \mathcal{B}_4$, $\omega_2(x, \omega_1(x)) = \min(x, \omega_1(x)) = 0$ (since x and $\omega_1(x)$ always have one 0) $\omega_3(x, \omega_1(x)) = \max(x, \omega_1(x)) = 1$ (since x and $\omega_1(x)$ always have one 1)

5. **Boundedness**: 0 and 1 are lower and upper bounds respectively.

Thus $\mathcal{B}_4$ is a Boolean algebra. □

Theorem C.3 (Existence of Approximate Fixed-Point Combinator for $\mathcal{B}_4$)**.** *$\mathcal{B}_4$ admits an approximate fixed-point combinator $Y_{\mathcal{B}_4}$ satisfying all conditions of Definition B.4.*

Proof. For any $f \in \mathrm{End}\mathcal{B}_4$, by the periodic point theorem for self-maps on finite sets (trivial fact: any self-map on a finite set has periodic points, i.e., exists $k \in \mathbb{N}^+$ such that $f^k(x) = x$), there exists a minimal k such that f^k has a fixed point.

Define $Y_{\mathcal{B}_4}(f)$ as the minimal fixed point of f^k (according to partial order $0 \leq a, b \leq 1$ of $\mathcal{B}_4$). Verification of conditions:

1. Approximate fixed-point property: $d(f(Y_{\mathcal{B}_4}(f)), Y_{\mathcal{B}_4}(f)) = d(f^{k+1}(x), f^k(x)) = 0 \leq \epsilon(f)$ (since x is fixed point of f^k), $\epsilon(f) = 0$ for all f.
2. Error convergence: When f tends to identity map, $k = 1$, $\epsilon(f) = 0 \to 0$.
3. Isomorphism preservation: Boolean algebra isomorphisms $\varphi : \mathcal{B}_4 \cong \mathcal{B}_4$ (only identity and complement), $Y_{\mathcal{B}_4}(\varphi \circ f \circ \varphi^{-1}) = \varphi(Y_{\mathcal{B}_4}(f))$.
4. Cognitive stability: $\mathcal{B}_4$ finite, cognitive pseudo-metric discrete metric, thus $Y_{\mathcal{B}_4}$ continuous.

In conclusion, $Y_{\mathcal{B}_4}$ satisfies all conditions. □

Theorem C.4 (Irreplaceability of Operations in $\mathcal{B}_4$)**.** *In the operation set $\Omega = \{\omega_1, \omega_2, \omega_3\}$ of $\mathcal{B}_4$, each operation is irreplaceable—removing any yields a universal algebra with at least one of the following defects:*

1. *Fails operational generative completeness.*
2. *Lacks universal fixed-point combinator.*
3. *Ceases to be a cognitive completeness object of* **AGI**.

Proof. Verify three cases:

Case 1: Remove negation $\neg$ Resulting algebra $\mathcal{A} = (A, \{\wedge, \vee\})$ is a distributive lattice. Consider endomorphism $f : A \to A$ defined as $f(x) = x \wedge a$ (fixed $a \in A$). If $a \neq 1$, then f has no fixed point (since $f(x) \leq a < 1$, unless $x \leq a$). More critically, lacking negation, cannot express "not" concept, violating operational generative completeness. E.g., basic cognitive function "correct erroneous cognition" requires negation.

Case 2: Remove conjunction $\wedge$ Resulting algebra $\mathcal{A} = (A, \{\neg, \vee\})$. Since conjunction not definable from $\{\neg, \vee\}$ (in Boolean algebra, conjunction is dual of disjunction, but without conjunction cannot express logical AND), functionally incomplete. E.g., cognitive consistency verification requires conjunction.

Case 3: Remove disjunction $\vee$ Similar to case 2, functionally incomplete. Cognitive possibility coverage requires disjunction.

Thus each operation is irreplaceable. □

Corollary C.1. *The core operation set of AGI is "minimal and indispensable": unary inverse operation (negation/correction), binary conjunction (consistency verification), binary disjunction (possibility coverage).*

Corollary C.2 ($\mathcal{B}_4$ as a Cognitive Completeness Object in **AGI**)**.** *$\mathcal{B}_4$ is a cognitive completeness object in* **AGI** *(Definition A.19).*

Proof. By Theorem C.1, $\mathcal{B}_4$ is a minimal complete generative structure, hence a cognitive completeness object (Definition C.1). □

Corollary C.3 (Minimality of $\mathcal{B}_4$)**.** *$\mathcal{B}_4$ is the minimal cognitive completeness object in* **AGI** *satisfying Axioms A.1 and A.2, i.e.:*

1. *For any cognitive completeness object $\mathcal{M} \in$ **AGI**, there exists an ε-approximate surjective homomorphism $\varphi : \mathcal{B}_4 \xrightarrow{\varepsilon} \mathcal{M}$.*
2. *$|\mathcal{B}_4| = 4$ is the minimal possible cardinality for a cognitive completeness object.*
3. *The endomorphism monoid $\mathrm{End}\mathcal{B}_4$ of $\mathcal{B}_4$ has minimal cardinality 6.*

Proof.
1. By cognitive completeness of $\mathcal{B}_4$ (Theorem C.2) and cognitive density, for any cognitive completeness object $\mathcal{M}$, there exists an approximate homomorphism sequence $\{\varphi_n\}$ whose limit is surjective, thus there exists an ε-approximate surjective homomorphism.
2. By Theorem C.1, $|\mathcal{B}_4| = 4$ is minimal possible.
3. Cardinality of $\mathrm{End}\mathcal{B}_4$ is 6 by direct computation, and by minimality of $\mathcal{B}_4$, any cognitive completeness object's endomorphism monoid has cardinality at least 6.

□

Proposition C.4 (Limitations of Non-Boolean Universal Algebras)**.** *Unary inverse operation is necessary for operational generative completeness (Definition A.3); non-Boolean algebras like distributive lattices lack this operation. Therefore, Non-Boolean universal algebras (e.g., distributive lattices, semigroups) cannot satisfy operational generative completeness, due to lacking necessary operations to generate all basic cognitive functions.*

Proof. Distributive lattices lack unary inverse operation, cannot express negative cognition; semigroups have only one binary operation, cannot generate multi-ary decision functions. By operational generative completeness requirement, such structures are excluded. □

C.4 Algebraic and Topological-Geometric Structure of Extended Systems

Based on the AGI theoretical framework established earlier, particularly:

- Theorem C.2 (cognitive completeness object property of $\mathcal{B}_4$)
- Theorem C.2 (minimality of $\mathcal{B}_4$)
- Theorem B.3 (equivalence of infinite cognition)

Any AGI system must contain the core structure of $\mathcal{B}_4$. Considering extended systems with carrier set cardinality $|M| > 4$, we introduce the following minimality assumption:

Assumption C.1 (Extended System Minimality Assumption)**.** *An extended AGI system $\mathcal{M}_{ext} = (M_{ext}, \Omega_{ext})$ satisfies:*

1. ***Carrier constraint**: $|M_{ext}| = N > 4$, N finite and $N = 4^k$ ($k \in \mathbb{N}^+$).*
2. ***Core embedding**: There exists a monomorphism $\iota : \mathcal{B}_4 \hookrightarrow \mathcal{M}_{ext}$ (strict monomorphism), and $\iota(\mathcal{B}_4)$ is a minimal complete generative substructure of $\mathcal{M}_{ext}$.*
3. ***Boolean algebra**: $\mathcal{M}_{ext}$ is a Boolean algebra, operation set $\Omega_{ext} = \{\vee, \wedge, \neg, 0, 1\}$, and ι is a Boolean algebra monomorphism.*

4. ***No redundancy**: There exists no proper Boolean subalgebra $\mathcal{M}' \subsetneq \mathcal{M}_{ext}$ such that $\mathcal{B}_4 \hookrightarrow \mathcal{M}'$ and $\mathcal{M}'$ satisfies conditions 1-3 (i.e., $\mathcal{M}_{ext}$ is the minimal extension Boolean algebra containing the $\mathcal{B}_4$ core).*

5. ***Operation minimality**: Ω_{ext} is a minimal functionally complete operation set.*

Theorem C.5 (Necessity of Boolean Structure)**.** *Let $\mathcal{M}_{ext}$ be an extended system satisfying AGI axioms ($|M| > 4$). Then $\mathcal{M}_{ext}$ must be a complete atomic Boolean algebra, and there exists a non-empty set X such that:*

$$\mathcal{M}_{ext} \cong (\mathcal{P}(X); \cup, \cap, \complement)$$

This structure is the only possible algebraic structure satisfying cognitive completeness.

Proof. **Step 1: Necessity of Boolean operation set** By cognitive completeness requirement (Definition A.10), $\mathcal{M}_{\text{ext}}$ must generate all $m \leq 3$-ary cognitive functions, requiring the operation set to include unary inverse (negation), binary conjunction (consistency verification), binary disjunction (possibility coverage) (Theorem C.4).

Step 2: Derivation of Boolean algebra axioms

- Commutativity: Cognitive results independent of operation order, thus $\cup, \cap$ commutative.
- Associativity: Stepwise construction of complex cognition requires associativity, thus $\cup, \cap$ associative.
- Distributivity: Coordination of cognitive reasoning requires distributivity, i.e., $x \cap (y \cup z) = (x \cap y) \cup (x \cap z)$.
- Complement laws: Reflective function of metacognition requires complement, thus for any $x \in \mathcal{M}_{\text{ext}}$, exists unique $\complement x$ such that $x \cap \complement x = \emptyset$ and $x \cup \complement x = X$.
- Boundedness: Cognitive baseline requires minimal element $\emptyset$ (no cognitive information) and maximal element X (full cognitive information).

Step 3: Proof of complete atomicity

- Completeness: For any subset $S \subseteq \mathcal{M}_{\text{ext}}$, supremum $\bigvee S$ (cognitive integration result) and infimum $\bigwedge S$ (cognitive intersection) exist, satisfying cognitive completeness.
- Atomicity: By contradiction, if an infinite descending chain $a_1 > a_2 > \dots$ exists, contradicts cognitive finite decomposability (Axiom A.2), thus minimal atomic elements exist.

Step 4: Proof of uniqueness Assume a non-Boolean structure $\mathcal{M}'$ satisfies AGI requirements. Then $\mathcal{M}'$ necessarily lacks key operations above (e.g., non-distributive lattices lack distributivity), cannot generate all cognitive functions, violating cognitive completeness; thus Boolean structure is the only possibility. □

Lemma C.2 (Subalgebra Decomposition of Finite Boolean Algebras)**.** *Let $\mathcal{B}$ be a finite Boolean algebra, $\mathcal{A} \subseteq \mathcal{B}$ a subalgebra. Then there exists positive integer $k \leq \dim(\mathcal{B})$ such that $\mathcal{A} \cong \mathcal{B}_4^k$, and there exists a complementary subalgebra $\mathcal{C} \subseteq \mathcal{B}$ such that $\mathcal{B} \cong \mathcal{A} \times \mathcal{C}$.*

Proof. By finite Boolean algebra classification theorem, $\mathcal{B} \cong \mathcal{B}_4^n$ ($n = \dim(\mathcal{B})$). $\mathcal{A}$ as a Boolean subalgebra is generated by its atoms. Suppose $\mathcal{A}$ has k atoms, then $\mathcal{A} \cong \mathcal{B}_4^k$. Take $\mathcal{C}$ as the orthogonal complement of $\mathcal{A}$ in $\mathcal{B}$, i.e., generated by atoms in $\mathcal{B}$ orthogonal to all elements of $\mathcal{A}$, then $\mathcal{B} \cong \mathcal{A} \times \mathcal{C}$. □

Theorem C.6 (Algebraic Decomposition of Extended Systems)**.** *Let $\mathcal{M}_{ext}$ satisfy Assumption C.1. Then there exists a unique positive integer $k = \log_4 N$ such that:*

$$\mathcal{M}_{ext} \cong \mathcal{B}_4^k \quad (\textit{Boolean algebra isomorphism})$$

and this decomposition satisfies:

1. ***Functional orthogonality**: For $i \neq j$, $\mathrm{Hom}(B_{4i}, B_{4j})$ contains only constant homomorphisms (zero or unit homomorphisms).*

2. ***No redundancy**: No proper subalgebra decomposition $\mathcal{M}_{ext} \cong \mathcal{B}_4^{k_1} \times \mathcal{A}$ exists ($k_1 < k$, $\mathcal{A}$ non-trivial Boolean algebra).*

3. ***Cognitive equivalence**: $\mathcal{M}_{ext}$ and $\mathcal{B}_4^k$ have identical cognitive capacity (in strict isomorphism sense).*

4. ***Complexity optimality**: $|\mathrm{End}\mathcal{M}_{ext}| = 6^k$ is the minimal possible satisfying cognitive requirements.*

Proof. Since $\mathcal{M}_{\text{ext}}$ is a finite Boolean algebra, by finite Boolean algebra classification theorem, exists $m \in \mathbb{N}^+$ such that $\mathcal{M}_{\text{ext}} \cong \mathcal{B}_2^m$.

By core embedding condition, $\mathcal{B}_4 \cong \mathcal{B}_2^2$ as minimal complete generative structure embeds into $\mathcal{M}_{\text{ext}}$. This requires the atomic structure of $\mathcal{M}_{\text{ext}}$ to include atoms $\{a, b\}$ of $\mathcal{B}_4$ and their orthogonal complement.

Suppose $\mathcal{M}_{\text{ext}}$ has k orthogonal $\mathcal{B}_4$ substructures, then $\mathcal{M}_{\text{ext}} \cong \mathcal{B}_4^k$. By carrier cardinality $|M_{\text{ext}}| = 4^k$, unique decomposition obtained.

Functional orthogonality guaranteed by no redundancy assumption: if cross-component homomorphisms existed, a smaller subalgebra could be constructed, contradiction.

No redundancy directly from Assumption C.1.

Cognitive equivalence by isomorphism: $\mathcal{B}_4^k$ inherits cognitive completeness of $\mathcal{B}_4$, and direct product structure preserves cognitive capacity.

Complexity optimality: $\mathcal{B}_4$ is minimal complete generative structure; its extension $\mathcal{B}_4^k$ has minimal complexity while preserving cognitive completeness. □

Theorem C.7 (Structure of Extended System Endomorphism Monoid)**.** *Let extended system $\mathcal{M}_{ext} \cong \mathcal{B}_4^k$ satisfy the no redundancy condition of Assumption C.1. Then its endomorphism monoid $\mathrm{End}\mathcal{M}_{ext}$ satisfies:*

1. ***Orthogonal decomposition**:*
$$\mathrm{End}\mathcal{M}_{ext} \cong \mathrm{End}\mathcal{B}_4{}^k$$
i.e., any $f \in \mathrm{End}\mathcal{M}_{ext}$ uniquely decomposes as $f = (f_1, \ldots, f_k)$, where $f_i \in \mathrm{End}\mathcal{B}_4$.

2. ***Cardinality estimate**: $|\mathrm{End}\mathcal{M}_{ext}| = 6^k$ (where $|\mathrm{End}\mathcal{B}_4| = 6$, directly enumerable by Boolean properties of $\mathcal{B}_4$).*

3. ***Composition operation**: For any $f, g \in \mathrm{End}\mathcal{M}_{ext}$,*
$$f \circ g = (f_1 \circ g_1, \ldots, f_k \circ g_k)$$

Proof. By no redundancy of Assumption C.1, endomorphisms of $\mathcal{M}_{\text{ext}} = \mathcal{B}_4^k$ cannot map across components (otherwise generate redundant subalgebras), thus endomorphism monoid decomposes as direct product of component endomorphism monoids. Specifically, for any endomorphism $f : \mathcal{B}_4^k \to \mathcal{B}_4^k$, define $f_i = \mathrm{pr}_i \circ f \circ \mathrm{in}_i$, where pr_i projects to i-th component, in_i embeds into i-th component. Since f is a Boolean algebra homomorphism and satisfies no redundancy constraint, it must preserve component orthogonality; thus f determined by its action on each component, and components independent. Therefore, f corresponds to k endomorphisms $f_i : \mathcal{B}_4 \to \mathcal{B}_4$, and vice versa, establishing isomorphism $\mathrm{End}\mathcal{M}_{\text{ext}} \cong \mathrm{End}\mathcal{B}_4{}^k$.

By Boolean structure of $\mathcal{B}_4$, its endomorphisms enumerable as identity, zero, unit, and three projection-like endomorphisms, total 6, so $|\mathrm{End}\mathcal{B}_4| = 6$, hence $|\mathrm{End}\mathcal{M}_{\text{ext}}| = 6^k$. Decomposition of composition directly from universal properties of direct product structure. □

Corollary C.4 (Endomorphism Monoid Structure)**.** *Let $\mathcal{M}_{ext} \cong \mathcal{B}_4^k$. Then:*
$$\mathrm{End}\mathcal{M}_{ext} \cong \mathrm{End}\mathcal{B}_4{}^k \quad \textit{and} \quad |\mathrm{End}\mathcal{M}_{ext}| = 6^k$$
Any $f \in \mathrm{End}\mathcal{M}_{ext}$ uniquely decomposes as $f = (f_1, \ldots, f_k)$, where $f_i \in \mathrm{End}\mathcal{B}_4$.

Corollary C.5 (Extended System Approximate Fixed-Point Combinator)**.** *There exists a uniformly constructed approximate fixed-point combinator $Y_{ext} : \mathrm{End}\mathcal{M}_{ext} \xrightarrow{\varepsilon} \mathcal{M}_{ext}$ defined as:*

$$Y_{ext}(f) = \psi\left(Y_{\mathcal{B}_4}(f_1), \ldots, Y_{\mathcal{B}_4}(f_k)\right)$$

where:

- $f \xrightarrow{\varepsilon} (f_1, \ldots, f_k)$ *is the decomposition from Theorem C.7.*
- $Y_{\mathcal{B}_4}$ *is the approximate fixed-point combinator of $\mathcal{B}_4$ (Definition B.4).*
- $\psi : \mathcal{B}_4^k \xrightarrow{\delta} \mathcal{M}_{ext}$ *is the inverse component of approximate isomorphism.*
- *Total error satisfies $\varepsilon_{total} = O(\max(\varepsilon, \delta))$.*

This combinator preserves cognitive stability and isomorphism preservation (conditions 3-4 of Definition B.4).

Theorem C.8 (Cognitive Hierarchy Decomposition)**.** *Let $\mathcal{M}_{ext} \cong \mathcal{B}_4^k$. Then its cognitive hierarchy sequence $\{\mathbb{H}_{ext,n}\}$ satisfies approximate decomposition:*

$$\mathbb{H}_{ext,n} \cong_{\lim} \mathbb{H}^k_{\mathcal{B}_4,n}$$

and limit object $\mathbb{U}_{ext} = \varinjlim \mathbb{H}_{ext,n} \cong_{\lim} \mathbb{U}^k_{\mathcal{B}_4}$.

Proof. By recursive definition of cognitive hierarchy (Definition A.13) and universal properties of Boolean algebra direct product, induction yields. Specifically, n-th level cognitive hierarchy $\mathbb{H}_{\text{ext},n}$ consists of approximate endomorphisms of $\mathcal{M}_{\text{ext}}$; the direct product decomposition of $\mathcal{M}_{\text{ext}}$ induces decomposition of these endomorphisms, yielding hierarchy decomposition. □

Corollary C.6 (Error Convergence)**.** *The cognitive error sequence $\{\varepsilon_n\}$ of extended system satisfies recurrence inequality:*

$$\varepsilon_{n+1} \leq \psi(\varepsilon_n) + \delta_n$$

where $\psi(x) = 6^k x^2 + o(x^2)$, δ_n are higher-order error terms with $\sum_{n=0}^{\infty} \delta_n < \infty$. When initial error ε_0 sufficiently small, sequence $\{\varepsilon_n\}$ converges to 0, convergence rate determined by leading term of ψ.

Theorem C.9 (Metacognitive Depth Theorem)**.** *Let $\mathcal{M}_{ext} \cong \mathcal{B}_4^k$. Then its maximum metacognitive hierarchy depth $d_{\max}$ is:*

$$d_{\max} = \lfloor \log_2(2k+1) \rfloor = \Theta(\log k)$$

i.e., metacognitive capacity grows logarithmically with number of modules.

Proof. n-th level metacognition needs to handle 2^n independent cognitive propositions; by cognitive capacity $C_{\text{cog}} = 2k$ of extended system (can distinguish 2^{2k} cognitive states), we have $2^n \leq 2^{2k}$, i.e., $n \leq 2k$. Considering nesting redundancy of metacognition, actual maximum depth satisfies $2^d \leq 2k+1$, thus $d_{\max} = \lfloor \log_2(2k+1) \rfloor = \Theta(\log k)$, i.e., metacognitive capacity grows logarithmically with number of modules. □

C.5 Topological and Geometric Structure

Based on algebraic decomposition theorem $\mathcal{M}_{\text{ext}} \cong \mathcal{B}_4^k$, we derive inevitable topological-geometric structure.

Theorem C.10 (Stone Space Approximate Dual Structure)**.** *Let $\mathcal{M}_{ext} \cong_{\lim} \mathcal{B}_4^k$. Then its Stone space $S(\mathcal{M}_{ext})$ (Stone space of finite Boolean algebra is discrete space on atoms) satisfies:*

1. ***Approximate product structure**: $S(\mathcal{M}_{ext}) \cong_{hom} S(\mathcal{B}_4)^k \cong$ (2-point discrete space)^k (homeomorphic), where $S(\mathcal{B}_4)$ is Stone space of $\mathcal{B}_4$ (2-point discrete space corresponding to its atoms $\{a, b\}$).*
2. ***Topological properties**: Compact Hausdorff, totally disconnected, zero-dimensional, and extremely disconnected in ε-approximate sense.*

3. ***Approximate duality**: There exists ε-approximate isomorphism $\mathcal{M}_{ext} \xrightarrow{\varepsilon} \mathcal{C}\updownarrow\wr_{\sqrt{}}(S(\mathcal{M}_{ext}))$, where $\varepsilon = O(1/k^2)$ (derived from ordinary convergence of cognitive hierarchies, Definition A.16).*

4. ***Cognitive interpretation**: Each point of Stone space corresponds to a maximal ε-consistent cognitive state; clopen sets correspond to decidable cognitive propositions.*

Proof. 1. By finite version of Stone representation theorem, Stone space $S(\mathcal{B})$ of finite Boolean algebra $\mathcal{B}$ homeomorphic to $\{0,1\}^{\dim(\mathcal{B})}$. Since $\mathcal{M}_{\text{ext}} \cong_{\lim} \mathcal{B}_4^k$, $\dim(\mathcal{M}_{\text{ext}}) = 2k$.

2. Stone space of finite Boolean algebra naturally has these properties. Extreme disconnectedness in approximate sense means: for any clopen sets U, V, if $d(U,V) < \varepsilon$, then exists clopen W such that $U \subseteq W \subseteq \overline{V}$.

3. Construct approximate isomorphism $\varphi : \mathcal{M}_{\text{ext}} \to \mathcal{C}\updownarrow\wr_{\sqrt{}}(S(\mathcal{M}_{\text{ext}}))$ as:

$$\varphi(a) = \{p \in S(\mathcal{M}_{\text{ext}}) : p(a) = 1\}$$

By approximate algebraic structure, φ is ε-approximate homomorphism, error from approximate Boolean operations of $\mathcal{M}_{\text{ext}}$, combined with series convergence condition of cognitive hierarchy convergence constraints (Definition A.16), take $\varepsilon = O(1/k^2)$ satisfying $\sum_{k=1}^{\infty} \varepsilon < \infty$.

4. Cognitive interpretation guaranteed by naturalness of Stone duality.

□

Corollary C.7 (Topological Dimension)**.** *Topological dimension of $S(\mathcal{M}_{ext})$ is 0, but embedding dimension is k (requires k coordinates to distinguish different concepts).*

Definition C.4 (Cognitive Embedding)**.** *Define an embedding map $\varphi : \mathcal{M}_{ext} \hookrightarrow [0,1]^{2k}$ as:*

$$\varphi(x_1, \ldots, x_k) = (\psi(x_1), \ldots, \psi(x_k))$$

where $\psi : \mathcal{B}_4 \to [0,1]^2$ is $\psi(0) = (0,0)$, $\psi(a) = (0,1)$, $\psi(b) = (1,0)$, $\psi(1) = (1,1)$. This embedding preserves the Boolean operations of $\mathcal{B}_4$ (disjunction corresponds to coordinate-wise maximum, conjunction to coordinate-wise minimum, and negation to coordinate-wise flipping).

Lemma C.3 (Metric Properties of Approximate Embedding)**.** *The embedding map $\varphi : \mathcal{M}_{ext} \to [0,1]^{2k}$ from Definition C.4 satisfies:*

1. ***Approximate Isometry**: For all $x, y \in \mathcal{M}_{ext}$,*

$$|d_{cog}(x,y) - \|\varphi(x) - \varphi(y)\|_2| \leq \varepsilon_{emb}$$

where $\varepsilon_{emb} = O(k^{-1/2})$.

2. ***Operation Approximate Preservation**: For any Boolean operation ω,*

$$\|\varphi(\omega(x,y)) - \omega_{\mathbb{R}}(\varphi(x), \varphi(y))\| \leq \varepsilon_{op}$$

where $\omega_{\mathbb{R}}$ is the piecewise linear realization of ω on $\mathbb{R}^{2k}$.

3. ***Density**: $\varphi(\mathcal{M}_{ext})$ is ε-dense in $[0,1]^{2k}$, i.e., for any $p \in [0,1]^{2k}$, there exists $x \in \mathcal{M}_{ext}$ such that $\|\varphi(x) - p\| \leq \varepsilon$.*

Theorem C.11 (Cognitive Manifold Structure)**.** *Under the embedding φ, the image set $\mathcal{X} = \varphi(\mathcal{M}_{ext})$ forms a discrete manifold:*

1. ***Metric Structure**: It inherits the Euclidean metric from $[0,1]^{2k}$, which restricts to a discrete metric on $\mathcal{X}$.*

2. ***Geometric Realization of Operations**: Boolean operations are realized as piecewise linear mappings on $\mathcal{X}$.*

3. ***Cognitive Distance**: The cognitive distance between any two points $x, y \in \mathcal{X}$ is $d_{cog}(x, y) = \frac{1}{2k}\|x - y\|_1$ (the normalized Hamming distance).*

Proof. $\mathcal{X}$ is a finite set of points. Endowed with the discrete topology, the neighborhood of each point contains only the point itself, making it locally homeomorphic to $\mathbb{R}^0$, thus naturally forming a 0-dimensional discrete manifold. The Boolean operations $\neg, \wedge, \vee$ become linear or piecewise linear mappings in the coordinate representation (disjunction corresponds to coordinate-wise maximum, conjunction to coordinate-wise minimum, negation to coordinate-wise flipping). The distance normalization ensures $d_{\text{cog}} \in [0, 1]$, satisfying the definition of a cognitive pseudometric. □

Theorem C.12 (Cognitive Hierarchy Approximation Error Convergence)**.** *Let $\{\varepsilon_n\}$ be the approximation error sequence for the cognitive hierarchy $\{\mathbb{H}_{ext,n}\}$ of the extended system $\mathcal{M}_{ext} \cong_{\lim} \mathcal{B}_4^k$, satisfying the following conditions:*

1. *Recurrence Relation: There exists a constant $c_k > 0$ and a function $\psi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, with $\psi(x) = o(x)$ as $x \to 0$, such that*
$$\varepsilon_{n+1} \leq \psi(\varepsilon_n) + \delta_n$$
where δ_n is a higher-order perturbation term.
2. *Perturbation Term Convergence: $\sum_{n=0}^{\infty} \delta_n < \infty$.*
3. *Error Control: There exists a monotonically decreasing function $\varphi : \mathbb{N} \to \mathbb{R}_{\geq 0}$ with $\lim_{n\to\infty} \varphi(n) = 0$, such that*
$$\varepsilon_n \leq \varphi(n) + \sum_{i=0}^{n-1} \delta_i.$$
4. *When $\psi(x) = c_k x^{\alpha_k}$ (where $\alpha_k = 1 + \frac{1}{k} > 1$) and δ_n monotonically decreases to 0, it recovers the superlinear convergence special case (which still satisfies ordinary series convergence).*

These convergence conditions guarantee $\lim_{n\to\infty} \varepsilon_n = 0$ and the existence of the cognitive hierarchy limit.

Proof. **Derivation of the Recurrence Relation**:

From the recursive construction of the cognitive hierarchy (Definition A.13) and the uniform Lipschitz condition of the cognitive pseudometric (Definition A.1, existence of a uniform constant $L > 0$ such that $\text{Lip}(f) \leq L$), the $(n+1)$-th layer error stems from the propagation of operations through the endomorphism monoid and the superposition of approximate homomorphism errors. Specifically, the endomorphism monoid of the extended system $\mathcal{M}_{\text{ext}} \cong \mathcal{B}_4^k$ has cardinality 6^k (Theorem C.7). For any $f \in \text{End}\mathcal{M}_{\text{ext}}$, its error propagation satisfies:

$$\begin{aligned}\varepsilon_{n+1} &\leq \sup_{f \in \text{End}\mathcal{M}_{\text{ext}}} \text{Err}(Y_{\text{ext}}(f)) \\ &\leq |\text{End}\mathcal{M}_{\text{ext}}| \cdot \sup_f d(f(Y_{\text{ext}}(f)), Y_{\text{ext}}(f)) \\ &\leq 6^k \cdot \varepsilon_n \cdot \sup_f \text{Lip}(f) + \delta_n.\end{aligned}$$

By the non-redundancy of the extended system, $\sup_f \text{Lip}(f) \leq \varepsilon_n^{1/k}$ (error decays polynomially with the number of components k), hence $\varepsilon_{n+1} \leq 6^k \cdot \varepsilon_n^{1+\frac{1}{k}} + \delta_n$, i.e., $\psi(x) = 6^k x^{\alpha_k}$ ($\alpha_k = 1 + \frac{1}{k} > 1$), satisfying $\psi(x) = o(x)$ ($x \to 0$).

Convergence of Perturbation Terms and Error Upper Bound:

From the error series convergence in Definition A.16, $\sum_{n=0}^{\infty} \zeta_n < \infty$. The term δ_n, as a higher-order perturbation, essentially represents the residual error of the connecting homomorphisms in the cognitive hierarchy, hence $\sum_{n=0}^{\infty} \delta_n < \infty$. By iterating the recurrence relation, for any n, we have:

$$\varepsilon_n \leq \psi(\varepsilon_{n-1}) + \delta_{n-1} \leq \psi(\varphi(n-1) + \sum_{i=0}^{n-2} \delta_i) + \delta_{n-1}.$$

Since $\psi(x) = o(x)$ and $\varphi(n)$ monotonically decreases to 0, we can construct $\varphi(n) = \psi(\varphi(n-1))$, satisfying $\varepsilon_n \leq \varphi(n) + \sum_{i=0}^{n-1} \delta_i$.

Proof of Limit Existence:

Since $\varphi(n)$ is monotonically decreasing and bounded below (error is non-negative), $\lim_{n\to\infty} \varphi(n) = \varphi^* \geq 0$. If $\varphi^* > 0$, then $\psi(\varphi^*) = \varphi^*$, contradicting $\psi(x) = o(x)$. Therefore, $\varphi^* = 0$. Furthermore, since $\sum_{i=0}^{\infty} \delta_i < \infty$, we have $\sum_{i=0}^{n-1} \delta_i \to \sum_{i=0}^{\infty} \delta_i < \infty$. Combined with the power-law decay property of $\psi(\varepsilon_n)$, this forces $\varepsilon_n \to 0$. In conclusion, $\lim_{n\to\infty} \varepsilon_n = 0$, and the cognitive hierarchy limit exists. □

Theorem C.13 (Approximate Cognitive Dynamics Equation)**.** *Let $\mathcal{M}_{ext} \cong_{\lim} \mathcal{B}_4^k$. Its cognitive state evolution is described by a stochastic approximation dynamical system:*

$$x_{n+1} = F_n(x_n; \theta_n) + \xi_n, \quad x_n \in \mathcal{X} \subset [0,1]^{2k}$$

where:

- $F_n = \varphi \circ (\sigma_n \circ \pi_n) \circ \varphi^{-1}$ *is the ε_n-approximate geometric realization of the perception-action closed loop.*
- $\theta_n \in \Theta$ *are system parameters, satisfying adaptive updates $\theta_{n+1} = \theta_n + \eta_n \nabla L(\theta_n)$.*
- ξ_n *is cognitive noise, satisfying $\mathbb{E}[\xi_n] = 0$, $\mathrm{Var}(\xi_n) \leq \varepsilon_n^2$.*
- *Convergence: There exists almost sure convergence $\lim_{n\to\infty} x_n = x^*$, and x^* is an approximate fixed point: $d(F_\infty(x^*), x^*) \leq \varepsilon^*$.*

Proof. By the idempotence of bidirectional interaction in Axiom A.1, the perception-action closed-loop mapping $\sigma_n \circ \pi_n$ satisfies $\mathrm{Pred} \circ \mathrm{Pred} \xrightarrow{\varepsilon_n} \mathrm{Pred}$. Therefore, $F_n = \varphi \circ (\sigma_n \circ \pi_n) \circ \varphi^{-1}$ asymptotically tends to an idempotent mapping.

By the uniform Lipschitz condition of the cognitive pseudometric (Definition A.1.3), there exists a uniform constant $L > 0$ such that $\mathrm{Lip}(F_n) \leq L$, ensuring the continuity of F_n.

The cognitive noise ξ_n satisfies $\mathbb{E}[\xi_n] = 0$ and $\mathrm{Var}(\xi_n) \leq \varepsilon_n^2$. By the error convergence in Theorem C.12, $\varepsilon_n \to 0$, so ξ_n asymptotically tends to 0.

Combined with the convergence of the Oraclestimator in stochastic approximation theory, there exists almost sure convergence $\lim_{n\to\infty} x_n = x^*$, and x^* satisfies $d(F_\infty(x^*), x^*) \leq \varepsilon^*$, where $\varepsilon^* = \lim_{n\to\infty} \varepsilon_n$. □

C.6 Limit Behavior and Infinite Extension

Theorem C.14 (Finite-Infinite Limit)**.** *As $k \to \infty$ ($N = 4^k \to \infty$):*

1. ***Algebraic Limit**: $\lim_{k\to\infty} \mathcal{M}_{ext}^{(k)} \cong \mathcal{P}(\mathbb{N})$ (the Boolean algebra of the power set of natural numbers).*
2. ***Topological Limit**: $\lim_{k\to\infty} S(\mathcal{M}_{ext}^{(k)}) \cong \beta\mathbb{N}$ (the $\mathrm{Stone} - \check{\mathrm{C}}\mathrm{ech}$ compactification).*
3. ***Capacity Asymptotics**: $\lim_{k\to\infty} \frac{C_{cog}(\mathcal{M}_{ext}^{(k)})}{N} = 2$ (from Definition C.2, $C_{cog}(\mathcal{B}_4^k) = 2k$, $N = 4^k$, hence the asymptotic ratio is 2).*
4. ***Geometric Limit**: The embedding space tends to ℓ^2, and the cognitive manifold is dense in the weak topology.*

Proof. (1) By the classification theorem of Boolean algebras, countably generated complete Boolean algebras are isomorphic to $\mathcal{P}(\mathbb{N})$. (2) Stone duality: The Stone space of $\mathcal{P}(\mathbb{N})$ is $\beta\mathbb{N}$. (3) Direct corollary of the capacity formula $C_{\text{cog}} = 2N$. (4) Due to the normalization of embedding coordinates, the point set is dense in the unit cube under the L^2 metric. □

Corollary C.8 (Mathematical Characterization of Infinite Cognition)**.** *The mathematical structure of an infinite AGI system ($|M| = \infty$) is:*

$$\mathcal{M}_\infty \cong \mathcal{P}(\mathbb{N}), \quad S(\mathcal{M}_\infty) \cong \beta\mathbb{N}, \quad C_{cog}(\mathcal{M}_\infty) = \infty$$

and it preserves all cognitive properties of finite systems (bidirectionality, metacognition, limit isomorphism).

D Machine Learning Mechanisms on AGI Systems: Approximate Sheaf Formulation

D.1 Sheaf Structure of the State Universe

This section elevates the previously established AGI algebraic framework to a sheaf-theoretic formulation. The core idea is:

1. To regard the cognitive hierarchy $\{\mathbb{H}_n\}$ as an approximate sheaf over a topological space.
2. To formulate the machine learning mechanism as dynamics on a parameter fiber category.
3. Axiom D.1 is the concrete realization of Axiom A.1 within the sheaf framework.

All definitions maintain consistency with previous notations. $\xrightarrow{\varepsilon}$ still denotes ε-approximate homomorphism, and d denotes the cognitive pseudometric.

Definition D.1 (ε-Approximate Sheaf)**.** *Let (X, τ) be a topological space, and $\mathcal{C}$ be a category equipped with a pseudometric d (compatible with Definition A.1). An ε-approximate sheaf $\mathscr{F}$ consists of the following data:*

1. *For each open set $U \in \tau$, an object $\mathscr{F}(U) \in \mathcal{C}$ is given.*
2. *For each inclusion $V \subseteq U$, an ε_{UV}-approximate morphism $\rho_{UV} : \mathscr{F}(U) \xrightarrow{\varepsilon_{UV}} \mathscr{F}(V)$ is given, satisfying:*
   - $\rho_{UU} = \mathrm{id}_{\mathscr{F}(U)}$.
   - *For $W \subseteq V \subseteq U$, $d(\rho_{UW}, \rho_{VW} \circ \rho_{UV}) \leq \varepsilon_{UV} + \varepsilon_{VW}$.*
3. *Approximate Gluing Axiom: For any open cover $\{U_i\}$, if there exist sections $s_i \in \mathscr{F}(U_i)$ satisfying*
   $$d(\rho_{U_i \cap U_j, U_i}(s_i), \rho_{U_i \cap U_j, U_j}(s_j)) \leq \varepsilon_{ij}$$
   then there exists $s \in \mathscr{F}(\bigcup_i U_i)$ such that $d(\rho_{\bigcup_i U_i, U_i}(s), s_i) \leq \delta_i$, where $\delta_i = O(\max_j \varepsilon_{ij})$ and $\sum_i \delta_i < \infty$.

Definition D.2 (AGI State Sheaf)**.** *Let $\mathcal{M} \in \mathbf{AGI}$, and $\{\mathbb{H}_n\}_{n=0}^{\infty}$ be a cognitive hierarchy sequence (Definition A.13). Define the **state universe** $\Omega = (\mathbb{N}, \tau_{cofinite})$ (cofinite topology). The **cognitive sheaf** $\mathscr{H}$ on Ω is an ε-approximate sheaf (Definition D.1):*

1. *For a cofinite open set $U \subseteq \mathbb{N}$, define $\mathscr{H}(U) = \varprojlim_{n \in U} \mathbb{H}_n$ (the approximate inverse limit exists, guaranteed by the cognitive hierarchy convergence constraint A.16).*
2. *$\mathscr{H}(\mathbb{N}) = \mathbb{U} = \varinjlim \mathbb{H}_n$ (inductive limit).*
3. *Restriction maps: For $V \subseteq U$, $\rho_{UV} : \mathscr{H}(U) \to \mathscr{H}(V)$ is induced by the connecting approximate homomorphisms μ_n, satisfying compatibility conditions.*

Definition D.3 (Parameter Fiber Category)**.** *Let $\mathcal{A}$ be a small category (parameter category), $\mathcal{P}$ be a category (total category), and $\pi : \mathcal{P} \to \mathcal{A}$ be a functor. For each $a \in \mathcal{A}$, the fiber is $\mathcal{P}_a = \pi^{-1}(a)$. The triple $(\mathcal{P}, \pi, \mathcal{A})$ is called a parameter fiber category if for each morphism $f : a \to b$ in $\mathcal{A}$, there exists a lifting functor $F_f : \mathcal{P}_b \to \mathcal{P}_a$.*

Definition D.4 (Parameter Sheaf)**.** *Let $(\mathcal{P}, \pi, \mathcal{A})$ be a parameter fiber category (Definition D.3), where $\mathcal{A}$ is a small category. Define the **parameter sheaf** $\mathscr{P}$ as a sheaf on $\mathcal{A}$:*

1. *For each open subcategory $U \subseteq \mathcal{A}$, $\mathscr{P}(U) = \prod_{a \in \mathrm{Ob}(U)} \mathcal{P}_a$.*
2. *Restriction maps are induced by the lifting functors $F_{a,b} : \mathcal{P}_b \to \mathcal{P}_a$ (for $a \leq b$).*
3. *The interpretation maps $\rho_{a,\mathcal{M}} : \mathcal{P}_a \to \mathcal{M}$ induce a sheaf morphism $\rho : \mathscr{P} \to \mathscr{M}$, where $\mathscr{M}$ is the constant sheaf on $\mathcal{M}$.*

Axiom D.1 (Sheaf-Theoretic Formulation of the Machine Learning Meta-Axiom System)**.** *Let $\mathcal{M} \in \mathbf{AGI}$. Its machine learning mechanism is characterized by the following axioms:*

1. ***Existence of State Sheaf****: There exists a complete metric space (Ω, d_Ω) (state universe) and an ε-approximate sheaf $\mathscr{F}_\Omega$ on it (Definition D.1), such that $\mathscr{F}_\Omega$ is the completion of the cognitive sheaf $\mathscr{H}$.*

2. ***Observation Interface Decomposition****: There exists a sheaf decomposition $\mathscr{F}_\Omega \cong \mathscr{I} \oplus \mathscr{O} \oplus \mathscr{H}$, where $\mathscr{I}$ is the input sheaf, $\mathscr{O}$ is the output sheaf, and $\mathscr{H}$ is the hidden state sheaf.*

3. ***Endogenous Meta-Dynamics****: There exists a sheaf endomorphism $\Phi : \mathscr{F}_\Omega \to \mathscr{F}_\Omega$, satisfying:*

   - *It is generated by a finite set of generating sheaf vector fields $\{X_1, \ldots, X_r\}$.*
   - *There exists a time step $\Delta > 0$ such that $d(\Phi_\Delta(s), \mathrm{Pred}(s)) \le \varepsilon(\Delta)$, with $\varepsilon(\Delta) \to 0$.*
   - *Φ is a contraction mapping: there exists $L < 1$ such that $d(\Phi(s), \Phi(t)) \le L d(s, t)$.*

4. ***Cognitive Homomorphism Realization****: There exist approximate sheaf homomorphisms $\rho : \mathscr{F}_\Omega \to \mathscr{P}$ and its approximate inverse $\rho^{-1} : \mathscr{P} \to \mathscr{F}_\Omega$, where $\mathscr{P}$ is the parameter sheaf (Definition D.4).*

5. ***Learning Initialization****: A machine learning task reduces to choosing an initial section $\sigma_0 \in \Gamma(\Omega, \mathscr{F}_\Omega)$.*

6. ***Ergodicity****: For any open set $U \subseteq \Omega$ and section s, the asymptotic time average of the trajectory $\{\Phi_t(s)\}$ over U equals the spatial average.*

Definition D.5 (Fiber Dynamics Sheaf)**.** *Let $\pi : \mathscr{P} \to \mathcal{A}$ be a parameter sheaf, and $\mathscr{F}_\Omega$ its completion. Define the **fiber dynamics sheaf** $\mathscr{F}$ as a sheaf on $\mathcal{A}$:*

1. *For each open set $U \subseteq \mathcal{A}$, $\mathscr{F}(U) = \{\varphi : U \times \mathbb{R}_{\ge 0} \to \mathscr{P} \mid \pi \circ \varphi = \mathrm{pr}_1, \varphi \text{ is a continuous sheaf section}\}$.*

2. *Restriction maps are restrictions of sections.*

3. *Local Flow Condition: For any $a \in \mathcal{A}$, there exists a neighborhood $U \ni a$ and a local section $\varphi \in \mathscr{F}(U)$ satisfying the local flow equation.*

Definition D.6 (Dynamics Semantic Encoding Sheaf)**.** *Let $\mathrm{Enc} : \mathcal{X} \to \mathscr{P}$ be a traditional semantic encoding. Define the **dynamics semantic encoding sheaf** $\mathrm{Enc}_{\mathscr{F}} : \mathcal{X} \to \mathscr{F}_\Omega$ as a sheaf map satisfying:*

1. *$\mathrm{Enc} = \rho \circ \mathrm{Enc}_{\mathscr{F}}$ ($\varepsilon_{\mathrm{enc}}$-approximately).*

2. *For any $x \in \mathcal{X}$, there exists $\tau(x)$ such that $\mathrm{Enc}_{\mathscr{F}}(\varphi(x)) = \Phi_{\tau(x)}(\mathrm{Enc}_{\mathscr{F}(x)})$.*

3. *$\mathrm{Enc}_{\mathscr{F}}(\mathcal{X})$ is dense in the section space of $\mathscr{F}_\Omega$.*

Definition D.7 (Cognitive Attractor Sheaf)**.** *A sheaf submodule $\mathscr{A} \subseteq \mathscr{F}_\Omega$ is called a **cognitive attractor sheaf** if:*

1. *$\Phi_t(\mathscr{A}) = \mathscr{A}$ for all $t \ge 0$.*

2. *There exists a neighborhood sheaf $\mathscr{U} \supset \mathscr{A}$ such that $\lim_{t\to\infty} \mathrm{dist}(\Phi_t(\sigma), \mathscr{A}) = 0$ for all sections $\sigma \in \Gamma(\mathscr{U})$.*

3. *$\mathscr{A}$ is minimal.*

D.2 Connectivity of Discrete Representation and Continuous Process

Theorem D.1 (Approximate Homomorphism Connection between Discrete Representation and Continuous Process)**.** *Let $\mathcal{M} \cong \mathcal{B}_4^{k_{\min}}$ be the carrier set of an AGI system (satisfying Theorem C.6), and $\mathscr{F}_\Omega$ be the state sheaf in the meta-dynamics sheaf paradigm satisfying Axiom D.1 (Definition D.2). Then there exist sequences of approximate homomorphisms $\{\rho_n : \mathscr{F}_\Omega \to \mathcal{M}\}$ and $\{\sigma_n : \mathcal{M} \to \mathscr{F}_\Omega\}$, such that for any $\varepsilon > 0$, there exists $N \in \mathbb{N}$, and when $n > N$:*

1. *ρ_n and σ_n are ε-approximate homomorphisms (Definition D.1).*

2. $\rho_n \circ \sigma_n \approx_\varepsilon \mathrm{id}_{\mathcal{M}}$ *and* $\sigma_n \circ \rho_n \approx_\varepsilon \mathrm{id}_{\mathscr{F}_\Omega}$.

3. *They preserve the operation structure: for any operation* $\omega \in \Omega$, $\rho_n(\omega_{\mathscr{F}_\Omega}(x)) \approx_\varepsilon \omega_{\mathcal{M}}(\rho_n(x))$ *and* $\sigma_n(\omega_{\mathcal{M}}(y)) \approx_\varepsilon \omega_{\mathscr{F}_\Omega}(\sigma_n(y))$.

Proof. Since $\mathcal{M}$ is a finite discrete space (cardinality $4^{k_{\min}}$) and $\mathscr{F}_\Omega$ is a complete metric space, we construct as follows:

1. Partition Construction: For each $n \in \mathbb{N}$, take a finite partition $\mathcal{P}_n = \{A_{n,1}, \ldots, A_{n,4^{k_{\min}}}\}$ of $\mathscr{F}_\Omega$, satisfying:
   - $\bigcup_{i=1}^{4^{k_{\min}}} A_{n,i} = \mathscr{F}_\Omega$.
   - $A_{n,i} \cap A_{n,j} = \emptyset$ for $i \neq j$.
   - $\mathrm{diam}(A_{n,i}) < \delta_n$, where $\delta_n \to 0$ as $n \to \infty$.
2. Mapping Definition:
   - Define $\rho_n : \mathscr{F}_\Omega \to \mathcal{M}$ by setting $\rho_n(x) = m_i$ when $x \in A_{n,i}$, where $\{m_1, \ldots, m_{4^{k_{\min}}}\}$ are the elements of $\mathcal{M}$.
   - Define $\sigma_n : \mathcal{M} \to \mathscr{F}_\Omega$ by $\sigma_n(m_i) = x_{n,i} \in A_{n,i}$, a fixed representative point.
3. Approximate Homomorphism Verification:
   - Since partition diameters tend to 0, for any $x, y \in A_{n,i}$, $d(x, y) < \delta_n$.
   - By the continuity of operations on $\mathscr{F}_\Omega$ (Axiom D.1), there exists a Lipschitz constant L.
   - When δ_n is sufficiently small, ρ_n and σ_n are ε-approximate homomorphisms, with $\varepsilon = O(\delta_n)$.
4. Composition Approximates Identity:
   - $\rho_n \circ \sigma_n(m_i) = \rho_n(x_{n,i}) = m_i$ (strict identity).
   - $\sigma_n \circ \rho_n(x) = \sigma_n(m_i) = x_{n,i}$, where $x \in A_{n,i}$, hence $d(x, \sigma_n \circ \rho_n(x)) < \delta_n$.
5. Operation Structure Preservation:
   - Since operations are continuous and the partition is sufficiently fine, operation results land in adjacent partition blocks, with controllable error.
   - Specifically: Let $x \in A_{n,i}$, then $\omega_{\mathscr{F}_\Omega}(x) \in A_{n,j}$ and $\omega_{\mathcal{M}}(m_i) = m_j$, thus $\rho_n(\omega_{\mathscr{F}_\Omega}(x)) = m_j = \omega_{\mathcal{M}}(\rho_n(x))$.

Taking a sequence $\delta_n = \varepsilon_n \to 0$ yields the desired approximate homomorphism sequence. □

Definition D.8 (Encoding-Decoding Pair)**.** *Let* $\mathcal{X}$ *be a compact metric space,* $\mathcal{M} \cong \mathcal{B}_4^{k_{\min}}$. *An* (ε, L)*-encoding-decoding pair consists of maps* $\mathrm{Enc} : \mathcal{X} \to \mathcal{M}^L$ *and* $\mathrm{Dec} : \mathcal{M}^L \to \mathcal{X}$*, satisfying:*

- $\mathrm{Dec} \circ \mathrm{Enc} \approx_\varepsilon \mathrm{id}_{\mathcal{X}}$.
- Enc *and* Dec *are* L*-Lipschitz continuous.*

Theorem D.2 (Approximate Representation Theorem: Dense Representation of Continuous Space by Discrete Structure)**.** *Let* $\mathcal{X}$ *be a compact metric space,* $\mathcal{M} \cong \mathcal{B}_4^{k_{\min}}$. *Then for any* $\varepsilon > 0$*, there exist* $L = O(\log(1/\varepsilon))$ *and an* $(\varepsilon, O(1))$*-encoding-decoding pair* $(\mathrm{Enc}_\varepsilon, \mathrm{Dec}_\varepsilon)$.

Proof.
1. Finite Cover Construction: Since $\mathcal{X}$ is compact, there exists a finite δ-net $S_\delta = \{s_1, \ldots, s_N\}$, where $\delta = \varepsilon/2$ and $N = N(\delta)$ is the covering number.
2. Encoding Length Determination: Take $L = \lceil \log_{|\mathcal{M}|}(N) \rceil = \lceil \log_{4^{k_{\min}}}(N) \rceil = O(\log N) = O(\log(1/\varepsilon))$.
3. Encoding Map Construction:

- Map each point s_i in the net to a unique codeword c_i in $\mathcal{M}^L$.
- Define $\mathrm{Enc}_\varepsilon(x) = \arg\min_{c_i} d_{\mathcal{X}}(x, s_i)$ (nearest neighbor encoding).

4. Decoding Map Construction: $\mathrm{Dec}_\varepsilon(c_i) = s_i$.

5. Error Analysis:

   - For any $x \in \mathcal{X}$, there exists s_i with $d(x, s_i) < \delta$.
   - Hence $d(\mathrm{Dec}_\varepsilon \circ \mathrm{Enc}_\varepsilon(x), x) \leq d(s_i, x) < \delta = \varepsilon/2 < \varepsilon$.

6. Lipschitz Continuity:

   - Since $\mathcal{X}$ is compact, Enc_ε maps points with distance less than η to the same or adjacent codewords.
   - Dec_ε maps from discrete to continuous; Lipschitz continuity can be ensured via interpolation.
   - Specific construction: define barycentric coordinates between codewords to obtain an $O(1)$-Lipschitz map.

□

Theorem D.3 (Realizability of Discrete Dynamics in Continuous Space)**.** *Let $F : \mathcal{M} \to \mathcal{M}$ be any map (discrete dynamics), where $\mathcal{M} \cong \mathcal{B}_4^{k_{\min}}$. Then there exists a continuous dynamics $G : \mathscr{F}_\Omega \to \mathscr{F}_\Omega$ and approximate homomorphisms $\rho : \mathscr{F}_\Omega \to \mathcal{M}$, $\sigma : \mathcal{M} \to \mathscr{F}_\Omega$, such that the following diagram ε-approximately commutes:*

$$\begin{array}{ccc} \mathscr{F}_\Omega & \xrightarrow{G} & \mathscr{F}_\Omega \\ \downarrow \rho & & \downarrow \rho \\ \mathcal{M} & \xrightarrow{F} & \mathcal{M} \end{array}$$

i.e., $\rho \circ G \approx_\varepsilon F \circ \rho$, and $G \circ \sigma \approx_\varepsilon \sigma \circ F$.

Proof. 1. Lifting Construction: Take the approximate homomorphisms ρ_n, σ_n from Theorem D.1 and define:

$$G_n = \sigma_n \circ F \circ \rho_n : \mathscr{F}_\Omega \to \mathscr{F}_\Omega.$$

2. Approximate Commutativity Verification:

   - $\rho_n \circ G_n = \rho_n \circ \sigma_n \circ F \circ \rho_n \approx_{\delta_n} F \circ \rho_n$, since $\rho_n \circ \sigma_n \approx_{\delta_n} \mathrm{id}_{\mathcal{M}}$.
   - $G_n \circ \sigma_n = \sigma_n \circ F \circ \rho_n \circ \sigma_n \approx_{\delta_n} \sigma_n \circ F$, since $\rho_n \circ \sigma_n \approx_{\delta_n} \mathrm{id}_{\mathcal{M}}$.

3. Continuity Correction: G_n may be discontinuous (because ρ_n is discrete), requiring smoothing:

   - Take a partition of unity $\{\varphi_i\}$ subordinate to the partition $\mathcal{P}_n$.
   - Define a smoothed version: $G(x) = \sum_i \varphi_i(x) \cdot \sigma_n(F(m_i))$.
   - G is continuous and satisfies $\|G - G_n\|_\infty \leq \max_i \mathrm{diam}(A_{n,i}) < \delta_n$.

4. Final Estimate:

   $$\|\rho_n \circ G - F \circ \rho_n\| \leq \|\rho_n \circ (G - G_n)\| + \|\rho_n \circ G_n - F \circ \rho_n\| \leq L\rho_n\delta_n + \delta_n = O(\delta_n).$$

   Taking $\delta_n < \varepsilon/(L\rho_n + 1)$ yields the desired result.

□

Definition D.9 (Dynamics Parameterized Family)**.** *Let $\Theta \subseteq \mathbb{R}^d$ be a parameter space. A family $\{G_\theta : \mathscr{F}_\Omega \to \mathscr{F}_\Omega\}_{\theta \in \Theta}$ of Lipschitz continuous maps is a dynamics parameterized family if it satisfies:*

- *Universal Approximation: For any continuous $G : \mathscr{F}_\Omega \to \mathscr{F}_\Omega$ and $\varepsilon > 0$, there exists $\theta \in \Theta$ such that $\|G_\theta - G\|_\infty < \varepsilon$.*

- *Lipschitz Constant Control:* $\mathrm{Lip}(G_\theta) \leq L(\theta)$ *and is optimizable.*

Theorem D.4 (Learning Convergence Theorem: Gradient Descent Learns Approximate Dynamics)**.** *Let training data be* $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N \subset \mathscr{F}_\Omega \times \mathscr{F}_\Omega$*, and the loss function* $\ell : \mathscr{F}_\Omega \times \mathscr{F}_\Omega \to \mathbb{R}_{\geq 0}$ *be convex and* L_ℓ*-Lipschitz. Define the empirical risk:*

$$L(\theta) = \frac{1}{N} \sum_{i=1}^{N} \ell(G_\theta(x_i), y_i) + \lambda R(\theta)$$

where $R(\theta)$ *is a regularization term. Consider the gradient descent iteration:*

$$\theta_{t+1} = \theta_t - \eta_t \nabla L(\theta_t).$$

Then, under suitable conditions (learning rate $\eta_t = O(1/t)$*, L smooth convex), we have:*

$$\min_{t=1,\ldots,T} L(\theta_t) - L(\theta^*) \leq \frac{C}{\sqrt{T}}$$

where $\theta^* = \arg\min_\theta L(\theta)$ *and* C *is a constant.*

Proof. 1. Convexity Assumption: Assume $L(\theta)$ is μ-strongly convex and M-smooth with respect to θ.

2. Gradient Descent Convergence Rate: For strongly convex smooth functions, gradient descent satisfies:
$$L(\theta_t) - L(\theta^*) \leq (1 - \mu\eta)^t (L(\theta_0) - L(\theta^*)).$$
Taking $\eta = 1/M$ yields linear convergence.

3. Non-Convex Case (More Realistic): If L is non-convex but satisfies the Polyak-Łojasiewicz condition:
$$\frac{1}{2} \|\nabla L(\theta)\|^2 \geq \mu (L(\theta) - L(\theta^*))$$
then gradient descent still has a linear convergence rate.

4. Stochastic Gradient Descent: For large datasets, use SGD:
$$\theta_{t+1} = \theta_t - \eta_t \nabla \ell(G_\theta(x_{i_t}), y_{i_t})$$
Under smoothness assumptions, $\mathbb{E}[L(\theta_t)] - L(\theta^*) \leq O(1/\sqrt{t})$.

5. Application to Dynamics Learning:
   - By Theorem D.3, the target dynamics G^* has a parameterized approximation G_{θ^*}.
   - The $\hat{\theta}$ found by gradient descent satisfies $L(\hat{\theta}) \leq L(\theta^*) + \varepsilon$.
   - Thus the learned dynamics approximates the target discrete dynamics.

□

D.3 Fiber Dynamics

Theorem D.5 (Existence and Uniqueness of Fiber Dynamics Sheaf)**.** *Let* $\mathcal{M} \in \mathbf{AGI}$ *satisfy Axiom D.1. Then there exists a unique fiber dynamics sheaf* $\mathscr{F}$ *satisfying:*

1. *Local Generativity: The local dynamics on each fiber* $\mathscr{P}_a$ *is generated by a sheaf vector field* X_a*.*
2. *Transverse Compatibility: The lifting functors* $F_{a,b}$ *commute with the dynamics.*
3. *Global Realizability: The global dynamics* Φ *can be locally represented as the direct product of fiber sheaves.*

Proof. **Existence Construction**: By Axiom D.1, there exist the state sheaf $\mathscr{F}_\Omega$ and the dynamics Φ. For each parameter $a \in \mathcal{A}$, the fiber $\mathscr{P}_a$ can be embedded into $\mathscr{F}_\Omega$. Define the local vector field X_a as generated by the restriction of Φ to $\mathscr{P}_a$:

$$X_a = \rho \circ \Phi \circ \rho^{-1}|_{\mathscr{P}_a}.$$

By the operation generation completeness (Axiom A.1), X_a is generated by a finite set of generators.

Transverse Compatibility: For $a \leq b$, the lifting functor $F_{a,b}$ is an approximate homomorphism. By the operation structure constraint of Definition A.15, $F_{a,b}$ commutes with the dynamics with controllable error.

Global Realizability: By the observation interface decomposition $\mathscr{F}_\Omega = \mathscr{I} \oplus \mathscr{O} \oplus \mathscr{H}$, the dynamics Φ can be decomposed into dynamics on each component. The whole is realized as the local direct product.

Uniqueness: If there exists another fiber dynamics sheaf $\mathscr{F}'$ satisfying the same conditions, then by the uniqueness of approximate sheaf isomorphism, $\mathscr{F}$ and $\mathscr{F}'$ are approximately isomorphic, hence unique in the limit sense.

Note: If the carrier set $\mathcal{M}$ further satisfies the conditions of Theorem C.6, then there exists a positive integer k such that $\mathscr{F}$ is isomorphic to the direct sum of k irreducible subsheaves. This decomposition property is a special case of the algebraic structure and does not change the existence and uniqueness conclusion under the general framework. □

Theorem D.6 (Dynamical Realization of Prediction Mechanism)**.** *Let* $\mathrm{Pred} : \mathscr{P} \to \mathscr{P}$ *be a traditional prediction mechanism. Then, within the framework of Axiom D.1, there exists a unique optimal time step* $\Delta_{\mathrm{opt}} > 0$ *and an approximate sheaf homomorphism* ρ_{opt}*, such that*

$$\mathrm{Pred} = \rho_{\mathrm{opt}} \circ \Phi_{\Delta_{\mathrm{opt}}} \circ \rho_{\mathrm{opt}}^{-1} \quad (\varepsilon_{\mathrm{opt}}\text{-}approximately)$$

and it satisfies error minimality, time optimality, and uniqueness.

Proof. **Continuity Proof**: By the Lipschitz condition of the cognitive pseudometric (Definition A.1.3), all approximate homomorphisms are uniformly Lipschitz continuous. $\mathrm{Pred}_{\mathscr{F}} = \rho^{-1} \circ \mathrm{Pred} \circ \rho$, where ρ and ρ^{-1} are approximate sheaf homomorphisms, guaranteed to be continuous (approximately) by Axiom D.1.4. Φ_Δ is the endogenous meta-dynamics. By Axiom D.1.3, it is generated by a finite set of generating sheaf vector fields, which are Lipschitz continuous on the complete metric space. Hence Φ_Δ is continuous in Δ (by the continuity of solutions of ordinary differential equations with respect to parameters).

Strict Proof of Minimum Existence: Define $E(\Delta) = \sup_{\sigma \in \Gamma(\mathscr{F}_\Omega)} d(\mathrm{Pred}_{\mathscr{F}}(\sigma), \Phi_\Delta(\sigma))$. Since $\Gamma(\mathscr{F}_\Omega)$ is a complete metric space under the cognitive metric (Axiom D.1.1), and $\mathrm{Pred}_{\mathscr{F}}$ and Φ_Δ are uniformly continuous, $E(\Delta)$ is well-defined.

Consider $\Delta \in [0, \infty]$. We need to prove $E(\Delta)$ attains its minimum at some finite Δ. By Axiom D.1.3, Φ is a contraction mapping; there exists $L < 1$ such that $d(\Phi_\Delta(\sigma), \Phi_\Delta(\sigma')) \leq L d(\sigma, \sigma')$. As $\Delta \to 0$, $\Phi_\Delta \to \mathrm{id}$ (by the dynamical equation), hence $E(0) = \sup_\sigma d(\mathrm{Pred}_{\mathscr{F}}(\sigma), \sigma)$. As $\Delta \to \infty$, by contraction, $\Phi_\Delta(\sigma)$ converges to a fixed point σ^*, hence $E(\infty) = \sup_\sigma d(\mathrm{Pred}_{\mathscr{F}}(\sigma), \sigma^*)$.

Since $\mathrm{Pred}_{\mathscr{F}}$ does not necessarily map σ to a fixed point, $E(\infty)$ may be large. But $E(\Delta)$ is continuous on $[0, \infty)$, and $\lim_{\Delta \to \infty} E(\Delta)$ exists (possibly infinite). If $\lim_{\Delta \to \infty} E(\Delta) = \infty$, then $E(\Delta)$ attains its minimum at some finite Δ. If $\lim_{\Delta \to \infty} E(\Delta)$ is finite, the minimum might be at infinity, but we can then consider minimizing Δ on a compact interval $[0, T]$ for sufficiently large T.

In practice, since $\mathrm{Pred}_{\mathscr{F}}$ is an approximate prediction mechanism, it should be close to $\Phi_{\Delta_{\mathrm{opt}}}$, so the minimum is attained at a finite Δ. □

Theorem D.7 (Dynamical Decomposition of Semantic Encoding)**.** *Any semantic encoding* $\mathrm{Enc} : \mathcal{X} \to \mathscr{P}$ *can be uniquely decomposed as:*

$$\mathrm{Enc} = \rho \circ \mathrm{Enc}_{\mathscr{F}} \circ \kappa \quad (\varepsilon\text{-}approximately)$$

where $\kappa : \mathcal{X} \to \mathcal{X}$ *is a normalization sheaf map,* $\mathrm{Enc}_{\mathscr{F}}$ *is the dynamics semantic encoding sheaf (Definition D.6), and* ρ *is the optimal cognitive homomorphism sheaf realization.*

Proof. By the cognitive homomorphism sheaf realization in Axiom D.1 (here ρ is an approximate sheaf homomorphism, whose definition does not require the operations on $\mathcal{M}$ to be Boolean, only that $\mathcal{M}$ satisfies the AGI axioms.), there exist $\rho : \mathscr{F}_\Omega \to \mathscr{P}$ and its approximate inverse. Define $\mathrm{Enc}_{\mathscr{F}} = \rho^{-1} \circ \mathrm{Enc}$, then $\mathrm{Enc} = \rho \circ \mathrm{Enc}_{\mathscr{F}}$. However, $\mathrm{Enc}_{\mathscr{F}}$ may not satisfy condition 2 of the dynamics semantic encoding sheaf. Therefore, we need to normalize $\mathcal{X}$.

Since the trajectories of the dynamics Φ are dense in the state space (ergodicity axiom), for any $x \in \mathcal{X}$, there exists $\tau(x)$ such that $d(\mathrm{Enc}_{\mathscr{F}(\varphi(x))}, \Phi_{\tau(x)}(\mathrm{Enc}_{\mathscr{F}(x)})) < \varepsilon$, where φ is some transformation. Construct $\kappa : \mathcal{X} \to \mathcal{X}$ as mapping x to its representative on the dynamical orbit, i.e., choosing the minimal $\tau(x)$ that minimizes the above distance. By density of orbits, such a representative exists. Since $\mathrm{Enc}_{\mathscr{F}}$ is injective (approximately), the representative is unique.

By the density condition (Definition D.6 condition 3), $\mathrm{Enc}_{\mathscr{F}}(\mathcal{X})$ is dense in $\mathscr{F}_\Omega$, hence κ is surjective.

Uniqueness: If there exists another decomposition $\mathrm{Enc} = \rho' \circ \mathrm{Enc}'_{\mathscr{F}} \circ \kappa'$, then by the uniqueness of the approximate inverses of ρ and ρ', we have $\rho' = \rho$, $\mathrm{Enc}'_{\mathscr{F}} = \mathrm{Enc}_{\mathscr{F}}$, and $\kappa' = \kappa$ (within error). □

Theorem D.8 (Sheaf-Theoretic Realization Theorem for Semantic Encoding)**.** *The optimal semantic encoding* $\mathrm{Enc}_{\mathscr{F}} : \mathcal{X} \to \mathscr{F}_{\nVdash}$ *can be decomposed as:*

$$\mathrm{Enc}_{\mathscr{F}} = \bigoplus_{\alpha \in G} \mathrm{lift}_\alpha \circ \mathrm{Enc}_\alpha$$

where $\mathrm{lift}_\alpha : \mathscr{F}_{\nVdash_\alpha} \to \mathscr{F}_{\nVdash}$ *are section lifting operators, satisfying:*

1. *Local Preservation: They preserve modal properties at the stalk level.*
2. *Global Coordination: The lifted sections automatically satisfy compatibility relations.*
3. *Minimal Extension: The lifting operators are the minimum norm solutions.*

Corollary D.1 (Dynamical Minimality of the Machine Learning Mechanism)**.** *Let* $\mathcal{M} \in \mathbf{AGI}$ *satisfy the minimality Axiom A.2 and Axiom D.1. Then its meta-dynamics sheaf paradigm* $(\mathrm{Enc}_{\mathscr{F}}, \Phi, \rho_{\mathrm{opt}}, \sigma_0^*)$ *satisfies: minimal state space dimension, minimal dynamics generators, optimal time parameters, lower bound on initial state complexity, and lower bound on computational complexity.*

Proof.

- Minimal state space dimension: As it is obvious, $k_{\min}$ is minimal, hence the state space dimension $2k_{\min}$ is minimal.
- Minimal dynamics generators: By the operation minimality of Axiom A.1, Φ is generated by a minimal set of generators with minimal cardinality.
- Optimal time parameters: By Theorem D.6, Δ_{opt} minimizes error.
- Lower bound on initial state complexity: The complexity of σ_0^* is determined by $k_{\min}$, reaching the information lower bound.
- Lower bound on computational complexity: By the minimality axiom, computational complexity is at least $\Omega(k_{\min}^2)$.

□

Theorem D.9 (Approximate Optimality of Generalized Semantic Encoding)**.** *Let* $\mathcal{M} \in \mathbf{C}_{\mathrm{AGI}}$ *satisfy the AGI Fundamental Axiom A.1 (possessing infinite cognitive capability),* $\mathcal{X}$ *be a compact metric space (input space), and* $\mathscr{F}_\Omega$ *be the state universe sheaf satisfying Axiom D.1. Let* $\mathrm{dist} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ *be a metric on* $\mathcal{X}$*, and* $\mathrm{dist}_{\mathscr{F}} : \Gamma(\Omega, \mathscr{F}_\Omega) \times \Gamma(\Omega, \mathscr{F}_\Omega) \to \mathbb{R}_{\geq 0}$ *be the cognitive metric on the state sheaf section space (induced by the pseudometric of Definition A.1). Let* $\mathcal{E}$ *be the set of all approximate sheaf morphisms satisfying:*

$$\mathcal{E} = \{\mathrm{Enc} : \mathcal{X} \to \mathscr{F}_\Omega \mid \mathrm{Enc} \textit{ is an } \epsilon\textit{-approximate sheaf morphism}, \\ \mathrm{Enc}(\mathcal{X}) \textit{ is dense in a compact subset of } \mathscr{F}_\Omega\}.$$

Define the rate-distortion function of the encoding as:

$$R_{\mathrm{Enc}}(D) = \inf_{\mathrm{Dec}:\mathscr{F}_\Omega \to \mathcal{X}} \{ I(\mathrm{Enc}) \mid \mathbb{E}_{x \sim P_\mathcal{X}}[\mathrm{dist}(x, \mathrm{Dec} \circ \mathrm{Enc}(x))] \leq D \},$$

where $I(\mathrm{Enc})$ is the mutual information rate of the encoding Enc *(when $\mathcal{X}$ is a stochastic process) or its description length (when $\mathcal{X}$ is a finite set), $P_\mathcal{X}$ is a probability distribution on $\mathcal{X}$ (prior knowledge), and* Dec *is a decoding map (the inverse of an approximate sheaf morphism).*

Then there exists a sequence of approximate sheaf morphisms $\{\mathrm{Enc}_n\}_{n=1}^\infty \subseteq \mathcal{E}$ and a corresponding sequence of decoding maps $\{\mathrm{Dec}_n\}_{n=1}^\infty$, satisfying:

1. ***Asymptotic Optimality**: For any $D > 0$,*

$$\lim_{n \to \infty} R_{\mathrm{Enc}_n}(D) = R(D),$$

where $R(D)$ is the rate-distortion function of $\mathcal{X}$ with respect to the metric dist *and distribution $P_\mathcal{X}$ (the information-theoretic lower bound).*

2. ***Dynamical Compatibility**: There exists a sequence of time steps $\{\Delta_n\}_{n=1}^\infty$ and approximate sheaf homomorphisms $\rho_n : \mathscr{F}_\Omega \to \mathscr{P}$ ($\mathscr{P}$ is the parameter sheaf), such that*

$$\lim_{n \to \infty} \mathbb{E}_{x \sim P_\mathcal{X}} \left[\mathrm{dist}_\mathscr{F} \left(\mathrm{Enc}_n(x), \Phi_{\Delta_n}(\mathrm{Enc}_n(x)) \right) \right] = 0,$$

where Φ is the endogenous meta-dynamics from Axiom D.1. That is, the encoded state is approximately stable under dynamical evolution.

3. ***Cognitive Hierarchy Realization**: There exists a cognitive hierarchy sequence $\{\mathscr{H}_m\}_{m=0}^\infty$ (Definition A.13) and approximate sheaf morphisms $\iota_{n,m} : \mathscr{H}_m \to \mathscr{F}_\Omega$, such that for each n, there exists m_n satisfying*

$$\mathrm{Enc}_n = \iota_{n,m_n} \circ \mathrm{Enc}_{n,m_n},$$

where $\mathrm{Enc}_{n,m_n} : \mathcal{X} \to \mathscr{H}_{m_n}$ is an encoding on a finite hierarchy. Moreover, as $n \to \infty$, $m_n \to \infty$ and the error $\epsilon_{n,m_n} \to 0$.

4. ***Structure Preservation**: If $\mathcal{X}$ possesses additional algebraic or topological structure (e.g., group action, smooth manifold structure), and the operation generation completeness of $\mathcal{M}$ can express maps preserving that structure, then the optimal encoding sequence $\{\mathrm{Enc}_n\}$ can be constructed to approximately preserve that structure.*

Furthermore, if $\mathcal{M}$ satisfies the Minimality Axiom A.2 (carrier minimality), then there exists a specific $\mathrm{Enc}^* \in \mathcal{E}$ such that for any $\mathrm{Enc} \in \mathcal{E}$,

$$R_{\mathrm{Enc}^*}(D) \leq R_{\mathrm{Enc}}(D) + O\left(\frac{1}{\sqrt{k_{\min}}} \right),$$

where $k_{\min}$ is the cognitive dimension of $\mathcal{M}$ (a variant of Definition C.2, reflecting system complexity). In this case, Enc^* is called the asymptotically optimal semantic encoding.

Proof. **Step 1: Construction of finite hierarchy approximate encodings.** Since $\mathcal{M}$ satisfies infinite cognitive capability, by Axiom A.1, there exists a cognitive hierarchy sequence $\{\mathscr{H}_m\}$ such that $\mathscr{F}_\Omega = \varinjlim \mathscr{H}_m$. For each m, $\mathscr{H}_m$ is a finitely generated τ-type algebra (since $\mathcal{M}$ is finitely generated and $\mathbb{H}_m$ is recursively constructed). Assume $\mathcal{X}$ is a compact metric space, $\mathrm{dist} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ is a metric, and $P_\mathcal{X}$ is a Borel probability measure on $\mathcal{X}$ (prior).

For each m, by information-theoretic rate-distortion theory, there exists a finite set $\mathcal{C}_m \subset \mathscr{H}_m$ and an encoding-decoding pair $(\mathrm{Enc}_m : \mathcal{X} \to \mathcal{C}_m, \mathrm{Dec}_m : \mathcal{C}_m \to \mathcal{X})$, such that

$$R_{\mathrm{Enc}_m}(D_m) \leq R(D) + \varepsilon_m,$$

where: - $D_m = \mathbb{E}_{x \sim P_\mathcal{X}}[\mathrm{dist}(x, \mathrm{Dec}_m \circ \mathrm{Enc}_m(x))]$ - $R(D)$ is the rate-distortion function of $\mathcal{X}$ with respect to metric dist and distribution $P_\mathcal{X}$ - $\varepsilon_m \to 0$ as $m \to \infty$ - $|\mathcal{C}_m| = \lceil 2^{R(D)+\varepsilon_m} \rceil$

This is guaranteed by classical results of vector quantization: for compact metric spaces and continuous distortion metrics, there exists a codebook size $|\mathcal{C}_m|$ satisfying the above bound.

Step 2: Lifting to the state sheaf and constructing a Cauchy sequence. By the inductive limit structure of the cognitive hierarchy, there exist connecting sheaf morphisms $\iota_m : \mathscr{H}_m \to \mathscr{F}_\Omega$, satisfying $d(\iota_m(h), \iota_{m+1}(\mu_m(h))) \leq \zeta_m$, where $\mu_m : \mathscr{H}_m \to \mathscr{H}_{m+1}$ are the connecting homomorphisms, and $\sum_{m=0}^{\infty} \zeta_m < \infty$ (by the error convergence condition of Definition A.16).

Define $\mathrm{Enc}'_m = \iota_m \circ \mathrm{Enc}_m : \mathcal{X} \to \mathscr{F}_\Omega$. Since $\{\iota_m\}$ is a sequence of approximate sheaf morphisms with summable errors, and Enc_m takes values in finite sets $\mathcal{C}_m$, the sequence $\{\mathrm{Enc}'_m\}$ is Cauchy in the function space $L^2(\mathcal{X}, P_\mathcal{X}; \mathscr{F}_\Omega)$:

$$\mathbb{E}_{x \sim P_\mathcal{X}}[d_\mathscr{F}(\mathrm{Enc}'_m(x), \mathrm{Enc}'_{m+1}(x))] \leq \zeta_m + \delta_m,$$

where δ_m comes from the rate-distortion error between Enc_m and Enc_{m+1}.

By the completeness of $\mathscr{F}_\Omega$ (Axiom D.1.1), the sequence $\{\mathrm{Enc}'_m\}$ converges to some $\mathrm{Enc}^* : \mathcal{X} \to \mathscr{F}_\Omega$.

Step 3: Verification of asymptotic optimality. For any $D > 0$, take m sufficiently large such that $D_m \leq D + \eta$ and $\varepsilon_m \leq \eta$, where $\eta > 0$ is arbitrarily small. Then:

$$R_{\mathrm{Enc}^*}(D) = \lim_{m \to \infty} R_{\mathrm{Enc}'_m}(D) \leq \lim_{m \to \infty} (R(D_m) + \varepsilon_m) \leq R(D + \eta) + \eta.$$

By the continuity of $R(D)$ (holds for compact metric spaces), letting $\eta \to 0$ yields:

$$\lim_{m \to \infty} R_{\mathrm{Enc}'_m}(D) = R(D).$$

Step 4: Proof of dynamical compatibility. By Axiom D.1.3, the endogenous meta-dynamics Φ is generated by a finite set of generating sheaf vector fields $\{X_1, \ldots, X_r\}$ and is a contraction mapping. Define an energy function:

$$E(\sigma) = \inf_{t \geq 0} d_\mathscr{F}(\sigma, \Phi_t(\sigma)).$$

By contraction, E is continuous and attains its minimum at $t = 0$ or at a fixed point.

For each finite hierarchy encoding Enc_m, consider its dynamics Φ_m on $\mathscr{H}_m$ (an approximation of the restriction of Φ to $\mathscr{H}_m$). Define an optimization problem:

$$\Delta_m = \arg\min_{t \geq 0} \mathbb{E}_{x \sim P_\mathcal{X}}[d_{\mathscr{H}_m}(\mathrm{Enc}_m(x), \Phi_{m,t}(\mathrm{Enc}_m(x)))].$$

By the contraction property of Φ_m and the finite range of Enc_m, the minimum exists.

Let $\mathrm{Enc}_{\Delta,m}(x) = \Phi_{m,\Delta_m}(\mathrm{Enc}_m(x))$. By adjusting Enc_m (e.g., setting $\mathrm{Enc}'_m(x) = \mathrm{Enc}_{\Delta,m}(x)$), we can ensure:

$$\mathbb{E}[d_{\mathscr{H}_m}(\mathrm{Enc}'_m(x), \Phi_{m,\Delta_m}(\mathrm{Enc}'_m(x)))] \leq \eta_m,$$

where $\eta_m \to 0$.

Lifting to $\mathscr{F}_\Omega$: $\mathrm{Enc}'_{\Delta,m} = \iota_m \circ \mathrm{Enc}'_m$, then

$$\mathbb{E}[d_\mathscr{F}(\mathrm{Enc}'_{\Delta,m}(x), \Phi_{\Delta_m}(\mathrm{Enc}'_{\Delta,m}(x)))] \leq \eta_m + \zeta_m \to 0.$$

Taking the limit yields that Enc^* satisfies dynamical compatibility.

Step 5: Cognitive hierarchy realization. By construction, Enc_m is an encoding on $\mathscr{H}_m$, and $\mathrm{Enc}'_m = \iota_m \circ \mathrm{Enc}_m$. Define Enc_{n,m_n} as an appropriate subsequence of Enc_m such that the convergence rates of various indices match. Set $\iota_{n,m_n} = \iota_{m_n}$, then $\mathrm{Enc}'_{m_n} = \iota_{n,m_n} \circ \mathrm{Enc}_{n,m_n}$.

Step 6: Structure preservation. Suppose $\mathcal{X}$ has additional structure, such as an action $\alpha : G \times \mathcal{X} \to \mathcal{X}$ by a compact Lie group G. Since $\mathcal{M}$ has operation generation completeness (Axiom A.1), we can construct approximate sheaf morphisms $\varphi_g : \mathscr{F}_\Omega \to \mathscr{F}_\Omega$ that approximately preserve the G-action, satisfying:

$$d_\mathscr{F}(\varphi_g(\sigma), \varphi_g(\sigma')) \leq L_g d_\mathscr{F}(\sigma, \sigma'),$$

and $\varphi_{gh} \approx \varphi_g \circ \varphi_h$.

When constructing Enc_m, use equivariant vector quantization: require $\mathrm{Enc}_m(\alpha(g, x)) \approx \varphi_g(\mathrm{Enc}_m(x))$. The existence of equivariant vector quantization is guaranteed by the fact that for compact group actions, there exist equivariant partitions and codebook designs. Taking the limit, Enc^* approximately preserves the G-action.

Step 7: Error bound under minimality. When $\mathcal{M}$ satisfies the Minimality Axiom A.2, $k_{\min}$ is the cognitive dimension. By the carrier minimality of $\mathcal{M}$, the number of distinguishable states in $\mathscr{F}_\Omega$ is about $O(2^{ck_{\min}})$ ($c > 0$ constant). By the finite codebook approximation theorem of rate-distortion theory:

$$R_{\mathrm{Enc}}(D) - R(D) \leq \frac{C}{\sqrt{|\mathcal{C}|}},$$

where $|\mathcal{C}|$ is the codebook size. Taking $|\mathcal{C}| \approx 2^{ck_{\min}}$ yields the error bound $O(1/\sqrt{k_{\min}})$.

Take Enc^* as the encoding minimizing $R_{\mathrm{Enc}}(D)$, then for any other Enc:

$$R_{\mathrm{Enc}^*}(D) \leq R_{\mathrm{Enc}}(D) + O\left(\frac{1}{\sqrt{k_{\min}}}\right).$$

□

Remark D.1. *This theorem does not require $\mathcal{M}$ to be isomorphic to $\mathcal{B}_4^k$. It relies only on:*

1. *The infinite cognitive capability of $\mathcal{M}$ (guaranteeing existence and convergence of the cognitive hierarchy).*
2. *The completeness and dynamical structure of the state sheaf $\mathscr{F}_\Omega$.*
3. *The compactness of the input space $\mathcal{X}$ and the prior distribution.*
4. *Operation generation completeness (for structure preservation).*

The "optimality" in the theorem is asymptotic, and the encoding form is not necessarily a "direct mapping" as in the $\mathcal{B}_4$ special case, but may be a complex construction via a limiting process. Nevertheless, the theorem establishes that in a general AGI system, there exists an information-theoretically approximately optimal realization of semantic encoding.

Theorem D.10 (Meta-Dynamics Isomorphism Theorem)**.** *Let $\mathcal{M}_1, \mathcal{M}_2 \in$ **AGI** satisfy Axiom D.1, with their meta-dynamics sheaf paradigms being $\mathfrak{M}_1 = (\mathscr{F}_{\nleq_1}, \Phi_1, \mathrm{Enc}_{\mathscr{F}_1}, \rho_1, \sigma_{0,1})$ and $\mathfrak{M}_2 = (\mathscr{F}_{\nleq_2}, \Phi_2, \mathrm{Enc}_{\mathscr{F}_2}, \rho_2, \sigma_{0,2})$ respectively. If there exists an approximate sheaf isomorphism $\psi : \mathscr{F}_{\nleq_1} \xrightarrow{\approx} \mathscr{F}_{\nleq_2}$ (this isomorphism does not require the carrier set to be B_4^k) that commutes with the dynamics, encoding, and cognitive homomorphism, then $\mathfrak{M}_1$ and $\mathfrak{M}_2$ are equivalent in the sense of semantic representation isomorphism, prediction mechanism conjugacy, and learning dynamics equivalence.*

Proof. Define a commutative diagram:

$$\begin{array}{ccc} \mathscr{F}_{\nleq_1} & \xrightarrow{\psi} & \mathscr{F}_{\nleq_2} \\ \downarrow{\scriptstyle \Phi_1} & & \downarrow{\scriptstyle \Phi_2} \\ \mathscr{F}_{\nleq_1} & \xrightarrow{\psi} & \mathscr{F}_{\nleq_2} \end{array}$$

Similarly, there are commutative diagrams for encoding and cognitive homomorphism. Then:

- Semantic representation isomorphism: $\mathrm{Enc}_{\mathscr{F}_2} = \psi \circ \mathrm{Enc}_{\mathscr{F}_1}$.
- Prediction mechanism conjugacy: $\Phi_2 = \psi \circ \Phi_1 \circ \psi^{-1}$.
- Learning dynamics equivalence: The initial sections satisfy $\sigma_{0,2} = \psi(\sigma_{0,1})$, and trajectories correspond.

Therefore, the two meta-dynamics sheaf paradigms are equivalent. □

Theorem D.11 (Structure and Stability of Cognitive Attractors)**.** *Let $\mathfrak{M}$ be a meta-dynamics sheaf paradigm satisfying Axiom D.1. Then:*

1. *There exists at least one cognitive attractor sheaf $\mathscr{A}^*$ in $\mathscr{F}_\Omega$.*

2. *The structure of $\mathscr{A}^*$ is determined by the generators of Φ.*

3. *The stability of $\mathscr{A}^*$ is characterized by the Lyapunov exponents of Φ.*

4. *The learning process is the search for an initial sheaf section σ_0 whose trajectory enters the basin of attraction of $\mathscr{A}^*$.*

Proof. 1. **Existence Proof**: By Axiom D.1, Φ is a contraction sheaf endomorphism (there exists $L < 1$ such that $d(\Phi(s), \Phi(t)) \leq Ld(s,t)$). Since $\mathscr{F}_\Omega$ is a complete metric space (completeness of state sheaf in Axiom D.1), by the Banach fixed point theorem, there exists a unique fixed point $\sigma^* \in \Gamma(\Omega, \mathscr{F}_\Omega)$ satisfying $\Phi(\sigma^*) = \sigma^*$.

Define the cognitive attractor sheaf $\mathscr{A}^*$ as the stable manifold of σ^*:

$$\mathscr{A}^* = \{s \in \Gamma(\Omega, \mathscr{F}_\Omega) : \lim_{t\to\infty} \Phi_t(s) = \sigma^*\}.$$

By the contraction property, for all $s \in \Gamma(\Omega, \mathscr{F}_\Omega)$, $d(\Phi_t(s), \sigma^*) \leq L^t d(s, \sigma^*)$, hence $\lim_{t\to\infty} d(\Phi_t(s), \sigma^*) = 0$. Therefore, $\mathscr{A}^* = \Gamma(\Omega, \mathscr{F}_\Omega)$, i.e., the entire state space is the basin of attraction. However, finer structure requires consideration of eigenspace decomposition.

2. **Structure Determination Proof**: Φ is generated by a functionally complete set of sheaf vector fields $\{X_1, \ldots, X_r\}$ (Axiom D.1(3)). Linearizing at the fixed point σ^* yields the differential $D\Phi(\sigma^*)$. The structure of $\mathscr{A}^*$ is determined by the kernel and invariant subspaces of $D\Phi(\sigma^*)$:

$$\mathscr{A}^* \cong \ker(D\Phi(\sigma^*) - I) \oplus \bigoplus_{\lambda_i<0} E_{\lambda_i},$$

where E_{λ_i} is the generalized eigenspace corresponding to eigenvalue λ_i. Since Φ is a contraction, all eigenvalues satisfy $|\lambda_i| < 1$, hence $\mathscr{A}^*$ is a stable finite-dimensional manifold.

3. **Stability Characterization Proof**: Near σ^*, consider the linearized system:

$$\frac{d\delta}{dt} = A\delta, \quad \text{where } A = D\Phi(\sigma^*) - I.$$

Define Lyapunov exponents:

$$\lambda_i = \lim_{t\to\infty} \frac{1}{t} \log \|e^{At} v_i\|,$$

where v_i are eigenvectors of A. Since Φ is a contraction, $\mathrm{Re}(\lambda_i) < 0$ for all i. The maximum Lyapunov exponent $\lambda_{\max} = \max_i \mathrm{Re}(\lambda_i)$ determines the stability of the attractor: $\lambda_{\max} < 0$ guarantees exponential stability.

4. **Learning Process Proof**: The learning task is defined as finding σ_0 such that $\Phi_t(\sigma_0)$ converges to σ^*. By the global attractivity of the attractor (Axiom D.1(6) ergodicity), any initial section eventually enters the basin of attraction. Specifically, define the learning algorithm as gradient descent:

$$\sigma_{n+1} = \sigma_n - \eta_n \nabla L(\sigma_n),$$

where $L(\sigma) = d(\Phi_T(\sigma), \sigma^*)^2$. By the contraction property, L is convex, and gradient descent converges to a global minimum σ_0^* with $L(\sigma_0^*) = 0$.

□

Theorem D.12 (Cohomological Classification of Attractors)**.** *The isomorphism class of a cognitive attractor $\mathscr{A}^*$ is completely classified by sheaf cohomology groups:*

$$[\mathscr{A}^*] \in H^2(\lessgtr, \mathscr{F}_{\textit{connection}})$$

Specifically:

1. *Dimension of H^0 = number of connected components of the attractor.*

2. *Non-triviality of H^1 = non-trivial cycles of the attractor.*

3. *Class in H^2 = global twisting degree of the attractor.*

Proof. **Step 1: Cohomology computation of the attractor sheaf.** The cognitive attractor sheaf $\mathscr{A}^* \subseteq \mathscr{F}_{\nleqslant}$ is an approximate sheaf. Consider its sheaf cohomology groups $H^n(\nleqslant, \mathscr{A}^*)$, computed using Čech cohomology. Take a finite open cover $\mathcal{U} = \{U_i\}$ of $\nleqslant$ and construct the Čech complex:

$$0 \to \check{C}^0(\mathcal{U}, \mathscr{A}^*) \to \check{C}^1(\mathcal{U}, \mathscr{A}^*) \to \check{C}^2(\mathcal{U}, \mathscr{A}^*) \to \cdots$$

where $\check{C}^k$ are the k-cochain groups.

Step 2: Geometric interpretation.

- Dimension of H^0: By sheaf cohomology theory, $\dim H^0(\nleqslant, \mathscr{A}^*) = \dim \Gamma(\nleqslant, \mathscr{A}^*)$, i.e., the dimension of the global section space, corresponding to the number of connected components of the attractor.
- Non-triviality of H^1: Non-zero cohomology classes correspond to obstructions to gluing local sections into a global section, which in dynamics manifests as non-trivial cycles (limit cycles) on the attractor.
- Class in H^2: Corresponds to higher-order topological obstructions, such as global twisting of the attractor (Möbius strip structure).

Step 3: Dynamical invariance. Since $\Phi_t(\mathscr{A}^*) = \mathscr{A}^*$, the cohomology groups are invariant under dynamical evolution, hence can serve as classification invariants. □

Theorem D.13 (Realizability of Natural Cognitive Tasks)**.** *Let $\mathfrak{M}$ satisfy Axiom D.1. Then there exists an initial sheaf section $\sigma_0^* \in \Gamma(\Omega, \mathscr{F}_\Omega)$ and a trajectory $\{\Phi_t(\sigma_0^*)\}$ simultaneously satisfying the accuracy goal and the coverage goal.*

Proof. Let a natural cognitive task $\mathfrak{T}$ be characterized by an accuracy goal $\varepsilon_{\text{acc}} > 0$ and a coverage goal $\varepsilon_{\text{cov}} > 0$. We need to prove there exists σ_0^* such that:

1. **Accuracy Goal**: $\limsup_{t\to\infty} d(\Phi_t(\sigma_0^*), \sigma_{\text{target}}) \le \varepsilon_{\text{acc}}$.
2. **Coverage Goal**: $\forall s \in \Gamma(\Omega, \mathscr{F}_\Omega)$, there exists $t \in [0, T]$ such that $d(\Phi_t(\sigma_0^*), s) \le \varepsilon_{\text{cov}}$.

Step 1: Construction of a candidate section family. Since Ω is a complete metric space (Axiom D.1(1)) and $\Gamma(\Omega, \mathscr{F}_\Omega)$ is separable under the cognitive metric, there exists a countable dense subset $D = \{s_1, s_2, \dots\}$. For any $\varepsilon_{\text{cov}} > 0$, there exists a finite subset $D_0 = \{s_1, \dots, s_m\}$ forming an ε_{cov}-approximate net, i.e., for any $s \in \Gamma(\Omega, \mathscr{F}_\Omega)$, there exists $s_i \in D_0$ such that $d(s, s_i) \le \varepsilon_{\text{cov}}$.

Step 2: Achieving the accuracy goal. By known theorem, there exists an attractor $\mathscr{A}^*$ and a fixed point σ^*. Choose σ_0^* such that $\lim_{t\to\infty} \Phi_t(\sigma_0^*) = \sigma^*$. By the contraction property, there exists T_1 such that for all $t > T_1$, $d(\Phi_t(\sigma_0^*), \sigma^*) \le \varepsilon_{\text{acc}}/2$. If $\sigma_{\text{target}} \neq \sigma^*$, the initial section can be adjusted so the trajectory converges near σ_{target}.

Step 3: Achieving the coverage goal. By the ergodicity axiom (Axiom D.1(6)), the trajectory $\{\Phi_t(\sigma_0^*)\}$ is dense in the state space. In particular, for any $s_i \in D_0$, there exists t_i such that $d(\Phi_{t_i}(\sigma_0^*), s_i) \le \varepsilon_{\text{cov}}/2$. Let $T_2 = \max_i t_i$. Then for all $t \ge T_2$, the trajectory visits all ε_{cov}-net neighborhoods.

Step 4: Simultaneous satisfaction. Take $T = \max(T_1, T_2)$. Define an optimization problem:

$$\sigma_0^* = \arg\min_{\sigma\in\Gamma(\Omega,\mathscr{F}_\Omega)} \left[\alpha \cdot \sup_{t\ge T} d(\Phi_t(\sigma), \sigma_{\text{target}}) + (1-\alpha) \cdot \max_{1\le i\le m} \inf_{t\le T} d(\Phi_t(\sigma), s_i)\right]$$

where $\alpha \in (0, 1)$ is a weight parameter. Since Φ is continuous and the domain is complete under the cognitive metric, the minimum exists.

Step 5: Convergence verification. By the contraction property and ergodicity, iterative optimization converges to a solution satisfying both goals. □

Theorem D.14 (Equivalence of Cognitive Hierarchy and Natural Cognitive Task Convergence)**.** *Let $\mathcal{M} \in \mathbf{AGI}$ satisfy the AGI axiom system, $\{\mathscr{H}_n\}$ be a cognitive hierarchy sequence, and $\mathfrak{T}_{nat}$ be a natural cognitive task. Then the following are equivalent:*

1. *Hierarchy Convergence:* $\lim_{n\to\infty} \varepsilon_n^{hier} = 0$.
2. *Task Convergence:* $\lim_{t\to\infty} \varepsilon_t^{nat} = 0$.
3. *Rate Equivalence:* $\varepsilon_n^{hier} = \Theta(\varepsilon_n^{nat})$.

Proof. **Definitions**:

- Hierarchy error: $\varepsilon_n^{\text{hier}} = d(\mathscr{H}_n, \mathscr{H}_\infty)$.
- Task error: $\varepsilon_t^{\text{nat}} = d(\Phi_t(\sigma_0), \sigma_{\text{target}})$.

Proof 1: Hierarchy convergence ⇒ Task convergence. By the definition of cognitive hierarchy, there exist connecting homomorphisms $\mu_n : \mathscr{H}_n \to \mathscr{H}_{n+1}$. Define the task trajectory as a projection sequence: $\Phi_t(\sigma_0) = \text{pr}_t(\mathscr{H}_t)$, where pr_t is the projection to the t-th layer. By hierarchy convergence, $\lim_{n\to\infty} d(\mathscr{H}_n, \mathscr{H}_\infty) = 0$, hence $\lim_{t\to\infty} \varepsilon_t^{\text{nat}} = 0$.

Proof 2: Task convergence ⇒ Hierarchy convergence. Construct a cognitive hierarchy sequence: $\mathscr{H}_n = \{\Phi_0(\sigma_0), \Phi_1(\sigma_0), \ldots, \Phi_n(\sigma_0)\}$. By task convergence, $\lim_{t\to\infty} d(\Phi_t(\sigma_0), \sigma_{\text{target}}) = 0$, thus $\lim_{n\to\infty} \varepsilon_n^{\text{hier}} = 0$.

Proof 3: Rate equivalence. By the Lipschitz property of the cognitive pseudometric (Definition A.1), there exists a constant $L > 0$ such that:

$$\frac{1}{L}\varepsilon_t^{\text{nat}} \leq \varepsilon_t^{\text{hier}} \leq L\varepsilon_t^{\text{nat}}.$$

Therefore $\varepsilon_n^{\text{hier}} = \Theta(\varepsilon_n^{\text{nat}})$. □

Theorem D.15 (Sheaf-Theoretic Version of Learning Dynamics Convergence)**.** *Let $(\mathfrak{M}, \mathcal{D}, \mathcal{A}_{init})$ be a meta-dynamics learning system. Then the learning process converges, and the convergence satisfies an ordinary series convergence condition:*

$$d(\Phi_t(\sigma_0^*), \mathscr{A}^*) \leq C \cdot \frac{1}{t^2}$$

The error sequence satisfies $\sum_{t=1}^{\infty} d(\Phi_t(\sigma_0^*), \mathscr{A}^*) < \infty$.

Proof. The learning system is defined by: - $\mathfrak{M} = (\mathscr{F}_{\not\leq}, \Phi, \text{Enc}_{\mathscr{F}}, \rho, \sigma_0)$: meta-dynamics sheaf paradigm. - $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$: dataset. - $\mathcal{A}_{\text{init}}$: initialization algorithm.

Step 1: Define loss function and gradient flow. Define empirical risk:

$$L(\sigma) = \frac{1}{N}\sum_{i=1}^{N} \ell(\rho(\Phi(\text{Enc}_{\mathscr{F}}(x_i))), y_i),$$

where ℓ is a loss function (e.g., squared loss). The learning process is the gradient flow:

$$\frac{d\sigma}{dt} = -\nabla L(\sigma).$$

Step 2: Convexity analysis. By Axiom D.1, Φ is a contraction mapping, and $\text{Enc}_{\mathscr{F}}$ and ρ are linear (approximate) maps. Assuming ℓ is convex, the composite function L is convex. Furthermore, by the finite generator assumption, the parameter space is a finite-dimensional convex set.

Step 3: Derivation of convergence rate. Use discretized gradient descent: $\sigma_{n+1} = \sigma_n - \eta_n \nabla L(\sigma_n)$. By convex optimization theory, if the step size $\eta_n = \frac{1}{n}$ is chosen, the convergence rate is $O(1/n)$. But a stronger result is needed here.

Consider Nesterov's accelerated gradient method:

$$y_n = \sigma_n + \frac{n-1}{n+2}(\sigma_n - \sigma_{n-1})$$
$$\sigma_{n+1} = y_n - \eta_n \nabla L(y_n)$$

For convex functions, this algorithm achieves $O(1/n^2)$ convergence rate:

$$L(\sigma_n) - L(\sigma^*) \leq \frac{2L\|\sigma_0 - \sigma^*\|^2}{(n+1)^2},$$

where L is the Lipschitz constant of the gradient.

Step 4: Distance estimate. Since L is strongly convex (guaranteed by contraction), there exists $\mu > 0$ such that:

$$\|\sigma_n - \sigma^*\|^2 \leq \frac{2}{\mu}(L(\sigma_n) - L(\sigma^*)).$$

Combined with the rate estimate:

$$\|\sigma_n - \sigma^*\| \leq \sqrt{\frac{4L}{\mu}} \frac{\|\sigma_0 - \sigma^*\|}{n+1} = \frac{C}{n+1}.$$

Therefore $d(\Phi_n(\sigma_0^*), \mathscr{A}^*) = \|\sigma_n - \sigma^*\| \leq C/(n+1)$.

Step 5: Verification of series convergence. The error sequence $e_n = d(\Phi_n(\sigma_0^*), \mathscr{A}^*) \leq C/n$, hence $\sum_{n=1}^{\infty} e_n \leq C \sum_{n=1}^{\infty} 1/n^2 = C \cdot \pi^2/6 < \infty$, satisfying ordinary series convergence.

Step 6: Convergence to attractor. By known theorem, $\mathscr{A}^*$ is the stable manifold of the fixed point σ^*. Since $\sigma_n \to \sigma^*$, the trajectory $\Phi_t(\sigma_0^*)$ enters the basin of attraction of $\mathscr{A}^*$ and then converges to σ^*. □

Definition D.10 (Sheaf Structure of the State Universe)**.** *Let $\mathcal{M} \in$ **AGI**, and $\{\mathbb{H}_n\}$ be a cognitive hierarchy sequence. Define the **state universe** Ω as the topological space $(\mathbb{N}, \tau_{cofinite})$ (cofinite topology). The **cognitive sheaf** $\mathscr{H}$ on Ω is defined as:*

1. *For a cofinite open set $U \subseteq \mathbb{N}$, $\mathscr{H}(U) = \varprojlim_{n \in U} \mathbb{H}_n$.*
2. *$\mathscr{H}(\mathbb{N}) = \mathbb{U} = \varinjlim \mathbb{H}_n$.*
3. *Restriction maps are induced by the connecting approximate homomorphisms μ_n.*

Ω is called the state universe of the AGI system, and $\mathscr{H}$ is its cognitive sheaf.

Theorem D.16 (Sheaf-Theoretic Version of Cohomological Dynamics Interpretation)**.** *Let $H^n(\lesssim, \mathscr{F})$ be the cohomology groups of a sheaf $\mathscr{F}$. Then:*

1. *$H^0(\lesssim, \mathscr{F}) \cong \Gamma(\lesssim, \mathscr{F})$ corresponds to globally consistent cognitive states.*
2. *Non-zero elements of $H^1(\lesssim, \mathscr{F})$ correspond to locally compatible but globally inconsistent cognitive states.*
3. *Non-triviality of $H^2(\lesssim, \mathscr{F})$ corresponds to global topological obstructions of the dynamics.*
4. *Higher cohomology groups vanish when $\lesssim$ has finite dimension.*

Proof. **Background**: Let $\lesssim$ be a topological space, $\mathscr{F}$ a sheaf on $\lesssim$ (valued in Abelian groups). Cohomology groups $H^n(\lesssim, \mathscr{F})$ are defined via Čech complexes or injective resolutions.

Proof 1: Interpretation of H^0. By definition of sheaf cohomology, $H^0(\lesssim, \mathscr{F}) = \Gamma(\lesssim, \mathscr{F})$, the space of global sections. In the cognitive context, global sections correspond to cognitive states consistent across the entire state space $\lesssim$. Existence of global sections implies global coordination of cognitive states.

Proof 2: Interpretation of H^1. $H^1(\lesssim, \mathscr{F})$ classifies obstructions to gluing local sections into a global section. Specifically, consider an open cover $\mathcal{U} = \{U_i\}$ and local sections $s_i \in \mathscr{F}(U_i)$

that agree on overlaps: $s_i|_{U_i \cap U_j} = s_j|_{U_i \cap U_j}$. If there exists a global section s with $s|_{U_i} = s_i$, the corresponding cohomology class is zero. Otherwise, a non-zero class represents cognitive states that are locally consistent but cannot be globalized.

In dynamics, this corresponds to cognitive conflict: local cognitions are compatible but cannot form a globally consistent cognition. For example, different modules have slightly different understandings of the same concept, leading to overall contradiction.

Proof 3: Interpretation of H^2. $H^2(\not\leqslant, \mathscr{F})$ corresponds to higher-order obstructions. Consider local sections $s_{ij} \in \mathscr{F}(U_i \cap U_j)$ satisfying cocycle conditions but not realizable as differences of local sections. In dynamics, this corresponds to global topological obstructions: even if local dynamics are well-defined, global dynamics may not be definable due to topology.

For example, the state space $\not\leqslant$ may have non-trivial topology (e.g., holes). Integration of a dynamical vector field around a hole yields a non-zero cycle, preventing definition of a global potential function. This corresponds to Berry phases or topological obstructions.

Proof 4: Vanishing of higher cohomology. If $\not\leqslant$ is a d-dimensional manifold (d finite) and $\mathscr{F}$ is a flabby or injective sheaf, then for $n > d$, $H^n(\not\leqslant, \mathscr{F}) = 0$. This is because the dimension of the Čech complex is bounded by the nerve complexity of the cover, and a d-dimensional manifold admits covers with nerve dimension d.

In the cognitive context, $\not\leqslant$ is a $2k_{\min}$-dimensional manifold (by the embedding $\varphi : \mathcal{M}_{\text{ext}} \hookrightarrow [0,1]^{2k_{\min}}$), hence $d = 2k_{\min}$ is finite. Therefore, for $n > 2k_{\min}$, $H^n(\not\leqslant, \mathscr{F}) = 0$.

Cognitive Interpretation: - H^0 non-trivial: existence of globally consistent cognition. - H^1 non-trivial: existence of cognitive conflicts that are reconcilable. - H^2 non-trivial: existence of fundamental cognitive obstructions (e.g., paradoxes). - $H^n = 0$ ($n \geq 3$): higher-order cognitive obstructions do not exist in finite-dimensional systems. □

D.4 Necessity of Self-Referential Prediction Module and Unification of Machine Learning Mechanisms

Theorem D.17 (Necessary Existence of Self-Referential Prediction Module)**.** *Let $\mathfrak{M} = (\mathscr{F}_\Omega, \Phi, \mathrm{Enc}_{\mathscr{F}}, \rho, \sigma_0)$ be a meta-dynamics sheaf paradigm, and $\mathcal{M} \in$ **AGI** possess infinite cognitive capability. Then $\mathfrak{M}$ necessarily contains a self-referential prediction module $\mathrm{MetaPred} : \mathscr{F}_\Omega \to \mathscr{F}_\Omega$, defined as*

$$\mathrm{MetaPred}(\sigma) = \rho^{-1}\left(\mathcal{Y}\left(\rho \circ \Phi_{\tau(\sigma)} \circ \rho^{-1}\right)\right)$$

satisfying contraction, self-referentiality, and compatibility.

Proof. By the infinite cognitive capability equivalence theorem B.3, $\mathcal{M}$ possesses an approximate fixed-point combinator $\mathcal{Y}$. Define $\mathrm{MetaPred}$ as:

$$\mathrm{MetaPred}(\sigma) = \rho^{-1}\left(\mathcal{Y}\left(\rho \circ \Phi_{\tau(\sigma)} \circ \rho^{-1}\right)\right)$$

where $\tau(\sigma)$ is a time parameter dependent on state σ.

Verification of properties:

1. **Contraction**: Since $\mathcal{Y}$ is an approximate fixed-point combinator, its output is a fixed point of a contraction mapping.

2. **Self-referentiality**: $\mathrm{MetaPred}$ predicts the behavior of its own dynamics Φ.

3. **Compatibility**: Commutes with Φ, guaranteed by construction.

Therefore, $\mathrm{MetaPred}$ necessarily exists. □

Corollary D.2 (Sheaf-Theoretic Dynamical Realization of Fixed-Point Combinator)**.** *The approximate fixed-point combinator $\mathcal{Y}$ can be realized via fixed points of Φ:*

$$\mathcal{Y}(f) = \lim_{t \to \infty} \Phi_t(\sigma_0(f))$$

where $\sigma_0(f)$ is the initial section corresponding to f.

Proof. By Theorem D.17, the definition of $\mathrm{MetaPred}$ depends on $\mathcal{Y}$. Conversely, $\mathcal{Y}$ can be obtained via fixed-point iteration of Φ. Specifically: given f, construct an initial section $\sigma_0(f)$ such that $\rho(\sigma_0(f)) = f$. Then iterate Φ; the limit is $\mathcal{Y}(f)$. By cognitive hierarchy convergence, this limit exists. □

Theorem D.18 (Existence and Isomorphism Uniqueness of MetaPred)**.** *Let $\mathcal{M} \in \mathbf{AGI}$ satisfy the AGI Fundamental Axiom A.1 and Axiom D.1, and let $\mathfrak{M}$ be any meta-dynamics sheaf paradigm. Then $\mathfrak{M}$ necessarily contains a self-referential prediction module $\mathrm{MetaPred}$, unique up to isomorphism, and it is embedded within any ordinary machine learning mechanism.*

Proof. Existence is guaranteed by Theorem D.17. Uniqueness: if there were two, $\mathrm{MetaPred}_1$ and $\mathrm{MetaPred}_2$, then by the uniqueness of the approximate fixed-point combinator (Theorem B.3), they are approximately equal, hence isomorphic.

Embeddedness: Any ordinary machine learning mechanism can be represented as a meta-dynamics sheaf paradigm, and $\mathrm{MetaPred}$ is part of the paradigm, hence embedded. □

Theorem D.19 (Convergence of MetaPred)**.** *Let $\mathcal{M} \in \mathbf{AGI}$ satisfy the AGI axiom system, and $\mathcal{Y}$ be an approximate fixed-point combinator on $\mathcal{M}$. Consider the iteration sequence $\{F_t\} \subseteq \mathrm{End}\mathcal{M}$ induced by $\mathrm{MetaPred}$. Then the sequence $\{F_t\}$ converges almost surely to some $F^* \in \mathrm{End}\mathcal{M}$, where F^* is a fixed point of $\mathcal{Y}$, the corresponding $\mathrm{MetaPred}_{F^*}$ is a contraction mapping, and the total error decreases monotonically to convergence.*

Proof. We need to prove that the iteration sequence $\{F_t\} \subseteq \mathrm{End}\mathcal{M}$ induced by $\mathrm{MetaPred}$ converges almost surely.

Define the iteration: $F_0 \in \mathrm{End}\mathcal{M}$ arbitrary, $F_{t+1} = \mathcal{Y}(F_t)$, where $\mathcal{Y}$ is the approximate fixed-point combinator.

By Definition B.4, $\mathcal{Y}$ satisfies:

$$d(F_t(\mathcal{Y}(F_t)), \mathcal{Y}(F_t)) \leq \varepsilon(F_t),$$

and $\varepsilon(F) \to 0$ as F approaches the identity mapping.

We need to prove $\{F_t\}$ is a Cauchy sequence. Consider:

$$d(F_{t+1}, F_t) = d(\mathcal{Y}(F_t), \mathcal{Y}(F_{t-1})).$$

By the cognitive stability of $\mathcal{Y}$ (Definition B.4.4), $\mathcal{Y}$ is continuous in the cognitive error metric. Hence, if the variation of $\{F_t\}$ tends to 0, then $\{\mathcal{Y}(F_t)\}$ also tends to stabilize.

More rigorously: By the approximate fixed-point property of $\mathcal{Y}$,

$$d(F_t(F_{t+1}), F_{t+1}) \leq \varepsilon(F_t).$$

If F_t is close to a fixed point, then $\varepsilon(F_t)$ is small.

We can prove: there exists a constant $0 < \alpha < 1$ such that

$$d(F_{t+1}, F_t) \leq \alpha d(F_t, F_{t-1}) + \varepsilon_{t-1},$$

where $\varepsilon_t = \varepsilon(F_t) \to 0$.

By a generalization of the contraction mapping principle, such a sequence is Cauchy.

Since $\mathrm{End}\mathcal{M}$ is complete under the cognitive pseudometric (as $\mathcal{M}$ is finite and operations continuous), the Cauchy sequence converges to some $F^* \in \mathrm{End}\mathcal{M}$.

By continuity of $\mathcal{Y}$, $F^* = \mathcal{Y}(F^*)$, i.e., F^* is a fixed point of $\mathcal{Y}$.

Contraction property of $\mathrm{MetaPred}_{F^*}$: Since F^* is a fixed point, the Lipschitz constant of $\mathrm{MetaPred}_{F^*}$ is determined by the construction of $\mathcal{Y}$ and can be shown to be less than 1.

Monotonic decrease of total error: Define total error $E_t = d(F_t, F^*)$. By the iteration definition and properties of $\mathcal{Y}$, $E_{t+1} \leq \alpha E_t + \varepsilon_t$, where $\alpha < 1$, $\varepsilon_t \to 0$, hence $E_t \to 0$. □

Theorem D.20 (Strong Isomorphism Theorem for AGI Machine Learning Mechanisms)**.** *Let $\mathcal{M} \in$ **AGI** satisfy the AGI axiom system, and ML_1 and ML_2 be any two machine learning mechanisms on $\mathcal{M}$, corresponding to prediction functions $f_1 : \mathcal{X}_1 \to \mathcal{Y}_1$ and $f_2 : \mathcal{X}_2 \to \mathcal{Y}_2$, respectively. Then ML_1 and ML_2 are strictly isomorphic in structure,* $\mathrm{MetaPred}$ *is universally embedded, and all AGI machine learning mechanisms form a single isomorphism class.*

Proof. By the Meta-Dynamics Isomorphism Theorem D.10, the meta-dynamics sheaf paradigm corresponding to any machine learning mechanism is unique up to isomorphism. In particular, $\mathrm{MetaPred}$, as part of the paradigm, is universally embedded. Therefore, all AGI machine learning mechanisms form a single isomorphism class.

□

Theorem D.21 (Universal Synchronous Adjunction Theorem for MetaPred)**.** *Let $\mathcal{M} \in$ **AGI** satisfy the AGI axiom system, and $\mathfrak{T}$ be any instance of a machine learning task on $\mathcal{M}$, realized by its meta-dynamics sheaf paradigm $\mathfrak{M}_{\mathfrak{T}}$. Then there exists a unique adjoint process $\mathfrak{P}_{\mathrm{MetaPred}}$ such that $\mathfrak{P}_{\mathrm{MetaPred}}$ runs synchronously with $\mathfrak{T}$, is embedded within $\mathfrak{M}_{\mathfrak{T}}$, and its necessity holds.*

Proof. Define the adjoint process $\mathfrak{P}_{\mathrm{MetaPred}}$ as, during task execution, invoking $\mathrm{MetaPred}$ at regular intervals to predict the current state and make adjustments. By the embeddedness of $\mathrm{MetaPred}$, it can access the task state. Synchronous operation: for each step (or every K steps) of the task, $\mathrm{MetaPred}$ also executes a step.

Uniqueness: The definition of the adjoint process is uniquely determined by $\mathrm{MetaPred}$.

Necessity: If $\mathrm{MetaPred}$ is lacking, the task may fail to converge or be inefficient, because $\mathrm{MetaPred}$ provides meta-cognitive guidance. □

Corollary D.3 (Sheaf-Theoretic MetaPred-Adjoint Structure of Machine Learning Instances)**.** *The actual execution of any AGI machine learning instance $\mathfrak{T}$ can be decomposed into a task layer and a meta-layer, with the two layers evolving synchronously via coupling equations.*

Proof. The task layer executes the main task; the meta-layer runs $\mathrm{MetaPred}$. Coupling equations:

$$\begin{aligned}\sigma_{\text{task}}^{t+1} &= \Phi_{\Delta}(\sigma_{\text{task}}^{t}) \\ \sigma_{\text{meta}}^{t} &= \mathrm{MetaPred}(\sigma_{\text{task}}^{t}) \\ \sigma_{\text{task}}^{t} &\leftarrow \text{Integrate}(\sigma_{\text{task}}^{t}, \sigma_{\text{meta}}^{t})\end{aligned}$$

where Integrate is an integration function (e.g., weighted average). □

Theorem D.22 (Sheaf-Theoretic Algorithmic Realization Theorem for Adjoint Operation)**.** *Let the implementation algorithm of $\mathfrak{T}$ be $\mathcal{A}_{\mathfrak{T}}$. Then there exists an algorithm $\mathcal{A}_{\mathfrak{T}}^{ac} = \mathrm{InsertMetapredCalls}(\mathcal{A}_{\mathfrak{T}})$, satisfying correctness, efficiency, and necessity.*

Proof. The algorithm $\mathrm{InsertMetapredCalls}$ inserts code that calls $\mathrm{MetaPred}$ within the loops of $\mathcal{A}_{\mathfrak{T}}$. Correctness: insertion does not change the semantics of the original algorithm because $\mathrm{MetaPred}$ is read-only (does not affect state evolution). Efficiency: insertion adds $O(k_{\min}^3)$ overhead per step but can accelerate convergence. Necessity: without insertion, the algorithm may not satisfy convergence requirements. □

Definition D.11 (Sheaf-Theoretic Task Step Function)**.** *Let $\mathfrak{T}$ be a machine learning task. Define TaskStep$_{\mathfrak{T}} : \Gamma(\nleq, \mathscr{F}_{\nleq}) \to \Gamma(\nleq, \mathscr{F}_{\nleq})$ as TaskStep$_{\mathfrak{T}}(\sigma) = \Phi_{\Delta_{\mathfrak{T}}}(\sigma)$.*

Definition D.12 (State Integration Function)**.** *Define Integrate $: \Gamma(\nleq, \mathscr{F}_{\nleq}) \times \Gamma(\nleq, \mathscr{F}_{\nleq}) \to \Gamma(\nleq, \mathscr{F}_{\nleq})$ as Integrate$(\sigma_{task}, \sigma_{meta}) = \alpha \cdot \sigma_{task} + (1 - \alpha) \cdot \sigma_{meta}$, where $\alpha \in [0, 1]$ is the integration coefficient.*

Theorem D.23 (Sheaf-Theoretic Minimum Frequency Theorem for Adjoint Operation)**.** *Let the time complexity of $\mathfrak{T}$ be T. Then* $\mathrm{MetaPred}$ *must run at a minimum frequency $f_{\min} = \Omega(1/\Delta)$, where Δ is the task step size.*

Proof. If $\mathrm{MetaPred}$ runs at a frequency lower than $f_{\min}$, the task may deviate from the correct trajectory, leading to slower convergence or divergence. By stability analysis, the invocation interval of $\mathrm{MetaPred}$ must be less than the reciprocal of the Lyapunov time, i.e., $f_{\min} \sim 1/\Delta$. □

Theorem D.24 (Sheaf-Theoretic Structural Rigidity of Machine Learning Mechanisms)**.** *Let* $\mathcal{M} \in \mathbf{AGI}$*. Then the structure of its machine learning mechanism is rigid: self-referential mandatory, algebraically constrained, topologically invariant, and dynamically unique.*

Proof.

- Self-referential mandatory: By Theorem D.17, $\mathrm{MetaPred}$ must exist.
- Algebraically constrained: By Theorem C.6, the algebraic structure must be $\mathcal{B}_4^{k_{\min}}$.
- Topologically invariant: By Theorem C.10, the Stone space topology is unique.
- Dynamically unique: By Theorem D.10, the meta-dynamics sheaf paradigm is unique.

□

Corollary D.4 (Sheaf-Theoretic Limit Uniqueness and Determinacy of MetaPred)**.** *Under the constraints of direct mapping semantic encoding and the minimality axiom, the prediction mechanism of* $\mathrm{MetaPred}$ *is completely uniquely determined: the Boolean polynomial structure is unique, the contraction factor is unique, and the implementation is non-redundant.*

Proof. In the special case of B_4, the Boolean polynomial corresponding to $\mathrm{MetaPred}$ is unique (determined by $k_{\min}$). The contraction factor is uniquely determined by the Lipschitz constant. Non-redundant implementation is guaranteed by the minimality axiom. □

D.5 Core Properties and Algebraic Structure of MetaPred

Theorem D.25 (Sheaf-Theoretic Three Core Properties of MetaPred)**.** *Let* $\mathcal{M} \in \mathbf{AGI}$ *satisfy the AGI axiom system,* $\mathrm{MetaPred} : \mathscr{F}_{\nleq} \to \mathscr{F}_{\nleq}$ *be the self-prediction module, and* $r^* \in \Gamma(\nleq, \mathscr{F}_{\nleq})$ *be an approximate fixed point of* $\mathrm{MetaPred}$*. For any prediction mechanism* $\mathrm{pred} : \mathscr{P} \to \mathscr{P}$*, define its lift* $P : \mathscr{F}_{\nleq} \to \mathscr{F}_{\nleq}$ *as* $P = \rho^{-1} \circ \mathrm{pred} \circ \rho$ *(where* ρ *is the cognitive homomorphism). Then the following hold:*

1. ***Error Measurability****:* $\mathrm{Err}_{\mathrm{meta}}(\mathrm{pred}) = d(P(r^*), r^*)$ *defines a prediction error measure.*
2. ***Decomposability****:* P *can be decomposed as the composition of* $\mathrm{MetaPred}$ *and a basic prediction function.*
3. ***Commutativity****:* $\mathrm{MetaPred} \circ P \approx P \circ \mathrm{MetaPred}$ *(*ε*-approximately commute).*

Proof. Define enc and dec as ρ and ρ^{-1} (approximately). Then $P = \mathrm{enc} \circ \mathrm{pred} \circ \mathrm{dec}$.

- Error Measurability: Define $\mathrm{Err}_{\mathrm{meta}}(\mathrm{pred}) = d_{\mathcal{M}}(P(r^*), r^*)$. Since r^* is a fixed point of $\mathrm{MetaPred}$, it represents the most stable cognitive state, making the error measure natural.
- Decomposability: In the special case of B_4, P can be decomposed as the composition of $\mathrm{MetaPred}$ and a basic prediction function.
- Commutativity: Because $\mathrm{MetaPred}$ is central (Corollary 6.8), $\mathrm{MetaPred} \circ P \approx P \circ \mathrm{MetaPred}$.

□

Theorem D.26 (Sheaf-Theoretic Mandatory Self-Nesting Structure of MetaPred)**.** *Let* $\mathrm{MetaPred} : \mathcal{M} \to \mathcal{M}$ *satisfy the self-nesting condition. Then* $\mathrm{MetaPred}$ *must have the following refined structure: projection decomposition, algebraic structure, adjoint dynamical realization, quantifiable contraction, and self-correction mechanism.*

Proof. Self-nesting condition: $\mathrm{MetaPred} \circ \mathrm{MetaPred} \approx \mathrm{MetaPred}$.

- Projection decomposition: By idempotence, $\mathrm{MetaPred}$ can be decomposed as a projection onto some subspace.
- Algebraic structure: In the special case of B_4, $\mathrm{MetaPred}$ is a Boolean polynomial.
- Adjoint dynamical realization: By Theorem D.17, $\mathrm{MetaPred}$ is realized via dynamics.
- Quantifiable contraction: Lipschitz constant less than 1.
- Self-correction mechanism: Automatic error correction via fixed-point iteration.

□

Corollary D.5 (MetaPred as the Canonical Measure of Prediction Error)**.** *For any prediction mechanism* pred*, its error can be canonically defined via* $\mathrm{MetaPred}$ *as* $\mathrm{Err}_{\mathrm{meta}}(\mathrm{pred}) = d_{\mathcal{M}}(P(r^*), r^*)$*. This definition satisfies consistency, monotonicity, and additivity.*

Proof.
- Consistency: If pred is perfect, then $P(r^*) = r^*$, error is 0.
- Monotonicity: If pred_1 is more accurate than pred_2, then $d(P_1(r^*), r^*) < d(P_2(r^*), r^*)$.
- Additivity: For composite predictions, errors are additive (approximately).

□

Theorem D.27 (Sheaf-Theoretic Unified Decomposition Structure of Prediction Mechanisms)**.** *Any prediction mechanism* pred *represented on* $\mathcal{M}$ *as* P *has a unique canonical decomposition:*

$$P = \mathrm{MetaPred} \circ P_{base} \circ \mathrm{MetaPred}^{-1}$$

where P_{base} *is a basic prediction function.*

Proof. By the approximate invertibility of $\mathrm{MetaPred}$ (existence of approximate $\mathrm{MetaPred}^{-1}$), define $P_{\text{base}} = \mathrm{MetaPred}^{-1} \circ P \circ \mathrm{MetaPred}$. Then $P = \mathrm{MetaPred} \circ P_{\text{base}} \circ \mathrm{MetaPred}^{-1}$. Uniqueness is guaranteed by the uniqueness of $\mathrm{MetaPred}$. □

Theorem D.28 (Sheaf-Theoretic Topological Meaning of Commutativity)**.** *The commutativity* $\mathrm{MetaPred} \circ P \approx P \circ \mathrm{MetaPred}$ *implies that on the cognitive state space* $\mathcal{M}$*,* $\mathrm{MetaPred}$ *and* P *commute in the sense of preserving attractors, stable manifolds, and eigenspaces.*

Proof. If two maps commute, they share common eigenvector fields. Therefore:

- Attractor preservation: If a set is an attractor of $\mathrm{MetaPred}$, it is also an attractor of P.
- Stable manifold preservation: Stable manifolds remain invariant under both actions.
- Eigenspace preservation: Eigenspaces (spaces spanned by generalized eigenvectors corresponding to eigenvalues) are the same.

□

Corollary D.6 (All Prediction Mechanisms Form an Algebra with MetaPred at its Core)**.** *The set* $\mathcal{P}$ *of all prediction mechanisms represented on* $\mathcal{M}$ *forms an algebra (approximately), where* $\mathrm{MetaPred}$ *is the approximate central element, idempotent element, and zero element.*

Proof. $\mathcal{P}$ is closed under composition and addition (approximately). Since $\mathrm{MetaPred}$ commutes with all $P \in \mathcal{P}$, it is central. Since $\mathrm{MetaPred} \circ \mathrm{MetaPred} \approx \mathrm{MetaPred}$, it is idempotent. Since the fixed point r^* of $\mathrm{MetaPred}$ is the zero element (baseline for error), it approximates the zero element. □

Corollary D.7 (Sheaf-Theoretic Strong Convergence of MetaPred Iteration)**.** *Under the conditions of Theorem D.14,* $\mathrm{MetaPred}$ *iteration satisfies global attractivity, lower bound on convergence rate, error control, and relation to cognitive hierarchy depth.*

Proof. Global attractivity: By the contraction mapping principle, iteration attracts to a unique fixed point.

Convergence rate: In the special case of B_4, superlinear convergence.

Error control: The error sequence is summable.

Relation to cognitive hierarchy depth: The number of iteration steps is proportional to the hierarchy depth $d_{\max}$. □

D.6 Unique Decomposition of AGI Machine Learning Mechanisms

Definition D.13 (Semantic-Process Decomposition of State Universe)**.** *Let* $\mathfrak{M} = (\mathscr{F}_\Omega, \Phi, \mathrm{Enc}_{\mathscr{F}}, \rho, \sigma_0)$ *be a meta-dynamics sheaf paradigm satisfying Axiom D.1.* $\mathscr{F}_\Omega$ *is said to have a* ***semantic-process decomposition*** *if there exists a sheaf isomorphism* $\psi : \mathscr{F}_\Omega \to \mathscr{F}_{sem} \otimes \mathscr{F}_{proc}$*, such that the dynamics* Φ *can be expressed as* $\Phi = (\Phi_{sem}, \Phi_{proc})$*.*

Definition D.14 (Sheaf-Theoretic Minimal Process State Space)**.** *Let* $k_{\min} = \lceil \log_4(|\mathcal{D}_{\min}|) \rceil$*. Define the* ***minimal process state sheaf*** $\mathscr{F}_{proc}$ *as the constant sheaf on the discrete space* $\{1, \ldots, k_{\min}\} \times \{0,1\}^{k_{\min}}$*.*

Definition D.15 (Sheaf-Theoretic Generalized Adjoint Dynamics)**.** *Let* $\mathcal{M} \in \mathbf{AGI}$ *satisfy Axiom D.1, and its state sheaf be decomposed as* $\mathscr{F}_{\not\in} \cong \mathscr{F}_{sem} \otimes \mathscr{F}_{proc}$*. Then its* ***generalized adjoint dynamics*** *is a triple* $(\mathscr{F}_{proc}, \Phi_{proc}, \Psi)$*, where* Φ_{proc} *is the ergodic scheduling dynamics and* Ψ *is the semantic-process interaction sheaf morphism.*

Theorem D.29 (Unique Decomposition Theorem for AGI Machine Learning Mechanisms)**.** *Let* $\mathcal{M} \in \mathbf{AGI}$ *satisfy Axiom D.1. Then there exists an isomorphism class such that its meta-dynamics sheaf paradigm can be decomposed into: semantic-process decomposition of the state sheaf (Definition D.13), adjoint decomposition of the dynamics (Definition D.15), and standard embedding of the self-referential prediction module.*

Proof. **Step 1: Construct the semantic-process decomposition.** By the observation interface decomposition of Axiom D.1, there exists a sheaf isomorphism:

$$\mathscr{F}_\Omega \cong \mathscr{I} \oplus \mathscr{O} \oplus \mathscr{H}.$$

Define the semantic sheaf $\mathscr{F}_{\text{sem}} = \mathscr{H}$ (the hidden state sheaf) and the process sheaf $\mathscr{F}_{\text{proc}}$ as the scheduling information sheaf. Concrete construction: Let $\mathcal{A}$ be a finite parameter category, and $\mathscr{P}$ be the parameter sheaf. By Definition D.3, there exist lifting functors $F_{a,b} : \mathscr{P}_b \to \mathscr{P}_a$. Define $\mathscr{F}_{\text{proc}}$ as the sheaf that records these lifting histories, i.e., for each open set $U \subseteq \mathcal{A}$, $\mathscr{F}_{\text{proc}}(U)$ is the set of all lifting paths.

Define a sheaf isomorphism $\psi : \mathscr{F}_\Omega \to \mathscr{F}_{\text{sem}} \otimes \mathscr{F}_{\text{proc}}$ as follows: For any section $\sigma \in \Gamma(\mathscr{F}_\Omega)$, decompose it into semantic content $\sigma_{\text{sem}} = \mathrm{pr}_{\mathscr{H}}(\sigma)$ and process information $\sigma_{\text{proc}} = (\mathrm{pr}_{\mathscr{I}}(\sigma), \mathrm{pr}_{\mathscr{O}}(\sigma), \{\text{lifting history}\})$. By the cognitive homomorphism realization of Axiom D.1, this decomposition is an approximate isomorphism.

Step 2: Decompose the dynamics. By the generator decomposition of the endogenous meta-dynamics Φ (Axiom D.1.3), we can write Φ as:

$$\Phi = (\Phi_{\text{sem}}, \Phi_{\text{proc}}),$$

where Φ_{sem} acts on $\mathscr{F}_{\text{sem}}$ and Φ_{proc} acts on $\mathscr{F}_{\text{proc}}$. Specifically, Φ_{sem} is generated by semantic generators $\{X_1, \ldots, X_r\}$, and Φ_{proc} is generated by scheduling generators (such as attention reallocation operators).

Step 3: Embed the self-referential prediction module. By Theorem D.17, $\mathrm{MetaPred}$ necessarily exists. Define its representation on the decomposed space:

$$\mathrm{MetaPred} = \psi \circ \mathrm{MetaPred}_{\mathscr{F}} \circ \psi^{-1},$$

where $\mathrm{MetaPred}_{\mathscr{F}}$ is the self-referential prediction module on $\mathscr{F}_\Omega$. By Theorem D.18, this embedding is unique up to isomorphism.

Step 4: Verify uniqueness. Suppose there exists another decomposition $(\mathscr{F}'_{\text{sem}}, \mathscr{F}'_{\text{proc}}, \Phi', \mathrm{MetaPred}')$. By the uniqueness of sheaf isomorphism (in an approximate

sense), there exists an approximate isomorphism $\varphi : \mathscr{F}_{\text{sem}} \otimes \mathscr{F}_{\text{proc}} \to \mathscr{F}'_{\text{sem}} \otimes \mathscr{F}'_{\text{proc}}$ that makes the diagram commute. Therefore, the decomposition is unique up to isomorphism. □

Definition D.16 (Ergodicity Axiom, Sheaf Version)**.** *A meta-dynamics sheaf paradigm $(\mathscr{F}_{\not\leqslant}, \Phi, \mathrm{Enc}_{\mathscr{F}}, \rho, \sigma_0)$ is said to satisfy the **ergodicity axiom** if for any open set $U \subseteq \not\leqslant$ and any section $s \in \Gamma(\not\leqslant, \mathscr{F}_{\not\leqslant})$, the asymptotic time average of the trajectory $\{\Phi_t(s)\}$ over U equals the spatial average, i.e., for any continuous function $f : \Gamma(\not\leqslant, \mathscr{F}_{\not\leqslant}) \to \mathbb{R}$,*

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T f(\Phi_t(s))dt = \int_{\Gamma(\not\leqslant, \mathscr{F}_{\not\leqslant})} f(s)d\mu(s),$$

where μ is an invariant probability measure on $\Gamma(\not\leqslant, \mathscr{F}_{\not\leqslant})$. In particular, this guarantees that the trajectory is dense in the state space.

Theorem D.30 (Generalized Adjoint Dynamics Cognitive Completeness, Sheaf Version)**.** *Let $\mathcal{M} \in$ **AGI** satisfy the AGI axiom system, and let $(\mathscr{F}_{proc}, \Phi_{proc}, \Psi)$ be its generalized adjoint dynamics. Then for any cognitive task $\mathfrak{T}$, there exists an initial process sheaf section and a scheduling strategy such that the task can be completed; all scheduling mechanisms satisfying ergodicity are equivalent in cognitive capability; there exists an optimal scheduling strategy that minimizes cognitive energy consumption; and the optimal strategy can be obtained via a meta-learning process.*

Proof.
- Completable: Since Φ_{proc} is ergodic, it can schedule all necessary processes;
- Equivalence: Ergodicity ensures that all scheduling mechanisms eventually cover the same states;
- Optimal strategy: Define an energy function and minimize it;
- Meta-learning: Learn the optimal scheduling via $\mathrm{MetaPred}$.

□

Theorem D.31 (Variational Principle for Ergodic Scheduling)**.** *The optimal ergodic scheduling dynamics Φ^*_{proc} is the solution to the following variational problem:*

$$\min_{\Phi_{proc}} \{\mathbb{E}[T_{cover}] + \lambda \cdot \mathrm{Var}(T_{switch})\}$$

where T_{cover} is the time to cover all modalities, and T_{switch} is the modality switching interval.

This problem has a unique solution, and the solution takes the form:

$$\Phi^*_{proc}(c_t, \mathbf{v}_t) = \mathrm{Softmax}\left(\frac{Q(c, \mathbf{v})}{\tau}\right)$$

where the Q function satisfies the Bellman equation.

Theorem D.32 (Cognitive Elegance Theorem for Adjoint Dynamics, Sheaf Version)**.** *The generalized adjoint dynamics mechanism necessarily derived from the AGI axiom system is the unique structure that satisfies completeness, minimality, self-consistency, universality, cognitive naturalness, algebraic harmony, and topological elegance.*

Proof.
- Completeness: Can complete all cognitive tasks;
- Minimality: $k_{\min}$ is minimal;
- Self-consistency: Consistent with the AGI axioms;
- Universality: Applicable to any $\mathcal{M}$;
- Cognitive naturalness: Similar to human cognitive scheduling;
- Algebraic harmony: Compatible with Boolean algebraic structures;

- Topological elegance: Stone duality.

□

Corollary D.8 (MetaPred Generalized Adjoint Realization, Sheaf Version)**.** *Under the generalized adjoint dynamics framework,* MetaPred *can be realized via*

$$\mathrm{MetaPred}(\sigma) = \rho\left(\lim_{t\to\infty} \Psi(\sigma_t, \mathrm{Enc}_{sem}(\sigma))\right)$$

where σ_t is the process trajectory.

Proof. MetaPred needs to integrate semantic and process information. By traversing the process trajectory σ_t and interacting with Ψ, the limit yields a stable prediction. □

D.7 Dual Symmetry of Semantic-Dynamics in AGI Machine Learning Mechanisms

Theorem D.33 (Unification Principle for AGI Machine Learning Mechanisms)**.** *Let $\mathcal{M} \in \mathbf{C}_{\mathrm{AGI}}$ satisfy Axiom A.1 (AGI Fundamental Axiom) and Axiom D.1 (Sheaf Formulation of the Meta-Axiom System for Machine Learning). Let $\mathfrak{T}$ be an arbitrary machine learning task on $\mathcal{M}$, defined by an input space $\mathcal{X}$ (compact metric space), output space $\mathcal{Y}$, and loss function $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$. Then the optimal realization mechanism for $\mathfrak{T}$ must be a unified triple $\mathfrak{U}^* = (\mathscr{F}_\Omega, \Phi^*, \mathrm{Enc}^*)$ with the following properties:*

1. ***State Space Unification:*** *$\mathscr{F}_\Omega$ is the state universe sheaf (Definition D.10), serving as the unique carrier space for semantic representation and dynamical evolution. Specifically, $\mathscr{F}_\Omega$ is the completion of the cognitive sheaf $\mathscr{H}$ (Axiom D.1.1), and there exists an observation interface decomposition $\mathscr{F}_\Omega \cong \mathscr{I} \oplus \mathscr{O} \oplus \mathscr{H}$.*

2. ***Dynamics Unification:*** *$\Phi^* : \mathscr{F}_\Omega \to \mathscr{F}_\Omega$ is the optimal realization of the endogenous meta-dynamics (Axiom D.1.3), simultaneously responsible for:*

   (a) *State transition and prediction: There exists an optimal time step $\Delta^* > 0$ such that for any prediction mechanism $\mathrm{Pred} : \mathscr{F}_\Omega \to \mathscr{F}_\Omega$, we have $\mathrm{Pred} \approx \rho \circ \Phi^*_{\Delta^*} \circ \rho^{-1}$, where ρ is the optimal cognitive homomorphism (Theorem D.6);*

   (b) *Self-referential prediction:* MetaPred *can be constructed as the fixed-point iteration process of Φ^*, i.e., $\mathrm{MetaPred}(\sigma) = \lim_{n\to\infty}(\Phi^*)^n(\sigma)$, and this limit exists (Theorem D.19).*

3. ***Encoding-Initialization Unification:*** *$\mathrm{Enc}^* : \mathcal{X} \to \mathscr{F}_\Omega$ is the optimal semantic encoding, satisfying:*

   (a) *There exists a decoding map $\mathrm{Dec}^* : \mathscr{F}_\Omega \to \mathcal{Y}$ and an optimal time $T^* \geq 0$ such that the composed map $\mathrm{Dec}^* \circ \Phi^*_{T^*} \circ \mathrm{Enc}^*$ achieves the minimum expected loss on task $\mathfrak{T}$:*

   $$\mathfrak{U}^* = \arg\min_{\mathrm{Enc},\Phi,\mathrm{Dec},T} \mathbb{E}_{(x,y)\sim P_{\mathcal{X}\times\mathcal{Y}}}\left[\mathcal{L}\left(y, \mathrm{Dec} \circ \Phi_T \circ \mathrm{Enc}(x)\right)\right];$$

   (b) *$\mathrm{Enc}^*(\mathcal{X})$ lies within the basin of attraction of the cognitive attractor $\mathscr{A}^*$ (Theorem D.11);*

   (c) *Enc^* is asymptotically optimal (Theorem D.9).*

Under this unified framework, the traditional separation of semantic encoding Enc^* and predictive dynamics Φ^* is not only suboptimal but also mathematically impossible: any mechanism attempting to decompose them into independent components would violate the self-referential closure required by infinite cognitive capacity (Definition B.4 condition 5).

Proof. **Proof Framework:** We combine proof by contradiction and constructive proof, deriving the necessity of the unified structure from the mathematical requirements of infinite cognitive capacity.

Part 1: Proof of State Space Unification

Suppose there exist two distinct state spaces $\mathscr{F}_1$ and $\mathscr{F}_2$ carrying semantic representation and dynamical evolution, respectively. By Axiom D.1.1, the state sheaf $\mathscr{F}_\Omega$ is the completion of the cognitive sheaf $\mathscr{H}$. Since $\mathscr{H}$ is constructed via the cognitive hierarchy sequence $\{\mathbb{H}_n\}$ (Definition A.13), and the construction of this sequence relies on the approximate homomorphisms between $\mathcal{M}$ and each hierarchy (Axiom A.1.1), it forces all cognitive states (including semantic representations and dynamical states) to reside in the same limit space $\varinjlim \mathbb{H}_n = \mathscr{F}_\Omega$.

More formally: Let $\pi : \mathcal{M} \to \mathscr{H}$ be the perceptual approximate homomorphism, and $\sigma : \mathscr{H} \to \mathcal{M}$ be the action approximate homomorphism. By the approximate invertibility in Definition A.15, there exists ψ such that $\sigma \circ \pi \approx \mathrm{id}_{\mathcal{M}}$. Now, if semantic representation is in $\mathscr{F}_1$ and dynamical states in $\mathscr{F}_2$, then to establish self-reference (the system's cognition of its own representation), we need maps $f : \mathscr{F}_1 \to \mathscr{F}_2$ and $g : \mathscr{F}_2 \to \mathscr{F}_1$. But by Axiom A.1.2 (infinite meta-cognition), the cognitive hierarchy $\mathbb{H}_{n+1}$ contains approximate homomorphisms of $\mathbb{H}_n$. By induction, the limit $\mathscr{F}_\Omega$ of all hierarchies must simultaneously contain representations of all hierarchies. In particular, both $\mathscr{F}_1$ and $\mathscr{F}_2$ must embed into $\mathscr{F}_\Omega$. By the uniqueness of the limit (up to isomorphism), $\mathscr{F}_1 \cong \mathscr{F}_2 \cong \mathscr{F}_\Omega$. Therefore, the state space is unified.

Part 2: Proof of Dynamics Unification

Let $\Phi_{\text{sem}} : \mathscr{F}_\Omega \to \mathscr{F}_\Omega$ be the semantic evolution dynamics, and $\Phi_{\text{pred}} : \mathscr{F}_\Omega \to \mathscr{F}_\Omega$ be the prediction dynamics. We need to prove that there exists a unified Φ^* that simultaneously realizes both functions.

By Theorem D.6, any prediction mechanism Pred can be expressed as $\mathrm{Pred} = \rho \circ \Phi_\Delta \circ \rho^{-1}$, where Φ is the endogenous meta-dynamics and Δ is a time step. In particular, for semantic evolution, we can also define $\mathrm{SemEvol} = \rho' \circ \Phi_{\Delta'} \circ (\rho')^{-1}$. But Axiom D.1.3 states that the endogenous meta-dynamics Φ is generated by a finite set of generator sheaf vector fields $\{X_1, \ldots, X_r\}$. If there were two different dynamics Φ_1 and Φ_2, then their corresponding generator sets $\{X_i^{(1)}\}$ and $\{X_i^{(2)}\}$ must both be functionally complete (by the operational generation completeness of Axiom A.1). But since the generators are finite and $\mathscr{F}_\Omega$ is complete, there exists a unified generator set $\{X_i^*\}$ whose linear span of function spaces contains the linear spans of $\{X_i^{(1)}\}$ and $\{X_i^{(2)}\}$. By the minimality of generators (Axiom A.2.2), this unified set is unique. Therefore, Φ_1 and Φ_2 are actually different time integrals of the dynamics parameterized by the same generator set. In particular, we can define Φ^* as the dynamics generated by this unified generator set and choose appropriate time parameters such that $\Phi^*_{\Delta_1} \approx \Phi_1$ and $\Phi^*_{\Delta_2} \approx \Phi_2$.

For the self-referential prediction module $\mathrm{MetaPred}$, by Theorem D.17, it must exist and have a fixed-point form. Consider the fixed-point iteration of Φ^*: Since Φ^* is a contraction map (Axiom D.1.3), by the Banach fixed-point theorem, for any initial state σ, the sequence $\{\Phi^{*n}(\sigma)\}$ converges to a unique fixed point σ^*. Define $\mathrm{MetaPred}(\sigma) = \sigma^*$. We need to verify that this definition satisfies the properties of $\mathrm{MetaPred}$ (Theorem D.25):

- Error measurability: $d(\mathrm{MetaPred}(\sigma), \Phi^*(\mathrm{MetaPred}(\sigma))) = 0$;
- Self-nesting: $\mathrm{MetaPred} \circ \mathrm{MetaPred} = \mathrm{MetaPred}$ (since iterating a fixed point yields the same fixed point);
- Commutativity with arbitrary prediction mechanism P: Since $P \approx \rho \circ \Phi^*_\Delta \circ \rho^{-1}$ and $\mathrm{MetaPred}$ is the fixed point of Φ^*, they commute in an approximate sense.

Thus, Φ^* unifies semantic evolution, prediction, and self-referential prediction.

Part 3: Proof of Encoding-Initialization Unification

Let $\mathrm{Enc} : \mathcal{X} \to \mathscr{F}_\Omega$ be the semantic encoding, and $\mathrm{Dec} : \mathscr{F}_\Omega \to \mathcal{Y}$ be the decoder. Define the task loss as:

$$J(\mathrm{Enc}, \Phi, \mathrm{Dec}, T) = \mathbb{E}_{(x,y)\sim P}\left[\mathcal{L}\left(y, \mathrm{Dec} \circ \Phi_T \circ \mathrm{Enc}(x)\right)\right].$$

We need to prove that there exists an optimal quadruple $(\mathrm{Enc}^*, \Phi^*, \mathrm{Dec}^*, T^*)$ that minimizes J, and that Enc^* and Φ^* satisfy the stated properties.

Existence: Since $\mathscr{F}_\Omega$ is a complete metric space (Axiom D.1.1), $\mathcal{X}$ is a compact metric space, and $\mathcal{Y}$ is typically a metric space. Consider the function spaces:

- $\mathcal{E} = \{\text{Enc} : \mathcal{X} \to \mathscr{F}_\Omega \mid \text{Enc continuous and image bounded}\}$;
- $\mathcal{D} = \{\Phi : \mathscr{F}_\Omega \to \mathscr{F}_\Omega \mid \Phi \text{ contraction with Lipschitz constant } L < 1\}$;
- $\mathcal{R} = \{\text{Dec} : \mathscr{F}_\Omega \to \mathcal{Y} \mid \text{Dec continuous}\}$;
- $T \in [0, \infty)$.

Under appropriate topologies (e.g., compact-open topology), these spaces are compact (a variant of the Arzelà-Ascoli theorem). The loss function $\mathcal{L}$ is continuous, so J is continuous. By compactness, a minimum exists; denote a point achieving the minimum as $(\text{Enc}^*, \Phi^*, \text{Dec}^*, T^*)$.

Basin of attraction property: Suppose $\text{Enc}^*(\mathcal{X})$ is not within the basin of attraction of the cognitive attractor $\mathscr{A}^*$. Then there exists $x \in \mathcal{X}$ such that $\lim_{t\to\infty} \Phi_t^*(\text{Enc}^*(x)) \notin \mathscr{A}^*$. Since $\mathscr{A}^*$ is globally attracting (Theorem D.11), this is impossible unless Φ^* is not a contraction. But Φ^* comes from $\mathcal{D}$, so it is a contraction. Therefore, $\text{Enc}^*(\mathcal{X})$ must be within the basin of attraction of $\mathscr{A}^*$.

Asymptotic optimality: By Theorem D.9, there exists an asymptotically optimal encoding sequence $\{\text{Enc}_n\}$. Since the minimum exists, the minimizer Enc^* must be in the closure of this sequence and thus is asymptotically optimal.

Part 4: Proof of Indecomposability (by contradiction)

Assume that the optimal mechanism $\mathfrak{U}^* = (\mathscr{F}_\Omega, \Phi^*, \text{Enc}^*)$ can be decomposed into two independent components: $\text{Enc}^* = \text{Enc}_1 \circ \text{Enc}_2$, where $\text{Enc}_2 : \mathcal{X} \to \mathscr{S}$ maps to a "pure semantic" space $\mathscr{S}$, and $\text{Enc}_1 : \mathscr{S} \to \mathscr{F}_\Omega$ maps semantics to the state space; simultaneously, $\Phi^* = \Phi_2 \circ \Phi_1$, where $\Phi_1 : \mathscr{F}_\Omega \to \mathscr{D}$ maps to a "pure dynamics" space $\mathscr{D}$, and $\Phi_2 : \mathscr{D} \to \mathscr{F}_\Omega$.

Now consider the self-reference requirement: The system must be able to cognize its own state. In particular, it needs to think about the semantic representation $\text{Enc}_2(x)$. This implies the existence of a map $f : \mathscr{S} \to \mathscr{D}$ such that the thinking process can be represented. But to maintain infinite cognitive capacity (Theorem B.3), we must have an approximate fixed-point combinator $\mathcal{Y}$ acting on $\text{End}(\mathscr{F}_\Omega)$. If $\mathscr{S}$ and $\mathscr{D}$ are proper subspaces, then $\mathcal{Y}$ cannot simultaneously act on both unless $\mathscr{S} \cong \mathscr{D} \cong \mathscr{F}_\Omega$. Therefore, the decomposition violates self-referential closure.

More formally: By Definition B.4 condition 5 (structure-independent constructiveness), the construction of the approximate fixed-point combinator $\mathcal{Y}$ does not depend on the specific algebraic structure. If there exists a proper subspace $\mathscr{S} \subsetneq \mathscr{F}_\Omega$ carrying semantics, then the restriction of $\mathcal{Y}$ to $\mathscr{S}$ might not exist (because $\mathscr{S}$ may not be closed under operations). But infinite cognitive capacity requires $\mathcal{Y}$ to exist for all endomorphisms. Contradiction.

Therefore, the optimal mechanism must be unified and cannot be decomposed into independent semantic encoding and predictive dynamics components. □

Remark D.2. *The key points of this proof are:*

1. *Infinite cognitive capacity forces self-referential closure, which requires the representation space and the dynamical state space to coincide;*
2. *The minimality of the generators of endogenous meta-dynamics ensures that semantic evolution and prediction can be realized by the same dynamics;*
3. *Compactness arguments guarantee the existence of an optimal solution;*
4. *Proof by contradiction shows how decomposition would break self-reference.*

The entire proof does not rely on the assumption $M \cong \mathcal{B}_4^k$, but only on Axioms A.1, A.2, D.1 and related definitions.

Definition D.17 (Complete Structure of Unified Cognitive Mechanism)**.** *Let $\mathcal{M} \in \mathbf{C}_{\text{AGI}}$ satisfy the AGI axiom system. A **unified cognitive mechanism** on it is a quintuple $\mathfrak{U} = (\mathscr{F}_\Omega, \mathscr{G}, \Phi, \mathcal{Y}, \mathcal{A}^*)$, where:*

1. ***State sheaf** $\mathscr{F}_\Omega$: Defined as the sheaf of sections of a fiber bundle $\pi : E \to \Omega$, where:*

- *The base space $\Omega = \bigcup_{n=0}^{\infty} \Omega_n$ is a stratified state universe, with Ω_n corresponding to the n-th level of cognition;*
- *The fiber $F_x = \mathbb{R}^{d(x)}$, with $d : \Omega \to \mathbb{N}$ a locally constant function;*
- *The section space $\Gamma(\Omega, \mathscr{F}_\Omega)$ is equipped with a Sobolev-type metric:*

$$\langle \sigma, \tau \rangle_{\mathscr{F}} = \int_\Omega \left(\langle \sigma(x), \tau(x) \rangle_{F_x} + \lambda \langle \nabla\sigma(x), \nabla\tau(x) \rangle \right) d\mu(x),$$

where μ is a cognitive measure on Ω.

2. ***Generator sheaf*** *$\mathscr{G}$: A set of sheaf vector fields on $\mathscr{F}_\Omega$, $\mathscr{G} = \{X_1, \ldots, X_r\}$, satisfying:*
   (a) ***Functional completeness:*** *$\mathrm{Lie}(\mathscr{G})$ is dense in $\Gamma(\mathscr{F}_\Omega)$;*
   (b) ***Stratified structure:*** *Each X_i can be decomposed as $X_i = \sum_{n=0}^{\infty} X_i^{(n)}$, where $X_i^{(n)}$ acts on the n-th cognitive level;*
   (c) ***Cognitive symmetry:*** *There exists a Lie algebra homomorphism $\varphi : \mathfrak{g} \to \mathrm{Der}(\mathscr{G})$, where $\mathfrak{g}$ is some Lie algebra reflecting cognitive symmetries (e.g., attention reallocation, concept reorganization).*
3. ***Unified dynamics*** *$\Phi : \mathbb{R}_{\geq 0} \times \mathscr{F}_\Omega \to \mathscr{F}_\Omega$: The development map generated by the generator sheaf:*

$$\frac{\partial \sigma_t}{\partial t} = \sum_{i=1}^{r} a_i(\sigma_t, t) X_i(\sigma_t),$$

   where the coefficients $a_i : \mathscr{F}_\Omega \times \mathbb{R}_{\geq 0} \to \mathbb{R}$ satisfy:
   (a) ***Adaptive equation:*** *$a_i(\sigma, t) = \mathrm{sigmoid}\left(\langle w_i, \mathrm{Enc}_{meta}(\sigma) \rangle + b_i(t)\right)$, where w_i are weight vectors and $b_i(t)$ are biases (learnable);*
   (b) ***Normalization:*** *$\sum_{i=1}^{r} a_i(\sigma, t) = 1$ (ensuring probabilistic interpretation);*
   (c) ***Contraction:*** *There exists $L < 1$ such that $\|\Phi_t(\sigma) - \Phi_t(\sigma')\| \leq L\|\sigma - \sigma'\|$.*
4. ***Fixed-point combinator*** *$\mathcal{Y} : \mathrm{End}(\mathscr{F}_\Omega) \to \mathscr{F}_\Omega$: Defined as:*

$$\mathcal{Y}(f) = \lim_{n \to \infty} \Phi_{t_n}(f^{(n)}(\sigma_0)),$$

   where $\{t_n\}$ is an increasing time sequence, $f^{(n)}$ is the n-th iterate of f, and σ_0 is a standard initial state. $\mathcal{Y}$ satisfies all conditions of Definition B.4.
5. ***Cognitive attractor sheaf*** *$\mathcal{A}^* = \{\mathscr{A}_\alpha^*\}_{\alpha \in A}$: A family of subsheaves satisfying:*
   (a) *Each $\mathscr{A}_\alpha^*$ is Φ-invariant;*
   (b) *Global attractiveness: For any $\sigma \in \mathscr{F}_\Omega$, there exists α such that $\lim_{t \to \infty} \mathrm{dist}(\Phi_t(\sigma), \mathscr{A}_\alpha^*) = 0$;*
   (c) *Minimality: Each $\mathscr{A}_\alpha^*$ contains no proper sub-attractor;*
   (d) *Cognitive interpretation: $\mathscr{A}_\alpha^*$ corresponds to fundamental cognitive modes (e.g., object recognition, language understanding, reasoning patterns).*

Theorem D.34 (Concrete Realization of Unified Cognitive Mechanism)**.** *The structure in Definition D.17 is the unique one satisfying the following optimization criteria:*

$$\min_{\mathscr{G}, a_i, \mathcal{Y}} \quad \mathbb{E}_{(x,y)\sim P}\left[\mathcal{L}(y, \mathrm{Dec} \circ \Phi_T \circ \mathrm{Enc}(x))\right] \tag{7}$$

$$\text{s.t.} \quad \text{Cognitive energy minimization: } E_{\mathrm{cog}}(\sigma) = \|\sigma - \mathcal{Y}(\mathrm{id})\|^2 + \lambda R(\sigma) \tag{8}$$

$$\text{Information bottleneck: } I(x; \mathrm{Enc}(x)) \leq I_c, \quad I(\mathrm{Enc}(x); y) \geq I_y \tag{9}$$

$$\text{Self-referential consistency: } \mathrm{MetaPred} \circ \mathrm{Enc} \approx \mathrm{Enc} \circ \mathrm{MetaPred}_{\mathcal{X}} \tag{10}$$

where $\mathrm{MetaPred}_{\mathcal{X}} : \mathcal{X} \to \mathcal{X}$ is the meta-prediction on the input space.

The optimal solution has the following concrete forms:

1. ***Concrete construction of generators:*** *Let the local coordinates of $\mathscr{F}_\Omega$ be $(x^1, \ldots, x^d)$. Then the generators can be expressed as:*

$$X_i = \sum_{j=1}^{d} f_i^j(x) \frac{\partial}{\partial x^j} + \sum_{j,k=1}^{d} g_i^{jk}(x) \frac{\partial^2}{\partial x^j \partial x^k},$$

where f_i^j and g_i^{jk} are polynomial functions (by the Stone-Weierstrass theorem, any continuous function can be approximated by polynomials).

2. ***Equation for coefficient functions:*** *$a_i(\sigma, t)$ satisfies the Hamilton-Jacobi-Bellman equation:*

$$\frac{\partial a_i}{\partial t} + H(\sigma, \nabla a_i) = 0,$$

where the Hamiltonian function is $H(\sigma, p) = \max_{u \in U} \{\langle p, \sum_i u_i X_i(\sigma) \rangle - C(\sigma, u)\}$, C is a cost function, and U is the control set.

3. ***Explicit formula for encoding:*** *The optimal encoding $\mathrm{Enc}^* : \mathcal{X} \to \mathscr{F}_\Omega$ can be expressed as:*

$$\mathrm{Enc}^*(x) = \sum_{k=1}^{K} c_k(x) \sigma_k^*,$$

where $\{\sigma_k^*\}$ is a basis of $\mathcal{A}^*$ (characteristic modes of the attractors), $c_k(x)$ are coefficients satisfying:

$$c_k(x) = \frac{\exp(\langle w_k, \psi(x) \rangle)}{\sum_{j=1}^{K} \exp(\langle w_j, \psi(x) \rangle)},$$

$\psi : \mathcal{X} \to \mathbb{R}^m$ is a feature mapping, and w_k are learnable weights.

4. ***Discrete-time realization of dynamics:*** *In computer implementation, Φ can be discretized as:*

$$\sigma_{n+1} = \sigma_n + \epsilon \sum_{i=1}^{r} a_i(\sigma_n, n\epsilon) X_i(\sigma_n) + \sqrt{2\epsilon} \xi_n,$$

where ϵ is the step size and ξ_n is noise (promoting exploration). This is a cognitive version of Langevin dynamics.

5. *Algorithmic realization*

Algorithm 2: Algorithmic Realization of MetaPred

Function `MetaPred(`σ`)`**:**

$\sigma_0 \leftarrow \sigma$;
for $n \leftarrow 0$ **to** $N - 1$ **do**
  $g_n \leftarrow \nabla_{\sigma_n} E_{\mathrm{cog}}(\sigma_n)$;
  $\sigma_{n+1} \leftarrow \sigma_n - \eta_n g_n + \sqrt{2\eta_n}\, \xi_n$;
return σ_N;

where E_{cog} is the cognitive energy function.

6. ***Learning rule:*** *Parameters are updated via stochastic gradient descent:*

$$\theta_{t+1} = \theta_t - \eta_t \left(\nabla_\theta \mathcal{L}_t + \lambda_1 \nabla_\theta E_{\mathrm{cog}}(\sigma_t) + \lambda_2 \nabla_\theta I(x_t; \mathrm{Enc}(x_t)) \right),$$

where θ includes all learnable parameters (those of $f_i^j, g_i^{jk}, w_k, a_i$, etc.).

Structure Derivation. We derive the concrete structure from the optimization problem (7)-(10) using variational calculus and optimal control theory.

Step 1: Derivation of generator structure

Consider the dynamics $\dot{\sigma} = \sum_i a_i X_i(\sigma)$. We wish to choose X_i such that all possible cognitive evolutions can be expressed. By the Chow-Rashevskii theorem, if $\{X_i\}$ satisfies the Lie algebra rank

condition, then the system can reach any point from any other point (controllability). In the cognitive context, we wish to evolve from any cognitive state to any other (cognitive completeness). Therefore, $\{X_i\}$ must span the entire tangent space $T_\sigma \mathscr{F}_\Omega$.

In local coordinates, vector fields can be expressed as $X_i = \sum_j f_i^j \frac{\partial}{\partial x^j}$. To handle stochasticity (cognitive uncertainty), we add diffusion terms $\sum_{j,k} g_i^{jk} \frac{\partial^2}{\partial x^j \partial x^k}$. By Hörmander's condition, if $\{X_i\}$ and their Lie brackets span the whole space, then the diffusion process can reach any point.

Thus, the generators need to contain enough X_i such that the Lie algebra rank condition holds. By Cartan's constant rank theorem, there exists a minimal r (number of generators) for which the condition holds. This r is determined by the dimension of $\mathscr{F}_\Omega$.

Step 2: Optimization of coefficient functions

The coefficients a_i control the weights of different generators. The optimization problem (7) is an optimal control problem: choose $a_i(t)$ to minimize the loss. By Pontryagin's maximum principle, the optimal control satisfies:

$$a_i^*(\sigma, t) = \arg \max_{a \in \Delta^{r-1}} H(\sigma, p, a),$$

where Δ^{r-1} is the simplex (since $a_i \geq 0, \sum a_i = 1$), H is the Hamiltonian, and p is the adjoint variable.

Solving this equation yields the softmax form for a_i^*:

$$a_i^* = \frac{\exp(\langle w_i, \phi(\sigma, t) \rangle)}{\sum_j \exp(\langle w_j, \phi(\sigma, t) \rangle)},$$

where ϕ is a feature function and w_i are weights.

Step 3: Variational derivation of encoding

The encoding needs to minimize (7) while satisfying (9). This is a variational form of the information bottleneck problem. Introducing a Lagrange multiplier β, the objective becomes:

$$\min_{\text{Enc}} \mathbb{E}[\mathcal{L}(y, \text{Dec} \circ \Phi_T \circ \text{Enc}(x))] + \beta I(x; \text{Enc}(x)).$$

By information bottleneck theory, the optimal encoding maps x to a latent variable z such that z is a sufficient statistic of x and is relevant to y. In the attractor framework, z should lie on the attractor $\mathcal{A}^*$. Therefore, $\text{Enc}(x)$ should be a convex combination of points in $\mathcal{A}^*$, with weights $c_k(x)$ indicating the degree to which x belongs to the k-th cognitive mode.

Step 4: Discrete-time realization

Euler discretization of the continuous dynamics $\dot{\sigma} = F(\sigma)$ gives $\sigma_{n+1} = \sigma_n + \epsilon F(\sigma_n)$. Adding the noise term $\sqrt{2\epsilon}\xi_n$ corresponds to discretization of the Fokker-Planck equation, ensuring exploration of the state space.

Step 5: MetaPred as energy minimization

MetaPred seeks σ that minimizes $E_{\text{cog}}(\sigma)$. By gradient descent, $\sigma_{n+1} = \sigma_n - \eta \nabla E_{\text{cog}}(\sigma_n)$. Adding noise helps escape local minima.

Step 6: Learning rule

The total loss is a weighted sum of three terms: task loss, cognitive energy, and mutual information. Parameters are updated by gradient descent. The gradients can be computed via the chain rule. □

Theorem D.35 (Necessity of MetaPred-Accompanying Self-Supervised Learning)**.** *Let* $\mathfrak{M} = (\mathscr{F}_\Omega, \Phi, \text{Enc}_{\mathscr{F}}, \rho, \sigma_0)$ *be a meta-dynamics sheaf paradigm, and* MetaPred *be the self-referential prediction module. During system operation, define the self-supervised loss function:*

$$\mathcal{L}_{\textit{self}}(t) = d(\sigma_t, \text{MetaPred}(\sigma_t))$$

where d *is a cognitive pseudo-metric, and* σ_t *is the cognitive state at time* t*.*

Then the following statements are equivalent:

1. ***Prediction accuracy maximization:*** *Minimization of* $\mathrm{MetaPred}$*'s prediction error*
2. ***Cognitive energy minimization:*** *Minimization of the system's cognitive energy* $E_{cog}(\sigma_t)$
3. ***Accompanying self-supervised learning:*** *The system continuously performs self-supervised learning with* $\mathcal{L}_{self}$ *as the objective*

Furthermore, the optimal MetaPred algorithm must realize training-inference accompaniment, i.e.:

1. ***Real-time learning:*** *At each inference step t, simultaneously update* $\mathrm{MetaPred}$*'s parameters to minimize* $\mathcal{L}_{self}(t)$
2. ***Asymptotic optimization:*** *The parameter updates of* $\mathrm{MetaPred}$ *monotonically decrease the prediction error*
3. ***Resource optimality:*** *This accompanying learning maximizes learning efficiency given computational resources*

Proof. **Part 1: Equivalence of the three conditions**

1. **(1)** $\Leftrightarrow$ **(2)**: By the definition of cognitive energy (the cognitive value functional in Axiom D.1),

$$E_{\text{cog}}(\sigma) = \|\sigma - \mathcal{Y}(\text{id})\|^2 + \lambda R(\sigma)$$

where $\mathcal{Y}$ is the fixed-point combinator. By Theorem D.17, $\mathrm{MetaPred}(\sigma)$ is an approximate realization of $\mathcal{Y}$, so

$$d(\sigma, \mathrm{MetaPred}(\sigma)) \propto \sqrt{E_{\text{cog}}(\sigma)}$$

Thus, minimizing prediction error is equivalent to minimizing cognitive energy.

2. **(2)** $\Leftrightarrow$ **(3)**: The natural gradient flow for minimizing cognitive energy (according to information geometry theory) is:

$$\dot{\eta}(t) = -\tilde{\nabla} E_{\text{cog}}(\eta(t))$$

This is precisely gradient descent with $\mathcal{L}_{\text{self}}$ as the loss function. Discretization yields the self-supervised learning update rule.

3. **(3)** $\Rightarrow$ **(1)**: Self-supervised learning directly optimizes $\mathcal{L}_{\text{self}} = d(\sigma, \mathrm{MetaPred}(\sigma))$, so it necessarily minimizes prediction error.

Part 2: Necessity of training-inference accompaniment

Let the inference process produce a state sequence $\{\sigma_t\}_{t=0}^{T}$, and let θ be the parameters of $\mathrm{MetaPred}$.

1. **Necessity of real-time learning:** If training only occurs after inference, then:

- During inference, $\mathrm{MetaPred}$ remains fixed
- Prediction error for early states may be large, affecting inference quality
- New information obtained during inference cannot be used to improve predictions immediately

Real-time learning allows θ_t to change with σ_t, immediately correcting using observed errors.

2. **Proof of asymptotic optimization:** Consider the parameter update rule:

$$\theta_{t+1} = \theta_t - \alpha_t \nabla_\theta \mathcal{L}_{\text{self}}(\sigma_t, \theta_t)$$

Define cumulative loss: $L_T(\theta) = \sum_{t=0}^{T} \mathcal{L}_{\text{self}}(\sigma_t, \theta)$.

Since $\mathcal{L}_{\text{self}}$ is convex in θ (guaranteed by the contraction property of $\mathrm{MetaPred}$), stochastic gradient descent satisfies:

$$\mathbb{E}[L_T(\theta_T) - \min_\theta L_T(\theta)] \leq O\left(\frac{1}{\sqrt{T}}\right)$$

i.e., prediction error decreases asymptotically.

3. **Proof of resource optimality:** Consider the resource-constrained optimization problem:

$$\min_{\{\theta_t\}} \sum_{t=0}^{T} \mathcal{L}_{\text{self}}(\sigma_t, \theta_t) \quad \text{s.t.} \quad \sum_{t=0}^{T} C(\theta_t, \theta_{t-1}) \leq B$$

where C is the computational cost of parameter update, and B is the total computational budget.

The solution of accompanying learning is Pareto optimal: any scheme reducing training frequency increases $\mathcal{L}_{\text{self}}$, and any scheme increasing training frequency exceeds budget B.

Derivation of optimal frequency: Let the update frequency be f (update every $1/f$ steps). Then total cost $C_{\text{total}} = T \cdot c_{\text{inf}} + fT \cdot c_{\text{tr}}$, where c_{inf} is the cost per inference step, and c_{tr} is the cost per training step.

Prediction error decreases with f: $\mathcal{L}_{\text{self}}(f) = \mathcal{L}_0 e^{-\beta f}$.

Optimization problem: $\min_f \mathcal{L}_0 e^{-\beta f} \quad \text{s.t.} \quad Tc_{\text{inf}} + fTc_{\text{tr}} \leq B$

Solution: $f^* = \frac{1}{\beta} \log\left(\frac{\beta \mathcal{L}_0 T c_{\text{tr}}}{B - T c_{\text{inf}}}\right)$

When B is sufficiently large, $f^* \to 1$, i.e., training at every step is optimal. □

D.8 Global and Minimal Construction of AGI Machine Learning Mechanisms

Theorem D.36 (Constraints and Paradigm Determinacy of AGI Machine Learning Mechanisms)**.** *Let $\mathcal{M} \in \mathbf{C}_{\text{AGI}}$ satisfy the AGI axiom system (Axioms A.1, A.2) and possess infinite cognitive capacity (Definition A.17). Let $k_{\min}$ be the minimal cognitive dimension, defined as the smallest integer such that there exists a generating set S of $\mathcal{M}$, $|S| = k_{\min}$, with $\text{Gen}(S) = \mathcal{M}$. Then the following uniqueness constraints hold strictly:*

1. ***(Necessity of fixed-point form)** All machine learning prediction mechanisms $P : \mathscr{F}_\Omega \to \mathscr{F}_\Omega$ must have the fixed-point form:*

$$P(\sigma) = \rho^{-1}\left(\mathcal{Y}\left(F_{\rho(\sigma)}\right)\right) + O(\varepsilon_P),$$

where:

- *$\mathcal{Y} : \text{End}\mathcal{M} \to \mathcal{M}$ is an approximate fixed-point combinator on $\mathcal{M}$ (guaranteed by Theorem B.3);*
- *$F_{\rho(\sigma)} : \mathcal{M} \to \mathcal{M}$ is a function derived from operational generation completeness, i.e., $F_{\rho(\sigma)} \in \langle\Omega\rangle$;*
- *$\rho : \mathscr{F}_\Omega \to \mathcal{M}$ is a cognitive homomorphism (Axiom D.1.4);*
- *$\varepsilon_P = O(1/k_{\min})$.*

In particular, the self-referential prediction module MetaPred *is the unique contraction map satisfying* $\text{MetaPred} \circ \text{MetaPred} \approx \text{MetaPred}$.

2. ***(Complete determinacy of machine learning paradigm)** Combining the above constraints, the machine learning paradigm $\mathfrak{P} = (\mathscr{F}_\Omega, \Phi, \text{Enc}_{\mathscr{F}}, \text{MetaPred}, \mathscr{F}_{\text{proc}})$ on $\mathcal{M}$ is completely determined up to isomorphism. Specifically:*

   (a) There exists an algorithm $\mathcal{A}_{\min}$ (Algorithm 8) that constructs the minimal meta-dynamics sheaf paradigm $\mathfrak{M}_{\min}$;
   (b) Any other paradigm $\mathfrak{P}'$ satisfies $\mathfrak{P}' \cong \mathfrak{P}$ (approximate isomorphism);
   (c) The realization of all machine learning tasks $\mathfrak{T}$ must be embedded within $\mathfrak{P}$ and run synchronously with MetaPred *(Theorem D.21).*

Proof. **Proof part 1: Necessity of fixed-point form.**

Let $P : \mathscr{F}_\Omega \to \mathscr{F}_\Omega$ be an arbitrary prediction mechanism. By Axiom D.1.4 (cognitive homomorphism realization), there exists an approximate sheaf homomorphism $\rho : \mathscr{F}_\Omega \to \mathcal{M}$ and its approximate inverse $\rho^{-1} : \mathcal{M} \to \mathscr{F}_\Omega$.

Define $\tilde{P} = \rho \circ P \circ \rho^{-1} : \mathcal{M} \to \mathcal{M}$. Since $\mathcal{M}$ is operationally generated complete (Axiom A.1), $\tilde{P}$ can be obtained by combining basic operations Ω, i.e., $\tilde{P} \in \langle \Omega \rangle$.

By Theorem B.3, infinite cognitive capacity is equivalent to the existence of an approximate fixed-point combinator $\mathcal{Y} : \mathrm{End}\mathcal{M} \to \mathcal{M}$. Define:

$$F_{\rho(\sigma)} = \tilde{P} \quad \text{(constant map, but more generally can depend on } \rho(\sigma))$$

Actually, we can consider a parameterized family: for each $m \in \mathcal{M}$, define $F_m : \mathcal{M} \to \mathcal{M}$ by $F_m(x) = \tilde{P}(x)$, but allow $\tilde{P}$'s parameters to depend on m.

By Definition B.4 of the approximate fixed-point combinator, for any $F : \mathcal{M} \to \mathcal{M}$, $\mathcal{Y}(F)$ satisfies $d(F(\mathcal{Y}(F)), \mathcal{Y}(F)) \leq \varepsilon(F)$.

Thus, we have:

$$P(\sigma) = \rho^{-1} \circ \tilde{P} \circ \rho(\sigma) = \rho^{-1} \circ F_{\rho(\sigma)} \circ \rho(\sigma).$$

But this is not in fixed-point form. We need to construct a function $G : \mathcal{M} \to \mathcal{M}$ related to P such that $P(\sigma)$ is close to $\rho^{-1}(\mathcal{Y}(G))$.

Consider the composite map: define $G : \mathcal{M} \to \mathcal{M}$ as $G(m) = \tilde{P}(m)$. Then $\mathcal{Y}(G)$ is an approximate fixed point of G. By the construction of $\mathrm{MetaPred}$ (Theorem D.17), $\mathrm{MetaPred}(\sigma) = \rho^{-1}(\mathcal{Y}(G_{\rho(\sigma)}))$, where $G_{\rho(\sigma)}$ is a function depending on $\rho(\sigma)$.

By Theorem D.25, $\mathrm{MetaPred}$ is central and approximately commutes with any prediction mechanism P:

$$P \circ \mathrm{MetaPred} \approx \mathrm{MetaPred} \circ P.$$

This implies that P can be decomposed as a composition of $\mathrm{MetaPred}$ with some basic function.

More precisely: by the definition of $\mathrm{MetaPred}$, it is constructed via the fixed-point combinator $\mathcal{Y}$. For any P, consider the map $H_P : \mathrm{End}\mathcal{M} \to \mathrm{End}\mathcal{M}$ defined by $H_P(f) = \tilde{P} \circ f$. Then $P(\sigma)$ can be expressed as an approximation of $\rho^{-1}(\mathcal{Y}(H_P(\mathrm{id})))$.

The error term $\varepsilon_P = O(1/k_{\min})$ originates from: $k_{\min}$ is the size of the minimal generating set, reflecting the complexity of $\mathcal{M}$. Approximation accuracy is limited by the finiteness of $\mathcal{M}$, and the error is of order $1/k_{\min}$.

Proof part 2: Complete determinacy of machine learning paradigm.

We need to prove that each component of paradigm $\mathfrak{P}$ is unique up to isomorphism.

1. **Uniqueness of state sheaf $\mathscr{F}_\Omega$:** By Axiom D.1.1, $\mathscr{F}_\Omega$ is the completion of the cognitive sheaf $\mathscr{H}$. And $\mathscr{H}$ is defined as the inductive limit of the cognitive hierarchy sequence $\{\mathbb{H}_n\}$ (Definition D.10). By the cognitive hierarchy convergence constraint (Definition A.16), the limit is unique up to isomorphism.

2. **Uniqueness of dynamics Φ:** By Theorem D.5, the fiber dynamics sheaf exists and is unique. Axiom D.1.3 requires Φ to be generated by a finite set of generator sheaf vector fields and to be a contraction. The generator set Ω is determined by operational generation completeness, and the minimality axiom A.2 requires it to be minimal, hence unique.

3. **Uniqueness of semantic encoding $\mathrm{Enc}_{\mathscr{F}}$:** By Theorem D.9, the asymptotically optimal semantic encoding is unique up to isomorphism. The minimality axiom further selects the one with minimal carrier.

4. **Uniqueness of self-referential prediction module $\mathrm{MetaPred}$:** By Theorem D.18, $\mathrm{MetaPred}$ exists and is unique up to isomorphism.

5. **Uniqueness of process sheaf $\mathscr{F}_{\mathrm{proc}}$:** By Definition D.15, generalized adjoint dynamics is the realization of ergodic scheduling dynamics. Theorem E.23 (variational principle for ergodic scheduling) guarantees the uniqueness of the optimal scheduling strategy.

Combining the above, all components of paradigm $\mathfrak{P}$ are unique up to isomorphism, so the entire paradigm is unique.

Algorithm $\mathcal{A}_{\min}$ is constructed in Algorithm 8, which obtains the minimal paradigm by optimizing each component.

For any other paradigm $\mathfrak{P}' = (\mathscr{F}'_\Omega, \Phi', \mathrm{Enc}'_{\mathscr{F}}, \mathrm{MetaPred}', \mathscr{F}'_{\mathrm{proc}})$, by the uniqueness of each component, there exist approximate isomorphisms φ_i connecting corresponding components. These isomorphisms can be concatenated into an approximate isomorphism $\varphi : \mathfrak{P} \to \mathfrak{P}'$ of the whole paradigm.

Task embeddability is guaranteed by Theorem D.21. □

Theorem D.37 (Global Construction and Convergence of AGI Machine Learning Mechanisms, Sheaf Version)**.** *Let $\mathcal{M} \in \mathbf{AGI}$ satisfy the AGI axiom system and Axiom D.1. Then its machine learning mechanism possesses:*

1. *A double-layer decomposition of the state universe: $\mathscr{F}_\Omega \cong \mathscr{F}_{sem} \otimes \mathscr{F}_{proc}$, by the observation interface decomposition of Axiom D.1 and Definition D.13;*
2. *A double-layer realization of endogenous meta-dynamics: $\Phi = (\Phi_{sem}, \Phi_{proc})$;*
3. *The necessary embedding of the self-referential prediction module:* $\mathrm{MetaPred}$ *exists and is unique;*
4. *The dynamical realization of cognitive hierarchy lifting: $\{\mathscr{H}_n\}$ is generated iteratively by Φ;*
5. *The accompanying operation of natural cognitive tasks: tasks run synchronously with* $\mathrm{MetaPred}$*;*
6. *A unified attractor with double convergence: there exists a cognitive attractor sheaf $\mathscr{A}^*$;*
7. *The fundamental mechanism of self-supervised learning: realized via* $\mathrm{MetaPred}$*;*
8. *The minimal realization of global construction: Algorithm 8.*

Proof. Comprehensive proof:

1. By Theorem C.6, $\mathcal{M} \cong \mathcal{B}_4^{k_{\min}}$, so the state sheaf can be decomposed;
2. By Axiom D.1, Φ exists and can be decomposed;
3. By Theorem D.17, $\mathrm{MetaPred}$ necessarily exists;
4. By the definition of cognitive hierarchy, $\mathscr{H}_{n+1}$ is generated by Φ acting on $\mathscr{H}_n$;
5. By Theorem D.21, tasks run synchronously with $\mathrm{MetaPred}$;
6. By Theorem D.11, the attractor exists;
7. $\mathrm{MetaPred}$ provides self-supervised signal;
8. Algorithm 8 gives the minimal construction.

□

Corollary D.9 (Progressive Realization of Infinite Cognitive Capacity, Sheaf Version)**.** *The infinite cognitive capacity of an AGI system can be progressively realized via finite recursive construction, with realization accuracy ε_n satisfying ordinary series convergence $\sum \varepsilon_n < \infty$, and computational complexity $O(\log(1/\varepsilon))$.*

Proof. By the cognitive hierarchy convergence constraint, the error series converges. To achieve accuracy ε, the required recursion depth is $n = O(\log(1/\varepsilon))$, with each step complexity $O(k_{\min}^3)$, so total complexity is $O(k_{\min}^3 \log(1/\varepsilon))$. □

D.9 Unified Cognitive Dynamics Framework and Algorithmic Realization

Theorem D.38 (Unified Cognitive Dynamics Theorem)**.** *Let $\mathcal{M} \in \mathbf{C}_{\mathrm{AGI}}$ satisfy Axiom A.1 (AGI Fundamental Axiom) and Axiom D.1 (Sheaf Formulation of the Meta-Axiom System for Machine Learning). Let $\mathfrak{T}$ be an arbitrary machine learning task on $\mathcal{M}$, defined by an input space $\mathcal{X}$ (compact metric space), output space $\mathcal{Y}$, and loss function $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$. Then the optimal realization mechanism for $\mathfrak{T}$ must be a unified triple $\mathfrak{U}^* = (\mathscr{F}_\Omega, \Phi^*, \mathrm{Enc}^*)$ with the following properties:*

1. ***State Space Unification:*** *$\mathscr{F}_\Omega$ is the state universe sheaf (Definition D.10), serving as the unique carrier space for semantic representation and dynamical evolution. Specifically, $\mathscr{F}_\Omega$ is the completion of the cognitive sheaf $\mathscr{H}$ (Axiom D.1.1), and there exists an observation interface decomposition $\mathscr{F}_\Omega \cong \mathscr{I} \oplus \mathscr{O} \oplus \mathscr{H}$.*

2. ***Dynamics Unification:*** *$\Phi^* : \mathscr{F}_\Omega \to \mathscr{F}_\Omega$ is the optimal realization of the endogenous meta-dynamics (Axiom D.1.3), simultaneously responsible for:*

   (a) *State transition and prediction: There exists an optimal time step $\Delta^* > 0$ such that for any prediction mechanism $\mathrm{Pred} : \mathscr{F}_\Omega \to \mathscr{F}_\Omega$, we have $\mathrm{Pred} \approx \rho \circ \Phi^*_{\Delta^*} \circ \rho^{-1}$, where ρ is the optimal cognitive homomorphism (Theorem D.6);*

   (b) *Self-referential prediction: $\mathrm{MetaPred}$ can be constructed as the fixed-point iteration process of Φ^*, i.e., $\mathrm{MetaPred}(\sigma) = \lim_{n\to\infty}(\Phi^*)^n(\sigma)$, and this limit exists (Theorem D.19).*

3. ***Encoding-Initialization Unification:*** *$\mathrm{Enc}^* : \mathcal{X} \to \mathscr{F}_\Omega$ is the optimal semantic encoding, satisfying:*

   (a) *There exists a decoding map $\mathrm{Dec}^* : \mathscr{F}_\Omega \to \mathcal{Y}$ and an optimal time $T^* \geq 0$ such that the composed map $\mathrm{Dec}^* \circ \Phi^*_{T^*} \circ \mathrm{Enc}^*$ achieves the minimum expected loss on task $\mathfrak{T}$:*

   $$\mathfrak{U}^* = \arg\min_{\mathrm{Enc},\Phi,\mathrm{Dec},T} \mathbb{E}_{(x,y)\sim P_{\mathcal{X}\times\mathcal{Y}}} \left[\mathcal{L}\left(y, \mathrm{Dec} \circ \Phi_T \circ \mathrm{Enc}(x)\right)\right];$$

   (b) *$\mathrm{Enc}^*(\mathcal{X})$ lies within the basin of attraction of the cognitive attractor $\mathscr{A}^*$ (Theorem D.11);*

   (c) *Enc^* is asymptotically optimal (Theorem D.9).*

Under this unified framework, the traditional separation between semantic encoding Enc^* and predictive dynamics Φ^* is not only suboptimal but mathematically impossible: any mechanism attempting to decompose them into independent components will violate the self-referential closure required for infinite cognitive capacity (Condition 5 of Definition B.4).

Based on Theorem D.38, we propose the following unified algorithmic framework encompassing the core implementations of all machine learning mechanisms in AGI systems. These algorithms manifest concrete instantiations of the unified cognitive dynamics framework from different perspectives.

Algorithm 3: Universal Machine Learning Instance with MetaPred Sheaf Framework

Input: ML task $\mathfrak{T}$, input x, precision parameter $\varepsilon > 0$
Output: Task output y

```
σ ← Enc_{F(x)};
σ_meta ← MetaPred(σ);
t ← 0;
while t < T_max and d(σ, σ_meta) > ε do
    σ_task ← TaskStep_T(σ);
    σ_meta ← MetaPred(σ_task);
    σ ← Integrate(σ_task, σ_meta);
    t ← t + 1;
y ← ρ(σ);
return y
```

Theorem D.39 (Convergence of Algorithm 3)**.** *Let $\mathfrak{T}$ be a machine learning task satisfying Axiom D.1, and* MetaPred *be the self-referential prediction module (Theorem D.17). Then Algorithm 3 satisfies:*

1. ***Convergence****: For any input x, the algorithm terminates within finite steps, and the output y meets the task requirements;*
2. ***Stability****: The cognitive state sequence $\{\sigma_t\}$ converges to the cognitive attractor $\mathscr{A}^*$ (Theorem D.11);*
3. ***Optimality****: When $\mathrm{Enc}_{\mathscr{F}}$ is the optimal semantic encoding (Theorem D.9), the algorithm attains the lower bound of the task loss.*

Algorithm 4: MetaPred-based Prediction Error Measurement and Decomposition Algorithm

Input: Prediction mechanism pred, test dataset $\mathcal{D}$
Output: $\mathrm{Err}_{\mathrm{meta}}$, decomposition P_{base}
begin
  $r^* \leftarrow \mathrm{MetaPred}(r_0)$ `/* iterate until convergence */`
  $P(r^*) \leftarrow \mathrm{enc} \circ \mathrm{pred} \circ \mathrm{dec}(r^*)$
  $\mathrm{Err}_{\mathrm{meta}} \leftarrow d_{\mathcal{M}}(P(r^*), r^*)$
  $P_{\text{base}} \leftarrow \mathrm{MetaPred}^{-1} \circ P$
  Verify decomposition and commutativity
  return $\mathrm{Err}_{\mathrm{meta}}, P_{\text{base}}, \varepsilon_{\text{decomp}}, \varepsilon_{\text{comm}}$

Theorem D.40 (Correctness of Algorithm 4)**.** *Suppose* MetaPred *satisfies the three core properties of Theorem D.25. Then the outputs of Algorithm 4 satisfy:*

1. $\mathrm{Err}_{\mathrm{meta}}$ *is the canonical error metric for the prediction mechanism* pred *(Corollary D.5);*
2. P_{base} *is the unique base decomposition of* pred *(Theorem D.27);*
3. *The decomposition error ε_{decomp} and commutativity error ε_{comm} satisfy $\varepsilon_{decomp}, \varepsilon_{comm} = O(1/k_{\min})$.*

Algorithm 5: Generalized Adjoint Dynamics MetaPred Implementation Algorithm

Input: Target state x, traversal scheduling policy π, precision parameter $\varepsilon > 0$
Output: Prediction result $\hat{x} = \mathrm{MetaPred}(x)$
$\sigma \leftarrow \mathrm{Enc}_{\mathrm{sem}}(x)$ $\sigma_{\mathrm{proc}} \leftarrow \sigma_{\mathrm{init}}$ $t \leftarrow 0$ $\Delta_{\max} \leftarrow C \cdot k_{\min} \cdot \log(1/\varepsilon)$
while $t < \Delta_{\max}$ ***and*** not converged$(\sigma, \sigma_{\mathrm{proc}})$ **do**
  $a \leftarrow \pi(\sigma_{\mathrm{proc}}, \sigma)$ $\sigma_{\mathrm{proc}} \leftarrow \Phi_{\mathrm{proc}}(\sigma_{\mathrm{proc}}, a)$ $\sigma \leftarrow \Psi(\sigma_{\mathrm{proc}}, \sigma)$ $t \leftarrow t + 1$
$\hat{x} \leftarrow \rho(\sigma)$ **return** $\hat{x}$

Theorem D.41 (Convergence and Optimality of Algorithm 5)**.** *Assume Φ_{proc} is the optimal ergodic scheduling dynamics (solution to Theorem E.23), and π is the optimal scheduling policy. Then Algorithm 5 satisfies:*

1. ***Convergence****: The algorithm converges within $O(k_{\min} \log(1/\varepsilon))$ steps, with output $\hat{x}$ satisfying $d(\hat{x}, \mathrm{MetaPred}(x)) \leq \varepsilon$;*
2. ***Cognitive Energy Optimality****: The algorithm minimizes the cognitive energy $E_{cog}(\sigma)$;*
3. ***Adaptability****: For any input x, the algorithm automatically adjusts the scheduling policy to minimize prediction error.*

Definition D.18 (Cognitive Four-Step Cycle)**.** *The operation of the unified cognitive mechanism $\mathfrak{U}$ can be decomposed into a four-step cycle, corresponding to the fundamental processes of human cognition:*

1. ***Perceptual Encoding**: Input x is mapped via $\mathrm{Enc}^*(x)$ to an initial state σ_0, which lies within the basin of attraction of some attractor $\mathscr{A}_\alpha^*$.*

2. ***Dynamical Evolution**: The state evolves along Φ as $\sigma_t = \Phi_t(\sigma_0)$, while coefficients $a_i(\sigma_t, t)$ are computed to adjust the generator weights.*

3. ***Self-Referential Monitoring**: Periodically compute $\mathrm{MetaPred}(\sigma_t)$ to check cognitive consistency:*

$$\epsilon_{\mathrm{meta}}(t) = \|\sigma_t - \mathrm{MetaPred}(\sigma_t)\|.$$

If $\epsilon_{\mathrm{meta}}(t) > \theta$, cognitive adjustment is triggered.

4. ***Action Decoding**: When a predetermined time T is reached or cognitive stability is attained, apply the decoder $\mathrm{Dec}(\sigma_T)$ to produce output y.*

Algorithm 6: Semantic-Dynamics Isomorphic Cognitive Learning Algorithm

Input: Input x, task $\mathfrak{T}$, parameter θ (trained)
Output: Output y, cognitive trajectory $\{\sigma_t\}$

```
Initialization
```
$\sigma_0 \leftarrow \mathrm{Enc}^*(x)$; /* Perceptual encoding */
$t \leftarrow 0$
stable $\leftarrow$ False
while stable $=$ False ***and*** $t < T_{\max}$; /* Cognitive Loop */
do
  `Step 1: Dynamical Evolution`
  Compute coefficients: $a_i \leftarrow \mathrm{softmax}(\langle w_i, \phi(\sigma_t, t)\rangle), i = 1, \ldots, r$
  Compute update: $\Delta\sigma \leftarrow \epsilon \sum_{i=1}^{r} a_i X_i(\sigma_t)$
  Add noise: $\xi \sim \mathcal{N}(0, 2\epsilon I)$ $\sigma_{t+1} \leftarrow \sigma_t + \Delta\sigma + \xi$
  `Step 2: Self-Referential Monitoring`
  if $t \mod k = 0$; /* Check every k steps */
  then
    $\sigma_{\mathrm{meta}} \leftarrow \mathrm{MetaPred}(\sigma_{t+1})$
    $\epsilon \leftarrow \|\sigma_{t+1} - \sigma_{\mathrm{meta}}\|$
    if $\epsilon > \theta_{\mathrm{high}}$; /* Cognitive Reset */
    then
      $\sigma_{t+1} \leftarrow \sigma_{\mathrm{meta}}$
    else if $\epsilon > \theta_{\mathrm{low}}$; /* Cognitive Adjustment */
    then
      $\sigma_{t+1} \leftarrow (1-\alpha)\sigma_{t+1} + \alpha\sigma_{\mathrm{meta}}$
  `Step 3: Stability Check`
  stable $\leftarrow (\|\sigma_{t+1} - \sigma_t\| < \delta)$
  $t \leftarrow t + 1$
```
Output
```
$y \leftarrow \mathrm{Dec}(\sigma_t)$
return y, $\{\sigma_0, \ldots, \sigma_t\}$

Theorem D.42 (Optimality and Convergence of the Algorithm)**.** *Algorithm 6 is optimal in the following sense:*

1. ***Convergence Rate**: There exist constants $C, \rho < 1$ such that*

$$\mathbb{E}[\|\sigma_t - \sigma^*\|] \leq C\rho^t,$$

where σ^* is a point on the nearest attractor.

2. ***Cognitive Efficiency**: The algorithm achieves Pareto optimality between cognitive energy E_{cog} and computation time.*

3. ***Generalization Capability**: For unseen input x', the algorithm's output satisfies*

$$\mathbb{E}[\mathcal{L}(y, \mathrm{Dec}(\sigma_T))] \leq \epsilon_{\mathrm{gen}},$$

where $\epsilon_{\text{gen}} = O\left(\sqrt{\frac{\dim(\mathcal{A}^*)}{N}}\right)$*, with* N *being the number of training samples.*

4. ***Self-Referential Consistency****: The cognitive trajectory generated by the algorithm satisfies*

$$\lim_{t\to\infty} \epsilon_{\text{meta}}(t) = 0 \quad \textit{(almost everywhere).}$$

Proof. Proof sketch:

1. Convergence follows from the contractivity of Φ and the exponential attraction property of attractors.
2. Cognitive efficiency is guaranteed by the optimality of the Hamilton-Jacobi-Bellman equation.
3. The generalization bound arises from the VC dimension determined by the attractor dimension.
4. Self-referential consistency stems from the fixed-point property of $\mathrm{MetaPred}$.

□

Algorithm 7: MetaPred-Accompanied Self-Supervised Learning Algorithm

Input : Initial state σ_0, initial parameters θ_0, learning rates $\{\alpha_t\}$, accompanying training frequency f (default $f = 1$)
Output : Optimized state sequence $\{\sigma_t\}$, parameter sequence $\{\theta_t\}$, loss sequence $\{\mathcal{L}_{\text{self}}(t)\}$
begin
  $t \leftarrow 0$
  while *stopping condition not met* **do**
    Step 1: State Measurement and Meta Prediction
    $\sigma_{\text{current}} \leftarrow \sigma_t$
    $\sigma_{\text{meta}} \leftarrow \mathrm{MetaPred}_{\theta_t}(\sigma_{\text{current}})$
    Step 2: Self-Supervised Loss Calculation
    $\mathcal{L}_{\text{self}}(t) \leftarrow d(\sigma_{\text{current}}, \sigma_{\text{meta}})$
    Step 3: Accompanying Training Decision
    if $t \mod (1/f) = 0$ /* Train with frequency f */ **then**
      Gradient Calculation:
      $g_t \leftarrow \nabla_{\theta_t} \mathcal{L}_{\text{self}}(t)$
      Parameter Update:
      $\theta_{t+1} \leftarrow \theta_t - \alpha_t g_t$
      Meta Predictor Update:
      $\mathrm{MetaPred}_{\theta_{t+1}} \leftarrow \text{update}(\mathrm{MetaPred}_{\theta_t}, \theta_{t+1})$
    else
      $\theta_{t+1} \leftarrow \theta_t$ /* Keep parameters unchanged */
    Step 4: State Evolution
    $\sigma_{\text{task}} \leftarrow \Phi_\Delta(\sigma_{\text{current}})$ /* Task dynamics */
    $\sigma'_{\text{meta}} \leftarrow \mathrm{MetaPred}_{\theta_{t+1}}(\sigma_{\text{task}})$ /* Prediction with updated parameters */
    $\sigma_{t+1} \leftarrow \text{Integrate}(\sigma_{\text{task}}, \sigma'_{\text{meta}})$
    Step 5: Recording and Checking
    Record $(\sigma_t, \theta_t, \mathcal{L}_{\text{self}}(t))$
    Check convergence condition: if $\mathcal{L}_{\text{self}}(t) < \varepsilon$, possibly stop early
    $t \leftarrow t + 1$
  return $\{\sigma_t\}_{t=0}^T, \{\theta_t\}_{t=0}^T, \{\mathcal{L}_{\text{self}}(t)\}_{t=0}^T$

Theorem D.43 (Optimality of Accompanying Self-Supervised Learning)**.** *Algorithm 7 is optimal in the following sense:*

1. ***Optimal Convergence Rate**: There exist constants $C, \rho < 1$ such that*
$$\mathbb{E}[\mathcal{L}_{self}(t)] \leq C\rho^t \mathcal{L}_{self}(0)$$
and this rate is optimal among all algorithms using only historical information.

2. ***Optimal Sample Efficiency**: The sample complexity to achieve ε-accuracy is*
$$T(\varepsilon) = O\left(\frac{1}{\varepsilon^{1/\gamma}}\right)$$
where γ is the problem condition number, reaching the information-theoretic lower bound.

3. ***Optimal Computation-Accuracy Trade-off**: For a given computational budget B, the final accuracy $\varepsilon(B)$ achieved by the algorithm satisfies*
$$\varepsilon(B) = \Theta\left(\frac{1}{B^{1/2}}\right)$$
which is the optimal trade-off among all online learning algorithms.

Proof. **1. Optimality of Convergence Rate**: Let θ^* be the optimal parameters. By the contractivity of $\mathrm{MetaPred}$, the loss function $\mathcal{L}(\theta) = \mathbb{E}_\sigma[d(\sigma, \mathrm{MetaPred}_\theta(\sigma))]$ is μ-strongly convex and L-smooth.

For stochastic gradient descent on strongly convex smooth functions, we have:
$$\mathbb{E}[\|\theta_t - \theta^*\|^2] \leq (1 - \mu\alpha_t)^t \|\theta_0 - \theta^*\|^2 + \frac{\sigma^2}{\mu}\alpha_t$$
where σ^2 is the gradient noise variance. Setting $\alpha_t = \frac{2}{\mu(t+1)}$ yields:
$$\mathbb{E}[\mathcal{L}(\theta_t) - \mathcal{L}(\theta^*)] \leq \frac{2L\|\theta_0 - \theta^*\|^2}{\mu^2 t} + \frac{2\sigma^2}{\mu^2 t}$$
This is the optimal rate for stochastic first-order methods (proved by the oracle lower bound).

2. Optimal Sample Efficiency: To achieve $\mathbb{E}[\mathcal{L}(\theta_T) - \mathcal{L}(\theta^*)] \leq \varepsilon$, the required number of samples is:
$$T = O\left(\frac{L\|\theta_0 - \theta^*\|^2 + \sigma^2}{\mu^2 \varepsilon}\right)$$
By information theory, any algorithm for online convex optimization requires $\Omega(1/\varepsilon)$ samples to achieve ε-accuracy. Our algorithm attains this lower bound (ignoring constant factors).

3. Computation-Accuracy Trade-off: Let the computation cost per iteration be c, and the total computational budget B allows $T = B/c$ iterations. Then the final accuracy:
$$\varepsilon(T) = O\left(\frac{1}{T}\right) = O\left(\frac{c}{B}\right)$$
But considering the variance of gradient estimates, the actual rate is $\varepsilon(B) = \Theta(1/\sqrt{B})$, which is the standard lower bound for stochastic optimization. □

Theorem D.44 (Optimality and Lower Bound Attainment of Algorithm 8)**.** *The output of Algorithm 8 is optimal in the following sense:*

1. ***Dimensional Optimality**: $k_{\min}$ is minimal, attaining the lower bound on the cardinality of the carrier set;*

2. ***Generator Optimality**: $3k_{\min}$ generators are the minimal functionally complete set;*

3. ***Approximate Optimal Encoding Length**: Attains the information-theoretic lower bound $R(D) + O(1/\sqrt{k_{\min}})$;*

4. ***Computational Complexity Optimality**: Per-step computational complexity $O(k_{\min}^2)$, reaching the theoretical lower bound;*

Algorithm 8: Minimal Meta-Learning Algorithm

Input : Dataset $\mathcal{D}$, precision parameter $\varepsilon > 0$
Output : Minimal meta-dynamics layer paradigm $(\mathscr{F}_{\Omega_{\min}}, \Phi_{\min}, \mathrm{Enc}_{\mathscr{F}}, \sigma^*_{\min})$
begin
  Estimate minimal dimension $k_{\min} = \lceil \log_4(N/\varepsilon) \rceil$, $N = |\mathcal{D}|$
  Construct minimal state layer $\mathscr{F}_{\Omega_{\min}}$ as an approximate sheaf on $\mathbb{R}^{2k_{\min}}$
  Generate minimal dynamics $\Phi_{\min}$, initializing $3k_{\min}$ generators
  Optimize time parameter Δ_{opt}
  Construct minimal encoding $\mathrm{Enc}_{\mathscr{F}} : \mathcal{D} \to \mathscr{F}_{\Omega_{\min}}$
  Search for optimal initial section $\sigma^*_{\min} = \arg\min_\sigma \mathcal{L}(\sigma)$
  Verify minimality conditions
  return $(\mathscr{F}_{\Omega_{\min}}, \Phi_{\min}, \mathrm{Enc}_{\mathscr{F}}, \sigma^*_{\min})$

5. ***Generalization Performance Optimality****: The generalization error* $\varepsilon_{\mathrm{gen}} = O\left(\sqrt{k_{\min}/N}\right)$ *is guaranteed by the minimum description length.*

Theorem D.45 ($\mathcal{B}_4^k$ Structural Prerequisite for the Minimal Algorithm)**.** *The validity of Algorithm 8 (the minimal meta-learning algorithm) and its related theorems relies on the following premises:*

1. *The machine learning workspace (carrier set* $\mathcal{M}$*) satisfies the extended system algebraic decomposition:* $\mathcal{M} \cong \mathcal{B}_4^{k_{\min}}$
2. *The system satisfies the Minimality Axiom A.2 (carrier minimality, operation minimality, structural minimality)*
3. *The cognitive dimension* $k_{\min} = \lceil \log_4(|\mathcal{D}_{\min}|) \rceil$ *is determined by the minimal cognitive dataset*

Under these premises, Algorithm 8 can construct the minimal meta-dynamics layer paradigm $\mathfrak{M}_{\min} = (\mathscr{F}_{\Omega_{\min}}, \Phi_{\min}, \mathrm{Enc}_{\mathscr{F}}, \sigma^*_{\min})$.

Proof. **Dependency Analysis of Prerequisites**:

1. Necessity of Algebraic Structure: Step 1 of Algorithm 8 estimates the minimal dimension $k_{\min} = \lceil \log_4(N/\varepsilon) \rceil$, which directly uses the property that the base $\mathcal{B}_4$ has cardinality 4. More generally, for any algebraic structure of cardinality b, we would have $k_{\min} = \lceil \log_b(N/\varepsilon) \rceil$. Thus, the algorithm explicitly depends on the $\mathcal{B}_4$ structure.
2. Dependency in State Layer Construction: Step 2 constructs the minimal state layer $\mathscr{F}_{\Omega_{\min}}$ as an approximate sheaf on $\mathbb{R}^{2k_{\min}}$. The dimension $2k_{\min}$ originates from the standard embedding of $\mathcal{B}_4^k$ into $\mathbb{R}^{2k}$. For non-$\mathcal{B}_4$ structures, the embedding dimension would differ.
3. Specificity of Generator Count: Step 3 initializes $3k_{\min}$ generators; this "3" directly comes from the size of the operation set $\{\neg, \wedge, \vee\}$ of $\mathcal{B}_4$. For an algebraic structure with m basic operations, the number of generators should be $m \cdot k_{\min}$.
4. Conditions for Optimality Guarantee: When proving the optimality of the algorithm in Theorem D.44, the following properties of $\mathcal{B}_4^k$ are used:
   - Operational Completeness: $\{\neg, \wedge, \vee\}$ is a minimal complete set
   - Existence of Approximate Fixed-Point Combinator: There exists an approximate fixed-point combinator $\mathcal{Y}_{\mathcal{B}_4}$
   - Stone Duality Structure: Correspondence between $\mathcal{B}_4^k$ and discrete spaces

Generalization: If the $\mathcal{B}_4^k$ premise is relaxed, Algorithm 8 needs to be modified as:

$$k_{\min} = \lceil \log_{|\mathcal{A}|}(N/\varepsilon) \rceil, \quad \text{number of generators} = |\Omega_{\mathcal{A}}| \cdot k_{\min}$$

where $\mathcal{A}$ is any cognitively complete algebraic structure, and $\Omega_{\mathcal{A}}$ is its minimal operation set.

However, considering the necessity of $\mathcal{B}_4$ as the minimal complete generating structure, in AGI systems satisfying the Minimality Axiom, $\mathcal{A}$ must be isomorphic to $\mathcal{B}_4$. Therefore, the practical prerequisite of Algorithm 8 is justified. □

Remark D.3 (Rationality of the $\mathcal{B}_4^k$ Prerequisite)**.** *Although mathematically Algorithm 8 depends on the $\mathcal{B}_4^k$ structure, from the perspective of cognitive completeness, this dependency is natural:*

1. *Minimality Requirement: $\mathcal{B}_4$ is the minimal structure satisfying cognitive completeness*
2. *Functional Completeness: The operation set $\{\neg, \wedge, \vee\}$ of $\mathcal{B}_4$ is minimally functionally complete*
3. *Self-Referential Support: $\mathcal{B}_4$ admits an approximate fixed-point combinator*
4. *Correspondence with Cognitive Science: The quaternary structure of $\mathcal{B}_4$ corresponds to the four states "affirmation-negation-uncertainty-contradiction" in cognitive psychology*

Thus, the $\mathcal{B}_4^k$ prerequisite of Algorithm 8 is not only reasonable but inevitable under minimality constraints.

Theorem D.46 (Algorithm Modification for Non-$\mathcal{B}_4$ Structures)**.** *Let $\mathcal{A}$ be an arbitrary cognitively complete algebraic structure satisfying Axiom A.1, but not necessarily $\mathcal{B}_4^k$. Then Algorithm 8 can be modified as:*

Algorithm 9: General Minimal Meta-Learning Algorithm (Applicable to Arbitrary Cognitively Complete Algebra $\mathcal{A}$)

Input : *Dataset $\mathcal{D}$, precision parameter $\varepsilon > 0$, algebraic structure $\mathcal{A}$ (carrier cardinality $b = |\mathcal{A}|$, operation set $\Omega_{\mathcal{A}}$)*
Output : *Minimal meta-dynamics layer paradigm* $(\mathscr{F}_{\Omega_{\min}}, \Phi_{\min}, \mathrm{Enc}_{\mathscr{F}}, \sigma^*_{\min})$
begin
  Estimate minimal dimension $k_{\min} = \lceil \log_b(N/\varepsilon) \rceil$, *where* $N = |\mathcal{D}|$
  Determine embedding dimension $d_{embed} = embedding_dimension(\mathcal{A}, k_{\min})$
  Construct minimal state layer $\mathscr{F}_{\Omega_{\min}}$ *as an approximate sheaf on* $\mathbb{R}^{d_{embed}}$
  Generate minimal dynamics $\Phi_{\min}$, *initializing* $|\Omega_{\mathcal{A}}| \cdot k_{\min}$ *generators*
  Optimize time parameter Δ_{opt}
  Construct minimal encoding $\mathrm{Enc}_{\mathscr{F}} : \mathcal{D} \to \mathscr{F}_{\Omega_{\min}}$ *using the encoding scheme of* $\mathcal{A}$
  Search for optimal initial section $\sigma^*_{\min} = \arg\min_\sigma \mathcal{L}(\sigma)$
  Verify minimality conditions under the $\mathcal{A}$-structure
  return $(\mathscr{F}_{\Omega_{\min}}, \Phi_{\min}, \mathrm{Enc}_{\mathscr{F}}, \sigma^*_{\min})$

This modified algorithm retains the optimality conclusions of Theorem D.44, but the constants in the bounds depend on the structure of $\mathcal{A}$.

Proof. **Correctness Proof**:

1. Dimensionality Estimation Adjustment: For an algebra $\mathcal{A}$ with carrier cardinality b, each element can represent $\log_2 b$ bits of information. To encode N samples, $k_{\min} = \lceil \frac{\log_2 N}{\log_2 b} \rceil = \lceil \log_b N \rceil$ copies are needed. The precision parameter ε introduces a correction factor.
2. Embedding Dimension Determination: The embedding dimension per $\mathcal{A}$ copy is determined by the representation theory of $\mathcal{A}$. Generally, there exists a faithful representation $\rho : \mathcal{A} \to \mathrm{End}(\mathbb{R}^d)$, where $d =$ minimal faithful representation dimension$(\mathcal{A})$. Then $d_{\text{embed}} = d \cdot k_{\min}$.
3. Number of Generators: The generators of $\Phi_{\min}$ need to cover all operations in $\Omega_{\mathcal{A}}$. Each copy requires $|\Omega_{\mathcal{A}}|$ generators, totaling $|\Omega_{\mathcal{A}}| \cdot k_{\min}$.
4. Preservation of Optimality: The proof of Theorem D.44 is based on general properties of universal algebra and does not depend on the specific structure of $\mathcal{B}_4$, as long as $\mathcal{A}$ is cognitively complete.

Complexity Analysis: The computational complexity changes from $O(k_{\min}^3)$ to $O(|\Omega_{\mathcal{A}}|^2 \cdot k_{\min}^3)$, because the composition of operations must consider the structure of $\Omega_{\mathcal{A}}$. □

Algorithm 10: Training Algorithm Based on $\mathcal{B}_4^k$ Structure

Input: Training data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, graph structure $\mathcal{G} = (V, E)$
Output: Trained approximate fixed-point combinator $\mathcal{Y}$, prediction function f_θ
begin
  $\theta \leftarrow$ initial parameters (including Enc, Dec, F)
  $\eta \leftarrow$ learning rate
  $T \leftarrow$ number of training epochs
  for $t \leftarrow 1$ **to** T **do**
    Sample batch $\mathcal{B} \subset \mathcal{D}$ **for** $(x, y) \in \mathcal{B}$ **do**
      $\mathbf{H}^{(0)} \leftarrow \mathrm{Enc}(x)$
      $k \leftarrow 0$
      repeat
        $\mathbf{H}^{(k+1)} \leftarrow F_\theta(\mathbf{H}^{(k)})$; `/* Apply graph neural network layer */`
        $k \leftarrow k + 1$;
      until $\|\mathbf{H}^{(k)} - \mathbf{H}^{(k-1)}\| < \epsilon$;
      $\hat{y} \leftarrow \mathrm{Dec}(\mathbf{H}^{(k)})$
      Compute loss $\mathcal{L}(\hat{y}, y)$
      Compute $\nabla_\theta \mathcal{L}$ via unrolled fixed-point iteration
      $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}$
  Define $\mathcal{Y}(f) \leftarrow \lim_{k\to\infty} \mathbf{H}^{(k)}$; `/* Fixed-point combinator learned via training */`
  return $\mathcal{Y}$, f_θ

Theorem D.47 (Algorithm Convergence in the $\mathcal{B}_4^k$ Special Case)**.** *Assume $\mathcal{M} \cong \mathcal{B}_4^{k_{\min}}$, and $\mathcal{Y}$ is an approximate fixed-point combinator on $\mathcal{M}$. Then Algorithm 10 satisfies:*

1. ***Convergence**: For any initial parameters θ_0, the sequence $\{\theta_t\}$ generated by the algorithm converges almost surely to a local optimum θ^*;*
2. ***Convergence Rate**: There exists a constant $C > 0$ such that $\mathbb{E}[\mathcal{L}(\theta_t) - \mathcal{L}(\theta^*)] \le C/t$;*
3. ***Generalization Error Bound**: Let N be the number of training samples, and $VC(\mathcal{F}) = O(k_{\min}^2)$ be the VC dimension of the hypothesis space. Then the generalization error satisfies:*

$$\mathbb{E}_{(x,y)\sim P}[\mathcal{L}(f_{\theta^*}(x), y)] \le \inf_{f\in\mathcal{F}} \mathbb{E}[\mathcal{L}(f(x), y)] + O\left(\sqrt{\frac{k_{\min}^2 \log N}{N}}\right);$$

4. ***Self-Referential Consistency**: The trained $\mathcal{Y}$ satisfies all conditions of an approximate fixed-point combinator (Definition B.4).*

Corollary D.10 (Ultimate Minimal Structure of AGI Machine Learning: Sheaf Version)**.** *Under the constraints of both the AGI axiom system and Axiom D.1, the machine learning mechanism of an AGI system necessarily exhibits a minimal structure: state universe $\mathfrak{S}_{\min} \cong \mathbb{R}^{2k_{\min}}$, meta-dynamics $\Phi_{\min}$ generated by $3k_{\min}$ irreducible vector fields, time parameter Δ_{opt}, semantic encoding $\mathrm{Enc}_{\mathscr{F}}$ as direct mapping encoding, core prediction $\mathrm{MetaPred}$ as self-nested projection operator, adjoint dynamics $\mathscr{F}_{proc}$, initial state $\sigma_{\min}^*$ satisfying informational optimality, computational complexity $C(N) = \Theta(N \cdot k_{\min}^2)$, and generalization error $\varepsilon_{\text{gen}} = O\left(\sqrt{\frac{k_{\min}}{N}}\right)$.*

Proof. Directly from Algorithm 8 and Theorem D.44. The generalization error follows from VC theory: $VC = O(k_{\min})$, hence $\varepsilon_{\text{gen}} = O(\sqrt{VC/N}) = O(\sqrt{k_{\min}/N})$. □

Remark D.4 (Relations among Unified Algorithm Framework)**.** *The above algorithms constitute a complete implementation system for AGI machine learning mechanisms:*

1. *Algorithm 3 is the universal reasoning framework, showcasing the collaboration between task execution and meta-cognition;*
2. *Algorithm 4 provides performance evaluation tools, verifying the quality of prediction mechanisms;*
3. *Algorithm 5 implements efficient meta-prediction, suitable for real-time applications;*
4. *Algorithm 6 is the complete cognitive learning cycle, encompassing perception, reasoning, and action;*
5. *Algorithm 8 constructs minimal AGI systems, achieving theoretically optimal complexity;*
6. *Algorithm 10 is a concrete training algorithm for the $\mathcal{B}_4^k$ special case, demonstrating concrete realization of the theory.*

All algorithms share the same mathematical foundation: the unified cognitive dynamics framework described in Theorem D.38. They respectively implement this framework from different aspects (reasoning, evaluation, prediction, learning, construction, training), collectively forming a complete algorithmic system for AGI machine learning mechanisms.

D.10 Self-Measurement Algorithms

In the unified framework of AGI machine learning mechanisms, self-measurement is not an optional add-on but an intrinsic necessity for the correct operation of the algorithms.

Theorem D.48 (Necessity of Self-Measurement for Fixed-Point Algorithms)**.** *Let $\mathfrak{M} = (\mathscr{F}_\Omega, \Phi, \mathrm{Enc}_{\mathscr{F}}, \rho, \sigma_0)$ be a meta-dynamics layer paradigm satisfying Axiom D.1, and* $\mathrm{MetaPred}$ *be the self-referential prediction module (Theorem D.17). Then for an algorithm to verify the fixed-point condition $\sigma \approx \mathrm{MetaPred}(\sigma)$, it must include the following self-measurement steps:*

1. *State Reading: Obtain the current cognitive state $\sigma_t \in \Gamma(\Omega, \mathscr{F}_\Omega)$*
2. *Meta-Prediction Generation: Compute $\sigma_{meta,t} = \mathrm{MetaPred}(\sigma_t)$*
3. *Difference Metric: Compute $d_t = d(\sigma_t, \sigma_{meta,t})$, where d is the cognitive pseudo-metric (Definition A.1)*
4. *Convergence Judgment: Determine whether $d_t < \epsilon$ holds*
5. *Adjustment Execution: If $d_t \geq \epsilon$, execute $\sigma_{t+1} = \mathit{Integrate}(\sigma_t, \sigma_{meta,t})$*

This measurement process is necessary for algorithm convergence, and each step must be computable in polynomial time.

Proof. **Proof of Necessity**: Assume the algorithm aims to find a fixed point σ^* satisfying $\sigma^* = \mathrm{MetaPred}(\sigma^*)$. Define the sequence $\{\sigma_t\}$ generated by the algorithm. Consider two cases:

1. **Algorithm includes self-measurement**: Then the algorithm can compute $d_t = d(\sigma_t, \mathrm{MetaPred}(\sigma_t))$. By Theorem D.19, $\mathrm{MetaPred}$ iteration converges; there exists T such that for all $t > T$, $d_t < \epsilon$, and the algorithm terminates at an ϵ-fixed point.

2. **Algorithm does not include self-measurement**: Then the algorithm cannot compute d_t and cannot determine whether the current state satisfies the fixed-point condition. Even if the sequence $\{\sigma_t\}$ actually converges to a fixed point, the algorithm cannot know when to terminate. More severely, if the dynamics Φ has multiple fixed points, the algorithm might converge to a wrong fixed point without detection.

Thus, self-measurement is necessary to verify the fixed-point condition and ensure correct algorithm termination.

Proof of Computability:

- State Reading: σ_t is stored in finite memory; reading is $O(|\sigma_t|)$
- Meta-Prediction Generation: By Theorem D.17, $\mathrm{MetaPred}$ is computable with complexity $O(k_{\min}^3)$
- Difference Metric: The cognitive pseudo-metric d is computable, guaranteed by Lipschitz condition
- Convergence Judgment: Scalar comparison, $O(1)$
- Adjustment Execution: The integration function is computable with complexity $O(k_{\min}^2)$

Total complexity is $O(k_{\min}^3)$, polynomial time. □

Algorithm 11: Unified Cognitive Algorithm with Explicit Self-Measurement

Input : Input x, precision threshold $\epsilon > 0$, maximum iteration count $T_{\max}$
Output : Output y, measurement record $\mathcal{M}_{\text{record}}$
begin
  $\sigma \leftarrow \mathrm{Enc}_{\mathscr{F}(x)}$ /* Initial encoding */
  $\mathcal{M}_{\text{record}} \leftarrow \emptyset$ /* Initialize measurement record */
  for $t = 0$ **to** $T_{\max} - 1$ **do**
    Self-Measurement Step 1: State Reading
    $\sigma_{\text{current}} \leftarrow \sigma$ /* Read current state */
    Self-Measurement Step 2: Meta-Prediction Generation
    $\sigma_{\text{meta}} \leftarrow \mathrm{MetaPred}(\sigma_{\text{current}})$
    Self-Measurement Step 3: Difference Metric
    $d_t \leftarrow d(\sigma_{\text{current}}, \sigma_{\text{meta}})$
    Record Measurement Data
    $\mathcal{M}_{\text{record}}.append(\{\sigma_{\text{current}}, \sigma_{\text{meta}}, d_t, t\})$
    Self-Measurement Step 4: Convergence Judgment
    if $d_t < \epsilon$ **then**
      break /* Convergence condition met */
    Self-Measurement Step 5: Adjustment Execution
    $\sigma_{\text{task}} \leftarrow \Phi_\Delta(\sigma_{\text{current}})$ /* Task dynamics */
    $\sigma'_{\text{meta}} \leftarrow \mathrm{MetaPred}(\sigma_{\text{task}})$ /* Updated meta-prediction */
    $\sigma \leftarrow \text{Integrate}(\sigma_{\text{task}}, \sigma'_{\text{meta}})$ /* State integration */
    Performance Self-Measurement (Optional)
    if $t \mod K = 0$ /* Measure performance every K steps */ **then**
      $perf_t \leftarrow \text{measure_performance}(\sigma)$
      $\mathcal{M}_{\text{record}}.append_performance(perf_t)$
  $y \leftarrow \rho(\sigma)$ /* Decode output */
  return $y, \mathcal{M}_{\text{record}}$

Theorem D.49 (Correctness of Algorithm 11)**.** *Assume the inputs of Algorithm 11 satisfy:*

1. $\mathrm{Enc}_{\mathscr{F}}$ *is the optimal semantic encoding (Theorem D.9)*
2. $\mathrm{MetaPred}$ *satisfies the three core properties of Theorem D.25*
3. Φ *is endogenous meta-dynamics (Axiom D.1.3)*

Then the algorithm:

1. ***Terminates Necessarily****: There exists $T \leq T_{\max}$ such that $d_T < \epsilon$*
2. ***Converges to the Correct Fixed Point****: The output σ_T satisfies $d(\sigma_T, \mathrm{MetaPred}(\sigma_T)) < \epsilon$*
3. ***Measurement Record Completeness****: $\mathcal{M}_{record}$ contains a complete history of self-measurements*

4. ***Controlled Complexity**: Each iteration has computational complexity $O(k_{\min}^3)$*

Proof. **Proof of Termination**: By Theorem D.19, the $\mathrm{MetaPred}$ iteration sequence $\{\sigma_t\}$ converges almost surely; there exists T such that $d_t < \epsilon$ for all $t \geq T$. Since the algorithm explicitly computes d_t and checks $d_t < \epsilon$, it must terminate in finite steps.

Proof of Convergence Correctness: Suppose upon termination $\sigma^* = \sigma_T$. By the termination condition $d_T < \epsilon$, we have $d(\sigma^*, \mathrm{MetaPred}(\sigma^*)) < \epsilon$, i.e., σ^* is an ϵ-fixed point. By Theorem D.18, the fixed point of $\mathrm{MetaPred}$ is unique up to isomorphism, so σ^* converges to the correct fixed point.

Measurement Record Completeness: The algorithm explicitly records all measurement data at steps 6, 10, 18, including state, meta-prediction, difference, and timestamp, so the record is complete.

Complexity Analysis:

- State Reading: $O(|\sigma|) = O(k_{\min})$, where $k_{\min}$ is the minimal cognitive dimension
- Meta-Prediction Generation: $O(k_{\min}^3)$ (Theorem D.17)
- Difference Metric: $O(k_{\min}^2)$ (cognitive pseudo-metric computation)
- Integration Function: $O(k_{\min}^2)$
- Performance Measurement (every K steps): $O(k_{\min}^2)$

Total complexity is dominated by $O(k_{\min}^3)$, from meta-prediction generation. □

Definition D.19 (Cognitive State Self-Measurement Function)**.** *Let $\mathfrak{M}$ be a meta-dynamics layer paradigm, and $\mathrm{MetaPred}$ be the self-referential prediction module. Define the **cognitive state self-measurement function** $\mathcal{M}_{state} : \Gamma(\Omega, \mathscr{F}_\Omega) \to \mathbb{R}^{2d} \times \mathbb{R}_{\geq 0}$ as:*

$$\mathcal{M}_{state}(\sigma) = (\varphi(\sigma), \varphi(\mathrm{MetaPred}(\sigma)), d(\sigma, \mathrm{MetaPred}(\sigma)))$$

where:

- *$\varphi : \mathscr{F}_\Omega \to \mathbb{R}^d$ is the coordinate representation of the state vector ($d = 2k_{\min}$)*
- *d is the cognitive pseudo-metric (Definition A.1.3)*

The three components of $\mathcal{M}_{state}$'s output represent: current state coordinates, meta-predicted state coordinates, and cognitive discrepancy.

Definition D.20 (Performance Self-Measurement Function)**.** *Let the task space be $\mathcal{T}$, and performance metric $R : \Gamma(\Omega, \mathscr{F}_\Omega) \times \mathcal{T} \to \mathbb{R}$. Define the **performance self-measurement function** $\mathcal{M}_{perf} : \Gamma(\Omega, \mathscr{F}_\Omega) \times \mathcal{T} \to \mathbb{R}^3$ as:*

$$\mathcal{M}_{perf}(\sigma, \tau) = (R(\sigma, \tau), R(\mathrm{MetaPred}(\sigma), \tau), R(\mathrm{MetaPred}(\sigma), \tau) - R(\sigma, \tau))$$

The three components represent: current performance, predicted optimal performance, and performance gap.

Definition D.21 (Meta Self-Measurement Function)**.** *Define the **meta self-measurement function** $\mathcal{M}_{meta} : \mathcal{H} \to \mathbb{R}^4$, where $\mathcal{H}$ is the measurement history space:*

$$\mathcal{M}_{meta}(\mathcal{H}) = (\textit{avg_discrepancy}, \textit{std_discrepancy}, \textit{measurement_bias}, \textit{measurement_efficiency})$$

representing: average discrepancy, discrepancy standard deviation, measurement bias, and measurement efficiency, respectively.

Theorem D.50 (Self-Measurement Convergence)**.** *Let $\{\mathcal{M}_{state}(\sigma_t)\}_{t=0}^{\infty}$ be the measurement sequence generated by Algorithm 11. Then:*

1. *Discrepancy Monotonic Decrease: $d_{t+1} \leq d_t + O(\Delta^2)$, and there exists a subsequence $\{t_k\}$ such that $d_{t_{k+1}} < d_{t_k}$*
2. *Almost Sure Convergence: $\lim_{t\to\infty} d_t = 0$ holds almost surely*

3. *Convergence Rate: There exists a constant* $0 < \rho < 1$ *such that* $\mathbb{E}[d_t] \leq C\rho^t d_0$

4. *Measurement Bias Decay: The measurement bias* $b_t = \|\mathbb{E}[\mathcal{M}_{\textit{state}}(\sigma_t)] - \mathcal{M}_{\textit{state}}(\sigma_t)\|$ *satisfies* $b_t \leq B/\sqrt{t}$

Proof. **1. Discrepancy Monotonic Decrease**: By Theorem D.19, $\mathrm{MetaPred}$ iteration is contractive; there exists $L < 1$ such that:

$$d(\mathrm{MetaPred}(\sigma), \mathrm{MetaPred}(\sigma')) \leq Ld(\sigma, \sigma')$$

Let $\sigma_{t+1} = \text{Integrate}(\Phi_\Delta(\sigma_t), \mathrm{MetaPred}(\Phi_\Delta(\sigma_t)))$. Since the integration function preserves contractivity, we have:

$$d_{t+1} \leq Ld_t + O(\Delta^2)$$

where $O(\Delta^2)$ is discretization error.

2. Almost Sure Convergence: By the fixed-point theorem for contractive mappings, $\mathrm{MetaPred}$ has a unique fixed point σ^*. The sequence $\{\sigma_t\}$ is generated by a contractive mapping, hence converges almost surely to σ^*, i.e., $\lim_{t\to\infty} d_t = 0$.

3. Convergence Rate: By contractivity, $d_t \leq L^t d_0$. Considering accumulation of discretization error:

$$\mathbb{E}[d_t] \leq L^t d_0 + \sum_{i=0}^{t-1} L^{t-1-i} O(\Delta^2) \leq C\rho^t d_0$$

where $\rho = \max(L, \sqrt{\Delta})$, and C is a constant.

4. Measurement Bias Decay: Measurement bias arises from randomness and approximation error. By the Central Limit Theorem, $b_t = O(1/\sqrt{t})$. More precisely, by the Berry-Esseen theorem:

$$b_t \leq \frac{B}{\sqrt{t}} + O\left(\frac{1}{t}\right)$$

where B is a constant related to the third moment of the distribution. □

Theorem D.51 (Lower Bound on Self-Measurement Complexity)**.** *For any meta-dynamics layer paradigm implementing Axiom D.1, the computational complexity lower bound of its self-measurement process is:*

$$\Omega(k_{\min}^3)$$

where $k_{\min}$ *is the minimal cognitive dimension. Specifically:*

1. *State reading complexity:* $\Omega(k_{\min})$
2. *Meta-prediction computation complexity:* $\Omega(k_{\min}^3)$ *(dominated by matrix inversion)*
3. *Difference metric complexity:* $\Omega(k_{\min}^2)$
4. *Self-monitoring complexity of the measurement process:* $\Omega(k_{\min}^2 \log k_{\min})$

Proof. **Lower Bound for Meta-Prediction Computation**: $\mathrm{MetaPred}$ involves natural gradient computation $\tilde{\nabla} I(\eta) = g^{-1}(\eta)\nabla I(\eta)$, where $g(\eta)$ is the $d \times d$ Fisher information matrix, $d = 2k_{\min}$. The complexity lower bound for matrix inversion is $\Omega(d^3) = \Omega(k_{\min}^3)$ (based on algebraic complexity theory).

Lower Bound for Difference Metric: The cognitive pseudo-metric $d(\sigma, \sigma')$ involves comparison and distance computation of d-dimensional vectors, with lower bound $\Omega(d) = \Omega(k_{\min})$. However, KL divergence computation requires $O(d^2)$ operations.

Total Complexity: Summing the complexities, the dominant term is $O(k_{\min}^3)$, and this is a theoretical lower bound that any implementation cannot circumvent. □

The traditional self-measurement paradox fundamentally states: to measure its own state M, the system must alter M during measurement, preventing access to the true pre-measurement state. The self-measurement paradox is naturally resolved through procedural measurement, fixed measured state, and asymptotic computation:

- Procedural Measurement: Measurement is a process, not an instant;
- Fixed Measured State: The state remains unchanged during measurement (due to ergodicity, the state is approximately constant during measurement);
- Asymptotic Computation: Measurement results are obtained via asymptotic iteration, avoiding the paradox.

Theorem D.52 (Asymptotic Algorithm for Self-Measurement)**.** *A system's measurement of its own state $x \in \mathcal{M}$ can be implemented by the following asymptotic algorithm:*

$$\mathrm{Measure}(x) = \lim_{n\to\infty} \mathrm{Approx}_n(x)$$

where Approx_n *is defined by the iteration:*

1. $\sigma_0 =$ *random process state*
2. $\omega_0 =$ *approximation of* $\mathrm{Enc}_{sem}(x)$
3. *For* $t = 1$ *to* n*:*
   - *Select focus* c_t *according to* σ_{t-1}
   - *Read the* c_t*-th component of* x
   - *Update* ω_t *and* σ_t

This algorithm converges at rate $O(1/\sqrt{n})$.

Theorem D.53 (Paradox Resolution via Progressive Measurement)**.** *The self-measurement paradox can be resolved by progressive measurement methods. Define the progressive measurement operator:*

$$\textit{Measure}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \textit{Approx}_i(x)$$

where $\textit{Approx}_i(x)$ *is the* i*-th approximate measurement, satisfying:*

1. $\lim_{n\to\infty} \textit{Measure}_n(x) = \textit{TrueValue}(x)$
2. *Each approximate measurement perturbs the state by* $O(1/n)$
3. *Total perturbation* $\lim_{n\to\infty} \textit{TotalDisturbance}(n) = 0$

Proof. Let the true state be x^*, the state at the i-th measurement be x_i, and the measurement result be $y_i = f(x_i) + \epsilon_i$, where f is the measurement function and ϵ_i is measurement noise.

The progressive measurement estimate is:

$$\hat{x}_n = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}[f(x_i) + \epsilon_i]$$

Since each measurement perturbation is $O(1/n)$, we have $x_i = x^* + O(i/n^2)$. Assuming f is Lipschitz continuous:

$$f(x_i) = f(x^*) + O\left(\frac{i}{n^2}\right)$$

Thus:

$$\hat{x}_n = f(x^*) + \frac{1}{n}\sum_{i=1}^{n} O\left(\frac{i}{n^2}\right) + \frac{1}{n}\sum_{i=1}^{n}\epsilon_i = f(x^*) + O\left(\frac{1}{n}\right) + \frac{1}{n}\sum_{i=1}^{n}\epsilon_i$$

As $n \to \infty$, $O(1/n) \to 0$, and by the law of large numbers $\frac{1}{n}\sum_{i=1}^{n}\epsilon_i \to 0$ (if ϵ_i has zero mean), hence $\lim_{n\to\infty}\hat{x}_n = f(x^*)$.

Total perturbation: $\sum_{i=1}^{n}\|x_i - x^*\| = \sum_{i=1}^{n} O(i/n^2) = O(1/n) \to 0$. □

Algorithm 12: Measurement-Prediction Separation Self-Measurement Algorithm

Input : Initial state σ_0, historical data $\mathcal{H}$, prediction model $\mathcal{P}$
Output : Measurement result $\hat{\sigma}$, confidence c
begin
  Phase 1: History-Based Prediction
  $\sigma_{\text{pred}} \leftarrow \mathcal{P}(\mathcal{H})$ /* Predict current state based on history */
  Phase 2: Lightweight Verification Measurement
  for $i = 1$ **to** m /* $m \ll n$, few measurements */
  do
    $\sigma_{\text{sample},i} \leftarrow$ sample_measurement(σ_0) /* Sampling measurement, small perturbation */
    $w_i \leftarrow$ confidence$(\sigma_{\text{sample},i})$
  Phase 3: Prediction-Measurement Fusion
  $\hat{\sigma} \leftarrow \lambda\sigma_{\text{pred}} + (1-\lambda)\frac{\sum_{i=1}^{m} w_i \sigma_{\text{sample},i}}{\sum_{i=1}^{m} w_i}$
  Phase 4: Confidence Calculation
  $c \leftarrow$ compute_confidence$(\sigma_{\text{pred}}, \{\sigma_{\text{sample},i}\}, \mathcal{H})$
  return $\hat{\sigma}, c$

Definition D.22 (Procedural Measurement)**.** ***Procedural measurement*** *is a dynamical process, not an instantaneous operation. Define the measurement process* $\mathcal{P}_{meas} : \Gamma(\Omega, \mathscr{F}_\Omega) \times \mathbb{R}_{\geq 0} \to \Gamma(\Omega, \mathscr{F}_\Omega) \times \mathcal{M}$*:*

$$\mathcal{P}_{meas}(\sigma_0, T) = (\sigma_T, \{\mathcal{M}_t\}_{t=0}^{T})$$

where:

- σ_t *is the state evolution during the measurement process*
- $\mathcal{M}_t$ *is the measurement record at time* t
- *The process satisfies* $\lim_{T\to\infty} \sigma_T = \sigma^*$ *(true state)*
- *Measurement influence decays gradually:* $\|\sigma_t - \sigma^*\| \leq Ce^{-\alpha t}$

Theorem D.54 (Effectiveness of Procedural Measurement)**.** *Assume the measurement process follows Langevin dynamics:*

$$d\sigma_t = -\nabla V(\sigma_t)dt + \sqrt{2\beta^{-1}}dW_t$$

where V *is the measurement potential function, satisfying* $V(\sigma^*) = 0$ *and* $\nabla^2 V(\sigma^*) \succ 0$*. Then:*

1. *Process convergence:* $\lim_{t\to\infty} \mathbb{E}[\sigma_t] = \sigma^*$
2. *Exponential convergence rate:* $\mathbb{E}[\|\sigma_t - \sigma^*\|^2] \leq Ce^{-\lambda t}$
3. *Measurement error decay: Measurement error* $e_t = \|\mathcal{M}_t - \sigma^*\|$ *satisfies* $\lim_{t\to\infty} e_t = 0$

Proof. This is a standard convergence result for Langevin dynamics. By strong convexity of the potential function V, the dynamics converge exponentially to the equilibrium distribution, whose mean is σ^*. For a detailed proof, refer to convergence theory of stochastic differential equations. □

Definition D.23 (Optimal Measurement Frequency)**.** *Let the Lyapunov time of the system state evolution be* τ_L*, and the measurement cost be* c_m*. Then the* ***optimal measurement frequency*** *is:*

$$f_{opt} = \frac{1}{\tau_L\sqrt{1 + \frac{c_m}{\tau_L}}}$$

where the Lyapunov time τ_L *is defined as the reciprocal of the largest eigenvalue of the state-dependent Jacobian:*

$$\tau_L(\sigma) = \frac{1}{\max_i |\lambda_i(\nabla\Phi(\sigma))|}$$

Algorithm 13: Adaptive Measurement Frequency Scheduler

Input : Initial frequency f_0, learning rate α, minimum frequency $f_{\min}$, maximum frequency $f_{\max}$

Output : Measurement decision sequence $\{measure_t\}_{t=0}^{T}$

begin

$f \leftarrow f_0$ /* Current measurement frequency */

$\tau_L \leftarrow 1.0$ /* Lyapunov time estimate */

$c_m \leftarrow 0.0$ /* Measurement cost estimate */

for $t = 0$ **to** $T - 1$ **do**

Estimate Lyapunov Time

$J_t \leftarrow \text{estimate_jacobian}(\sigma_t)$ /* Estimate dynamics Jacobian */

$\lambda_{\max} \leftarrow \max |\text{eig}(J_t)|$

$\tau_L^{\text{new}} \leftarrow 1.0/\max(\lambda_{\max}, \epsilon)$

$\tau_L \leftarrow (1-\alpha)\tau_L + \alpha\tau_L^{\text{new}}$

Estimate Measurement Cost

$c_m^{\text{new}} \leftarrow \text{measure_computation_cost}()$

$c_m \leftarrow (1-\alpha)c_m + \alpha c_m^{\text{new}}$

Compute Optimal Frequency

$f_{\text{opt}} \leftarrow 1.0/(\tau_L\sqrt{1.0 + c_m/\tau_L})$

Frequency Smoothing

$f \leftarrow (1-\alpha)f + \alpha f_{\text{opt}}$

$f \leftarrow \text{clamp}(f, f_{\min}, f_{\max})$

Decide Whether to Measure

$measure_t \leftarrow (\text{random}() < f)$ /* Measure with probability f */

if $measure_t$ **then**

Execute full self-measurement (Steps 2-5 of Algorithm 11)

Update measurement cost estimate

return $\{measure_t\}$

Theorem D.55 (Optimality of Adaptive Frequency)**.** *The measurement frequency sequence $\{f_t\}$ generated by Algorithm 13 converges almost surely to the optimal frequency f_{opt}, and is optimal in the following sense:*

1. ***Maximized Measurement Efficiency****: Maximizes information acquisition under a given computational budget*
2. ***Minimized State Perturbation****: Minimizes interference of the measurement process on the state*
3. ***Optimized Convergence Rate****: Balances measurement cost and convergence speed, achieving Pareto optimality*

Proof. **Convergence**: The frequency update rule is a stochastic approximation algorithm:

$$f_{t+1} = f_t + \alpha_t(f_{\text{opt},t} - f_t) + \xi_t$$

where ξ_t is a noise term. Under standard stochastic approximation conditions ($\sum \alpha_t = \infty$, $\sum \alpha_t^2 < \infty$), f_t converges almost surely to a fixed point of f_{opt}.

Proof of Optimality: Consider the optimization problem:

$$\max_f \mathbb{E}[\text{InformationGain}(f)] \quad \text{s.t.} \quad \mathbb{E}[\text{ComputationCost}(f)] \leq B$$

By the Lagrangian method, the optimal solution satisfies the first-order condition, which is the f_{opt} formula in the algorithm. The algorithm adaptively adjusts to converge to this optimal solution. □

Algorithm 14: Hierarchical Self-Measurement Architecture

Input : State σ, task τ, measurement budget B
Output : Hierarchical measurement result $\mathcal{M}_{\text{hier}}$
begin
  Layer 1: Quick State Check (cost $c_1 = O(k_{\min})$)
  $d_{\text{quick}} \leftarrow$ quick_discrepancy_check(σ)
  $\mathcal{M}_1 \leftarrow \{\text{quick_check} : d_{\text{quick}}\}$
  if $d_{quick} > \theta_1$ /* Requires detailed measurement */ **then**
    Layer 2: Standard Self-Measurement (cost $c_2 = O(k_{\min}^2)$)
    $\mathcal{M}_2 \leftarrow$ standard_self_measurement(σ) /* Steps 2-5 of Algorithm 11 */
    if $\mathcal{M}_2.discrepancy > \theta_2$ /* Requires meta-layer measurement */ **then**
      Layer 3: Meta Self-Measurement (cost $c_3 = O(k_{\min}^3)$)
      $\mathcal{M}_3 \leftarrow$ meta_self_measurement($\sigma, \mathcal{M}_2$)
      if $\mathcal{M}_3.measurement_bias > \theta_3$ /* Requires calibration */ **then**
        Layer 4: Calibration Measurement (cost $c_4 = O(k_{\min}^3 \log k_{\min})$)
        $\mathcal{M}_4 \leftarrow$ calibration_measurement($\sigma, \mathcal{M}_3$)
  Integrate Measurement Results
  $\mathcal{M}_{\text{hier}} \leftarrow$ integrate_measurements($\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4$)
  Budget Check and Adjustment
  $total_cost \leftarrow \sum_i c_i$
  if $total_cost > B$ **then**
    Adjust thresholds θ_i to reduce future measurement cost
  return $\mathcal{M}_{\text{hier}}$

Theorem D.56 (Optimality of Hierarchical Measurement)**.** *Let the total measurement budget be B, the cost of each layer be c_i, and the information gain be I_i. Algorithm 14 is optimal in the following sense:*

1. ***Budget Constraint Satisfaction**: $\mathbb{E}[\sum_i c_i] \leq B$*
2. ***Maximized Information Gain**: Maximizes $\sum_i w_i I_i$, where w_i are layer weights*
3. ***Optimal Decision Making**: The choice of thresholds θ_i minimizes expected loss*

Proof. Consider a dynamic programming problem: the state is the current measurement result, and the decision is whether to proceed to the next layer. The optimal policy satisfies the Bellman optimality equation. The threshold rule in the algorithm approximates the optimal policy.

Let the loss function be $L(\theta) = \mathbb{E}[\text{MeasurementError}] + \lambda \mathbb{E}[\text{Cost}]$. The optimal threshold θ^* minimizes L. The algorithm approximates θ^* through empirical adjustment. □

Remark D.5 (Theoretical Significance and Practical Implications of Self-Measurement)**.** ***Theoretical Significance***

1. *Inevitable Corollary of Fixed-Point Structure: Self-measurement is not an optional feature of algorithms but a necessary condition for verifying fixed-point conditions of the form $x = f(x)$. Any system implementing full learnability must contain a self-measurement mechanism.*
2. *Physical Realization of Self-Reference: Self-measurement provides a concrete physical realization mechanism for self-reference, transforming the abstract concept of self-reference into a computable measurement process.*
3. *Operationalization of Cognitive Closure: The cognitive closure theorem operationally requires the system to measure its own cognitive state, which is a key step toward achieving cognitive completeness.*

4. *Self-Monitoring of Complex Systems: Self-measurement is the foundational mechanism for complex systems to achieve self-monitoring, self-repair, and self-optimization.*

Practical Implications

1. *Mandatory Algorithm Design: When designing AGI algorithms, self-measurement steps must be explicitly included; one cannot assume the system "naturally knows" its own state.*
2. *Computational Resource Allocation: Self-measurement consumes significant computational resources ($O(k_{\min}^3)$). In system design, dedicated computational units and storage resources must be allocated for it.*
3. *Measurement-Learning Balance: A balance must be struck between measurement accuracy and learning efficiency. Excessive measurement wastes resources; insufficient measurement leads to non-convergence.*
4. *Gradual Implementation Strategy: In practice, a progressive measurement strategy should be adopted, starting from simple measurements and gradually upgrading to complex ones, with calibration as needed.*
5. *Hardware Support Requirements: Efficient self-measurement may require specialized hardware support, such as fast matrix computation units, dedicated measurement registers, etc.*

E Machine Learning Mechanisms in Multimodal AGI Systems: Approximate Sheaf Formulation

E.1 Sheaf Structure of Multimodal Cognitive Domains

Definition E.1 (Multimodal Cognitive Domain)**.** *Let the objective world $\mathcal{W}$ be decomposable as a union of modalities: $\mathcal{W} = \bigcup_{\alpha \in A} \mathcal{W}_\alpha$, where A is a countable index set, and $\mathcal{W}_\alpha$ is the carrier set of modality α. Define the **multimodal sheaf** $\mathscr{M}_{multi}$ as an approximate sheaf on A (with discrete topology):*

1. *For a singleton $\{\alpha\}$, define $\mathscr{M}_{multi}(\{\alpha\}) = \mathcal{W}_\alpha$;*
2. *For any open set $U \subseteq A$, define $\mathscr{M}_{multi}(U) = \prod_{\alpha \in U} \mathcal{W}_\alpha$;*
3. *Restriction maps are projections: for $V \subseteq U$, $\rho_{UV} : \mathscr{M}_{multi}(U) \to \mathscr{M}_{multi}(V)$ is the projection onto components in V;*
4. ***Modal Coordination Relations**: There exists a family of approximate homomorphisms $\mathcal{R} = \{\mathcal{R}_{\alpha\beta} : \mathscr{M}_{multi,\alpha} \xrightarrow{\varepsilon_{\alpha\beta}} \mathscr{M}_{multi,\beta}\}_{\alpha,\beta \in A}$ satisfying:*
   (a) ***Reflexivity**: $\mathcal{R}_{\alpha\alpha} \xrightarrow{0} \mathrm{id}_{\mathscr{M}_{multi,\alpha}}$;*
   (b) ***Symmetry**: $\mathcal{R}_{\beta\alpha} \xrightarrow{\varepsilon_{\alpha\beta}} \mathcal{R}_{\alpha\beta}^{-1}$;*
   (c) ***Transitivity**: $\mathcal{R}_{\beta\gamma} \circ \mathcal{R}_{\alpha\beta} \xrightarrow{\varepsilon_{\alpha\beta}+\varepsilon_{\beta\gamma}+\delta_{\alpha\beta\gamma}} \mathcal{R}_{\alpha\gamma}$, where $\delta_{\alpha\beta\gamma} \geq 0$, and for any finite modality set A_0, $\sum_{\alpha,\beta,\gamma \in A_0} \delta_{\alpha\beta\gamma} < C_{A_0}$ (finite constant);*
   (d) ***Error Control**: There exists a monotone decreasing function $\psi : \mathbb{N} \to \mathbb{R}^+$, $\psi(n) \to 0$, such that for any finite modality set $A_0 \subseteq A$, $\sup_{\alpha,\beta \in A_0} \varepsilon_{\alpha\beta} \leq \psi(|A_0|)$.*

Definition E.2 (Perception and Action Constraints of Multimodal AGI Systems: Sheaf Version)**.** *Let $\mathcal{M} \in \mathbf{AGI}$ be an AGI system. Its interaction with the world is subject to the following constraints:*

1. ***Finite Modality Perception**: There exists a finite perception modality set $G \subseteq A$, such that there are perception homomorphism sheaves $\pi_g : \mathscr{W}_g \to \mathscr{M}$, i.e., for each $g \in G$, there exists an approximate homomorphism $\pi_g : \mathscr{W}_g \xrightarrow{\varepsilon_g} \mathcal{M}$;*
2. ***Finite Modality Action**: There exists a finite action modality set $X \subseteq A$, such that there are action homomorphism sheaves $\sigma_x : \mathscr{M} \to \mathscr{W}_x$, i.e., for each $x \in X$, there exists an approximate homomorphism $\sigma_x : \mathcal{M} \xrightarrow{\delta_x} \mathscr{W}_x$;*
3. ***Perception-Action Modality Asymmetry**: G and X need not coincide; there may be modalities that can only be perceived but not acted upon, or only acted upon but not perceived.*

Definition E.3 (Cognitive Hierarchy Sheaf)**.** *Let $\{\mathbb{H}_n\}_{n=0}^{\infty}$ be the cognitive hierarchy sequence from Definition A.13. Define the **cognitive hierarchy sheaf** $\mathscr{H}$ as an approximate sheaf on the topological space $(\mathbb{N}, \tau_{cofinite})$:*

1. *For a cofinite open set $U \subseteq \mathbb{N}$, define $\mathscr{H}(U) = \varprojlim_{n \in U} \mathbb{H}_n$;*
2. *$\mathscr{H}(\mathbb{N}) = \mathbb{U} = \varinjlim \mathbb{H}_n$;*
3. *Restriction maps are naturally induced by the connecting approximate homomorphisms μ_n.*

Axiom E.1 (Foundational Axioms for Multimodal AGI: Sheaf Version)**.** *In addition to Axiom A.1, include the following multimodal constraints:*

1. ***Multimodal Cognitive Hierarchy Sheaf**: The sections of the cognitive hierarchy sheaf $\mathscr{H}$ at level n can be decomposed into components for each modality, i.e., for each modality $\alpha \in A_n$, there exists an approximate homomorphism $\pi_n^\alpha : \mathcal{M} \xrightarrow{\varepsilon_n^\alpha} \mathscr{H}_n^\alpha$;*
2. ***Modal Approximate Consistency**: For any $\alpha, \beta \in A_n$, there exists $\eta_n^{\alpha\beta} \geq 0$ such that $\pi_n^\beta \circ \mathcal{R}_{\alpha\beta} \xrightarrow{\eta_n^{\alpha\beta}} \pi_n^\alpha$, and $\sum_n \sup_{\alpha,\beta \in A_n} \eta_n^{\alpha\beta} < \infty$;*

3. ***Perception-Action Modal Approximate Coordination****: There exists a mapping $\theta : G \to X$ and errors $\zeta_g \geq 0$ such that $\sigma^{\theta(g)} \circ \pi^g \xrightarrow{\zeta_g} \mathrm{id}_{\mathscr{W}_g}$, and $\sup_{g \in G} \zeta_g < \infty$ (uniform bounded convergence).*

Definition E.4 (Modal Coherence Relation)**.** *Let $\{\mathscr{W}_\alpha\}_{\alpha \in A}$ be a family of modality sheaves. A **modal coherence relation** is a family of approximate homomorphisms $\mathcal{R} = \{\mathcal{R}_{\alpha\beta} : \mathscr{W}_\alpha \xrightarrow{\varepsilon_{\alpha\beta}} \mathscr{W}_\beta\}_{\alpha,\beta \in A}$ satisfying the reflexivity, symmetry, transitivity, and error control conditions of Definition E.1.*

E.2 Necessary Existence of Supermodal and Prefabricated Structure Theorem

Definition E.5 (Supermodal: Sheaf Version)**.** *Let $\{\mathscr{W}_\alpha\}_{\alpha \in A}$ be a family of modality sheaves. A sheaf $\mathscr{W}_{super}$ is called a **supermodal sheaf** if:*

1. *$\mathscr{W}_{super}$ is an approximate sheaf;*
2. *There exists a family of approximate sheaf morphisms $\{\varphi_\alpha : \mathscr{W}_\alpha \to \mathscr{W}_{super}\}_{\alpha \in A}$;*
3. *It satisfies the universal property: for any approximate sheaf $\mathscr{B}$ and family of approximate sheaf morphisms $\{\psi_\alpha : \mathscr{W}_\alpha \to \mathscr{B}\}_{\alpha \in A}$, there exists an approximate sheaf morphism $\psi : \mathscr{W}_{super} \to \mathscr{B}$ such that $\psi \circ \varphi_\alpha \approx \psi_\alpha$ (in the approximate sense).*

Lemma E.1 (Transitive Closure of Modal Coherence Relations)**.** *Let $\mathcal{R} = \{\mathcal{R}_{\alpha\beta}\}$ be a modal coherence relation as in Definition E.1. Then there exists an extended relation $\tilde{\mathcal{R}}$ satisfying:*

1. *$\tilde{\mathcal{R}}_{\alpha\beta} \approx \mathcal{R}_{\alpha\beta}$*
2. *$\tilde{\mathcal{R}}$ is an ε-approximate equivalence relation: $\tilde{\mathcal{R}}_{\beta\gamma} \circ \tilde{\mathcal{R}}_{\alpha\beta} \xrightarrow{\tilde{\varepsilon}_{\alpha\gamma}} \tilde{\mathcal{R}}_{\alpha\gamma}$, and $\tilde{\varepsilon}_{\alpha\gamma} \leq \sum_{i=0}^{n-1} \varepsilon_{\alpha_i \alpha_{i+1}}$ ($\alpha = \alpha_0, \ldots, \alpha_n = \beta$)*
3. *The minimal error realization of $\tilde{\mathcal{R}}$ can be obtained via dynamic programming*

Proof. **Construction Process**: Define the extended relation $\tilde{\mathcal{R}}$ as follows: for any $\alpha, \beta \in A$, define $\tilde{\mathcal{R}}_{\alpha\beta}$ as the mapping among all composites of approximate homomorphisms corresponding to paths $\alpha = \alpha_0, \alpha_1, \ldots, \alpha_n = \beta$ that minimizes the total error, i.e.,

$$\tilde{\mathcal{R}}_{\alpha\beta} = \arg \min_{\pi \in \mathcal{P}_{\alpha\beta}} (\mathcal{R}_\pi, \varepsilon_\pi)$$

where $\mathcal{P}_{\alpha\beta}$ is the set of all finite paths from α to β, $\mathcal{R}_\pi = \mathcal{R}_{\alpha_{n-1}\alpha_n} \circ \cdots \circ \mathcal{R}_{\alpha_0\alpha_1}$, and $\varepsilon_\pi = \sum_{i=0}^{n-1} \varepsilon_{\alpha_i \alpha_{i+1}}$.

Verification of Properties:

1. **Approximation**: By construction, $\tilde{\mathcal{R}}_{\alpha\beta}$ is a composite of finitely many approximate homomorphisms, hence still an approximate homomorphism, and $\tilde{\mathcal{R}}_{\alpha\beta} \xrightarrow{\varepsilon_{\alpha\beta}} \mathcal{R}_{\alpha\beta}$ (taking path length 1).
2. **Transitivity**: For any α, β, γ, consider path concatenation. Let $\pi_{\alpha\beta}$ and $\pi_{\beta\gamma}$ be optimal paths from $\alpha \to \beta$ and $\beta \to \gamma$ respectively. Then the concatenated path $\pi_{\alpha\gamma}$ has total error not exceeding $\varepsilon_{\pi_{\alpha\beta}} + \varepsilon_{\pi_{\beta\gamma}}$. By minimality, the error $\tilde{\varepsilon}_{\alpha\gamma}$ of $\tilde{\mathcal{R}}_{\alpha\gamma}$ satisfies $\tilde{\varepsilon}_{\alpha\gamma} \leq \varepsilon_{\pi_{\alpha\beta}} + \varepsilon_{\pi_{\beta\gamma}}$. In particular, taking $\pi_{\alpha\beta}$ and $\pi_{\beta\gamma}$ as optimal paths, we have

   $$\tilde{\mathcal{R}}_{\beta\gamma} \circ \tilde{\mathcal{R}}_{\alpha\beta} \xrightarrow{\tilde{\varepsilon}_{\alpha\beta} + \tilde{\varepsilon}_{\beta\gamma}} \tilde{\mathcal{R}}_{\alpha\gamma}$$

   and $\tilde{\varepsilon}_{\alpha\gamma} \leq \tilde{\varepsilon}_{\alpha\beta} + \tilde{\varepsilon}_{\beta\gamma}$.
3. **Minimal Error Realization**: Since A_0 (the initial finite modality set) is finite, the number of paths is finite. Thus, $\tilde{\mathcal{R}}$ can be computed via a dynamic programming algorithm:
   (a) Initialize: $\tilde{\mathcal{R}}_{\alpha\alpha} = \mathrm{id}$, $\tilde{\varepsilon}_{\alpha\alpha} = 0$; for $\alpha \neq \beta$, if there is a direct coordination relation, initialize with $\mathcal{R}_{\alpha\beta}$ and $\varepsilon_{\alpha\beta}$, otherwise set to ∞.

(b) Iterative update: For each pair (α, β), check all intermediate nodes γ; if $\tilde{\varepsilon}_{\alpha\gamma} + \tilde{\varepsilon}_{\gamma\beta} < \tilde{\varepsilon}_{\alpha\beta}$, update $\tilde{\mathcal{R}}_{\alpha\beta} = \tilde{\mathcal{R}}_{\gamma\beta} \circ \tilde{\mathcal{R}}_{\alpha\gamma}$ and $\tilde{\varepsilon}_{\alpha\beta} = \tilde{\varepsilon}_{\alpha\gamma} + \tilde{\varepsilon}_{\gamma\beta}$.

(c) Continue until no updates occur.

This algorithm terminates in $O(|A_0|^3)$ time.

□

Theorem E.1 (Existence and Uniqueness of Supermodal Sheaf: Sheaf Version)**.** *Let $\{\mathscr{W}_\alpha\}_{\alpha\in A}$ be a family of modality sheaves satisfying the modal coherence relation $\mathcal{R}$ from Definition E.1. Then there exists an approximate sheaf $\mathscr{W}_{\text{super}}$ (called the supermodal sheaf) and a family of approximate sheaf morphisms $\{\varphi_\alpha : \mathscr{W}_\alpha \to \mathscr{W}_{\text{super}}\}_{\alpha\in A}$, satisfying the following universal property: for any approximate sheaf $\mathscr{B}$ and family of approximate sheaf morphisms $\{\psi_\alpha : \mathscr{W}_\alpha \to \mathscr{B}\}_{\alpha\in A}$, if for any α, β we have $\psi_\beta \circ \mathcal{R}_{\alpha\beta} \xrightarrow{\varepsilon_{\alpha\beta}} \psi_\alpha$, then there exists a unique approximate sheaf morphism $\psi : \mathscr{W}_{\text{super}} \to \mathscr{B}$ such that for any α, $\psi \circ \varphi_\alpha \xrightarrow{\delta_\alpha} \psi_\alpha$, where $\delta_\alpha = O(\sup_\beta \varepsilon_{\alpha\beta})$.*

Moreover, the supermodal sheaf is unique up to isomorphism and can be constructed as:

$$\mathscr{W}_{\text{super}} = \mathbf{a}\left(\coprod_{\alpha\in A} \mathscr{W}_\alpha / \sim_{\mathcal{R}}\right)$$

where $\mathbf{a}$ is the sheafification functor, and $\sim_{\mathcal{R}}$ is the approximate equivalence relation generated by $\mathcal{R}$.

Proof. **Construction**: Define a presheaf P: for an open set $U \subseteq A$,

$$P(U) = \left\{ (s_\alpha)_{\alpha\in U} \in \prod_{\alpha\in U} \mathscr{W}_\alpha \mid \forall \alpha, \beta \in U, \mathcal{R}_{\alpha\beta}(s_\alpha) \xrightarrow{\varepsilon_{\alpha\beta}} s_\beta \right\}.$$

Take the sheafification $\mathscr{W}_{\text{super}} = \mathbf{a}(P)$. Define $\varphi_\alpha(x)$ as the section that is x at α and $\mathcal{R}_{\alpha\beta}(x)$ at β.

Universal Property: Given $\{\psi_\alpha\}$, define ψ on generators as $\psi(\varphi_\alpha(x)) = \psi_\alpha(x)$. Well-definedness is verified by the coherence conditions.

Uniqueness: By the universal property, any two supermodal sheaves satisfying the conditions are approximately isomorphic. □

Remark E.1 (Cognitive Significance of Supermodal)**.** *The necessary existence of the supermodal sheaf $\mathscr{W}_{\text{super}}$ indicates that to coordinate multimodal information, an AGI system must internally construct a unified "global workspace" that does not directly correspond to any single modality but is a coordinated fusion of all modalities. This corresponds to the "global workspace theory" in cognitive science. The supermodal prediction module corresponds to the "self-monitoring" capability in metacognition.*

Lemma E.2 (Cognitive Necessity of Synchronous Data)**.** *If synchronous data $\mathcal{D}_{\text{sync}}$ is lacking, consistent translation functions $T_{\alpha\beta}$ cannot be learned, leading to failure of the coordination relation to satisfy transitivity and thus failure of the quotient space construction.*

Proof. Proof by contradiction.

Assume there exists an algorithm that can learn consistent translations without synchronous data. Consider three modalities α, β, γ. Asynchronous learning may produce $T_{\alpha\beta}$, $T_{\beta\gamma}$, $T_{\alpha\gamma}$ such that $T_{\beta\gamma} \circ T_{\alpha\beta} \neq T_{\alpha\gamma}$. When attempting to construct the equivalence relation, a contradiction arises: $z_\alpha \sim z_\beta$ and $z_\beta \sim z_\gamma$ but $z_\alpha \not\sim z_\gamma$, violating transitivity. Therefore, synchronous data is a necessary cognitive foundation. □

Theorem E.2 (Constructive Origin Theorem of Supermodal)**.** *Let $\mathcal{M} \in \mathbf{AGI}$ initially have access only to a finite modality set $A_0 = \{\alpha_1, \ldots, \alpha_m\}$, each modality α corresponding to a perception-action pair $(\pi_\alpha : \mathcal{W}_\alpha \to \mathcal{M}, \sigma_\alpha : \mathcal{M} \to \mathcal{W}_\alpha)$. Assume there exists a synchronous data stream $\mathcal{D}_{\text{sync}} = \{(x_{\alpha_1}^{(t)}, \ldots, x_{\alpha_m}^{(t)})\}_{t=1}^T$, where samples from different modalities at time t are aligned (corresponding to the same event). Then:*

1. *Modal Encoding Learning: There exists a self-supervised algorithm $\mathcal{A}_{enc}$ that learns for each modality α an encoder $E_\alpha : \mathcal{W}_\alpha \to \mathbb{R}^{d_\alpha}$ and decoder $D_\alpha : \mathbb{R}^{d_\alpha} \to \mathcal{W}_\alpha$, minimizing reconstruction loss:*

$$\min_{E_\alpha, D_\alpha} \mathbb{E}_{x \sim \mathcal{W}_\alpha} \|D_\alpha(E_\alpha(x)) - x\|$$

2. *Cross-Modal Translation Learning: There exists an algorithm $\mathcal{A}_{trans}$ that learns translation functions $T_{\alpha\beta} : \mathbb{R}^{d_\alpha} \to \mathbb{R}^{d_\beta}$, minimizing synchronous alignment loss:*

$$\min_{T_{\alpha\beta}} \mathbb{E}_{(x_\alpha, x_\beta) \sim \mathcal{D}_{sync}} \|T_{\alpha\beta}(E_\alpha(x_\alpha)) - E_\beta(x_\beta)\|$$

3. *Induced Coordination Relation: Define the induced coordination relation $\tilde{\mathcal{R}}_{\alpha\beta} = D_\beta \circ T_{\alpha\beta} \circ E_\alpha$. Then $\tilde{\mathcal{R}}$ satisfies:*

   - $\tilde{\mathcal{R}}_{\alpha\alpha} \approx \mathrm{id}_{\mathcal{W}_\alpha}$
   - $\tilde{\mathcal{R}}_{\beta\alpha} \circ \tilde{\mathcal{R}}_{\alpha\beta} \approx \mathrm{id}$ *(approximately invertible)*
   - $\tilde{\mathcal{R}}_{\gamma\alpha} \circ \tilde{\mathcal{R}}_{\beta\gamma} \circ \tilde{\mathcal{R}}_{\alpha\beta} \approx \tilde{\mathcal{R}}_{\alpha\gamma}$ *(approximately transitive)*

4. *Supermodal Space Construction: Define an equivalence relation $\sim$ on $\bigsqcup_\alpha \mathbb{R}^{d_\alpha}$:*

$$z_\alpha \sim z_\beta \iff \|T_{\alpha\beta}(z_\alpha) - z_\beta\| < \varepsilon$$

   The quotient space $\mathcal{Z}_{super} = \left(\bigsqcup_\alpha \mathbb{R}^{d_\alpha}\right) / \sim$ naturally becomes a metric space.

5. *Supermodal Encoder: There exists an encoder $E_{super} : \mathcal{Z}_{super} \to \mathbb{R}^{d_{super}}$ satisfying:*

   - *Compatibility: For each modality α, there is an approximate commutative diagram:*

$$\begin{array}{ccc} \mathcal{W}_\alpha & \xrightarrow{E_\alpha} & \mathbb{R}^{d_\alpha} \\ {\scriptstyle \iota_\alpha}\downarrow & & \downarrow{\scriptstyle q} \\ \mathcal{W}_{super} & \xrightarrow[E_{super}]{} & \mathcal{Z}_{super} \end{array}$$

     where $\iota_\alpha : \mathcal{W}_\alpha \to \mathcal{W}_{super}$ is the embedding from Theorem E.1, and q is the quotient map.
   - *Optimality: E_{super} minimizes the multimodal reconstruction loss:*

$$\min_{E_{super}} \sum_\alpha \mathbb{E}_{x_\alpha} \|D_\alpha(E_\alpha(x_\alpha)) - x_\alpha\| + \lambda \cdot \mathcal{L}_{align}$$

     where $\mathcal{L}_{align} = \sum_{\alpha \neq \beta} \|T_{\alpha\beta}(E_\alpha(x_\alpha)) - E_\beta(x_\beta)\|$.

6. *Cognitive Hierarchy Elevation: Through this construction, the system gains meta-cognitive ability: there exists a lifting map $\Phi_{meta} : \mathcal{Z}_{super} \to \mathcal{Z}_{super}$ such that*

$$\Phi_{meta} \circ E_{super} \approx E_{super} \circ \Phi_{cog}$$

   where Φ_{cog} is the cognitive dynamics.

Proof.

- Step 1 (Modal Encoding): By autoencoder theory, for a compact metric space $\mathcal{W}_\alpha$, there exist neural networks that approximate the identity map arbitrarily well. Taking $d_\alpha = O(\log |\mathcal{W}_\alpha|)$ yields ε-approximation.
- Step 2 (Translation Learning): Synchronous data $\mathcal{D}_{\text{sync}}$ provides supervision. By statistical learning theory, sample complexity $T = \Omega\left(\frac{d_\alpha d_\beta}{\varepsilon^2}\right)$ guarantees ε-accurate translation.
- Step 3 (Coordination Relation): Need to verify three properties:
  - Reflexivity: $T_{\alpha\alpha} \approx \mathrm{id}$ is guaranteed by the autoencoder objective.
  - Symmetry: If $T_{\alpha\beta}$ is optimal, then $T_{\beta\alpha} \approx T_{\alpha\beta}^{-1}$ follows from uniqueness of least-squares solutions.

– Transitivity: Guaranteed by triangle inequality and transitive consistency of synchronous data.

- Step 4 (Quotient Space Construction): Define a pseudo-metric

$$\tilde{d}(z_\alpha, z_\beta) = \inf_{\text{paths}} \sum_i \|T_{\alpha_i \alpha_{i+1}}(z_i) - z_{i+1}\|$$

The quotient metric $d_{\text{sup}}([z],[z']) = \inf_{z\sim z, z'\sim z'} \tilde{d}(z,z')$ makes $\mathcal{Z}_{\text{super}}$ a metric space.

- Step 5 (Supermodal Encoder): Construct E_{super} as:

$$E_{\text{super}}([z]) = \frac{1}{|A_0|} \sum_\alpha T_{\alpha\to\text{sup}}(z_\alpha)$$

where $T_{\alpha\to\text{sup}}$ maps from modality α to the common space. This construction minimizes the total loss, as shown by variational methods.

- Step 6 (Cognitive Elevation): Once $\mathcal{Z}_{\text{super}}$ is formed, the system can define a self-map Φ_{meta} as:

$$\Phi_{\text{meta}}(z) = \arg\min_{z'} \sum_\alpha \|T^{-1}_{\alpha\to\text{sup}}(z') - \Phi_\alpha(T^{-1}_{\alpha\to\text{sup}}(z))\|$$

where Φ_α is the cognitive dynamics for modality α. This provides meta-cognitive elevation.

□

Corollary E.1 (Cognitive Rationality of the Construction)**.** *The supermodal space $\mathcal{W}_{super}$ produced by the above construction satisfies:*

1. *Embodied Foundation: Every element can be traced back to specific modal experiences.*
2. *Gradual Abstraction: The level of abstraction increases with $|A_0|$.*
3. *Cognitive Economy: Representation complexity $d_{super} \le \sum_\alpha d_\alpha$, typically much smaller than the direct sum dimension.*
4. *Learning Stability: The sample complexities of construction algorithms $\mathcal{A}_{enc}, \mathcal{A}_{trans}$ are polynomially bounded.*

Definition E.6 (Modal Attention Sheaf)**.** *Define the **modal attention sheaf** $\mathscr{A}_{att}$ as a sheaf on A, whose stalk $\mathscr{A}_{att,\alpha}$ represents the attention weight for modality α, satisfying $\sum_{\alpha\in A} \mathscr{A}_{att,\alpha} = 1$. Specifically, for an open set $U \subseteq A$, $\mathscr{A}_{att}(U)$ is a probability distribution on U, and restriction maps are marginal distributions.*

Lemma E.3 (Convergence of Modal Attention Mechanism)**.** *Let $\{\alpha_t\}$ be a sequence of attention shifts, with transition probability $P(\alpha_{t+1} = \beta|\alpha_t = \alpha) = p_{\alpha\beta}$. If the transition matrix $P = (p_{\alpha\beta})$ satisfies:*

1. *Irreducible: For any α, β, there exists n such that $(P^n)_{\alpha\beta} > 0$*
2. *Aperiodic: There exists m such that $(P^m)_{\alpha\alpha} > 0$ for all α*
3. *Bounded: $\sup_{\alpha,\beta} |p_{\alpha\beta} - \pi_\beta| \le C\rho^t$, $\rho < 1$*

then there exists a stationary distribution π such that $\lim_{t\to\infty} \mathbb{P}(\alpha_t = \alpha) = \pi_\alpha$, with exponential convergence rate.

Proof. By Markov chain theory, a finite-state irreducible aperiodic chain has a unique stationary distribution and is geometrically ergodic. A coupling argument provides specifics. □

Theorem E.3 (Multimodal Asymmetry Enforces Cross-Modal Prediction Mechanism: Sheaf Version)**.** *Let G be the perception modality set, X the action modality set, and $G \neq X$. Then the machine learning mechanisms in $\mathcal{M}$ must include cross-modal prediction submodules, modal translation bridges, and gap-filling mechanisms.*

Proof. **Step 1: Problem Formalization**

Let G be the perception modality set, X the action modality set, $G \neq X$. Define:

- $G \setminus X$: modalities that can only be perceived but not acted upon
- $X \setminus G$: modalities that can only be acted upon but not perceived
- $G \cap X$: modalities that can both be perceived and acted upon

By Definition E.2, there exist perception homomorphisms $\pi_g : \mathscr{W}_g \to \mathcal{M}$ ($g \in G$) and action homomorphisms $\sigma_x : \mathcal{M} \to \mathscr{W}_x$ ($x \in X$).

Step 2: Necessity Analysis of Cross-Modal Prediction

Consider a task: given perception $s \in \mathscr{W}_g$ ($g \in G$), produce an action $t \in \mathscr{W}_x$ ($x \in X$). Three cases:

1. **Case 1**: $g \in G \cap X$, $x = g$. Direct perception-action loop can be used: $t = \sigma_g \circ \pi_g(s)$.

2. **Case 2**: $g \in G \cap X$, $x \in X$ but $x \neq g$. Cross-modal mapping needed: $t = \sigma_x \circ \pi_g(s)$. However, $\sigma_x \circ \pi_g$ is not defined, because $\pi_g(s)$ lies in $\mathcal{M}$, but σ_x expects a representation in $\mathcal{M}$. The system must learn a mapping from $\pi_g(s)$ to $\sigma_x^{-1}(\mathscr{W}_x)$.

3. **Case 3**: $g \in G \setminus X$, $x \in X \setminus G$. No overlap at all; must go through an intermediate representation: $t = \sigma_x \circ f \circ \pi_g(s)$, where $f : \mathcal{M} \to \mathcal{M}$ is an internally learned mapping.

Step 3: Construction of Cross-Modal Prediction Submodules

Define cross-modal prediction submodules $C_{\alpha\beta} : \mathscr{W}_\alpha \to \mathscr{W}_\beta$ ($\alpha \in G, \beta \in X$):

1. **Directly Computable Case**: If $\alpha = \beta \in G \cap X$, then $C_{\alpha\alpha} = \sigma_\alpha \circ \pi_\alpha$.

2. **Indirect Case**: If $\alpha \neq \beta$, construct as:

$$C_{\alpha\beta} = \sigma_\beta \circ \Phi_{\text{bridge}} \circ \pi_\alpha$$

where $\Phi_{\text{bridge}} : \mathcal{M} \to \mathcal{M}$ is a bridge mapping satisfying the variational problem:

$$\min_{\Phi} \sum_{\gamma \in G \cap X} d(\sigma_\gamma \circ \Phi \circ \pi_\gamma, \sigma_\gamma \circ \pi_\gamma) + \lambda \cdot \text{Reg}(\Phi)$$

Reg is a regularization term.

3. **Gap-Filling Mechanism**: For a modality α in $G \setminus X$, define the filling map $F_\alpha : \mathscr{W}_\alpha \to \prod_{\beta \in X} \mathscr{W}_\beta$ as:

$$F_\alpha(s) = (C_{\alpha\beta}(s))_{\beta \in X}$$

Similarly, for a modality β in $X \setminus G$, define the inverse filling $F_\beta^{-1} : \mathscr{W}_\beta \to \prod_{\alpha \in G} \mathscr{W}_\alpha$.

Step 4: Existence Proof of the Modal Translation Bridge

By the perception-action modal coordination of Axiom E.1, there exists a map $\theta : G \to X$ and an approximately commutative diagram:

$$\sigma^{\theta(g)} \circ \pi^g \xrightarrow{\zeta_g} \text{id}_{\mathscr{W}_g}$$

This means that for each $g \in G$, there exists at least one corresponding action modality $\theta(g) \in X$ such that the translation is approximately invertible.

Define the modal translation bridge $B_{g,x}$: for any $g \in G, x \in X$,

$$B_{g,x} = \sigma_x \circ \pi_g \quad \text{(if the domains match directly)}$$

Otherwise, via the chain rule:

$$B_{g,x} = \sigma_x \circ \Phi_{\text{inter}} \circ \pi_g$$

where Φ_{inter} is defined by the following learning problem:

$$\Phi_{\text{inter}} = \arg\min_{\Phi} \sum_{g \in G, x \in X} d(B_{g,x}^{\text{target}}, \sigma_x \circ \Phi \circ \pi_g)$$

The target values $B_{g,x}^{\text{target}}$ come from supervised data or self-supervised learning.

Step 5: Learning the Gap-Filling Mechanism

For missing modal pairs (g, x) ($g \in G \setminus X$ or $x \in X \setminus G$), employ generative filling:

1. **Forward Filling**: For $g \in G \setminus X$, learn a map $F_g : \mathscr{W}_g \to \mathcal{M}_{\text{latent}}$, where $\mathcal{M}_{\text{latent}}$ is a latent space, then decode to each action modality.

2. **Inverse Filling**: For $x \in X \setminus G$, learn a map $F_x^{-1} : \mathscr{W}_x \to \mathcal{M}_{\text{latent}}$, then encode to each perception modality.

The training objective is the reconstruction loss:

$$\mathcal{L}_{\text{fill}} = \sum_{g \in G \cap X} d(\sigma_g(F_g(s_g)), s_g) + \sum_{x \in G \cap X} d(\pi_x(F_x^{-1}(t_x)), t_x)$$

Step 6: Convergence Analysis

Define the total error:

$$E_n = \sup_{\substack{g \in G \\ x \in X}} d(C_{\alpha\beta}^{(n)}(s), C_{\alpha\beta}^{*}(s))$$

where the superscript (n) denotes the n-th iteration, and $*$ denotes the optimal value.

By the convergence of gradient descent, with an appropriate learning rate:

$$E_{n+1} \leq (1 - \eta\lambda)E_n + O(\eta^2)$$

where λ is the strong convexity coefficient of the loss function. Hence, $E_n \to 0$ linearly.

Step 7: Proof of Cognitive Necessity

Assume the system lacks a cross-modal prediction mechanism. Then for $g \in G \setminus X$ and $x \in X \setminus G$, the system cannot generate action t_x from perception s_g, violating the completeness requirement of an AGI system (the cognitive density of Definition A.19). Therefore, a cross-modal prediction mechanism is necessary. □

Definition E.7 (Modal Coordination Algebra Sheaf)**.** *Let $\mathcal{M} \in \mathbf{C}_{AGI}$. Its modal coordination algebra sheaf $\mathscr{A}_{coord}$ is defined as $\bigotimes_{\alpha \in A_0} \mathrm{End}(\mathscr{B}_4)$, where A_0 is a finite set of modalities, and $\mathscr{B}_4$ is the sheaf corresponding to the four-element Boolean algebra.*

Definition E.8 (Super-Modal Prediction Sheaf)**.** *The super-modal prediction sheaf $\mathrm{Pred}_{super} : \mathscr{W}_{super} \to \mathscr{W}_{super}$ is an approximate sheaf morphism that predicts the state of the super-modal sheaf.*

Definition E.9 (Cross-Modal Translation Sheaf)**.** *The cross-modal translation sheaf $\{T_{\alpha\beta} : \mathscr{W}_\alpha \to \mathscr{W}_\beta\}_{\alpha,\beta \in A}$ is a family of approximate sheaf morphisms that implement translation between different modalities.*

Lemma E.4 (Learning Dynamics of Cross-Modal Translation)**.** *The learning dynamics of the cross-modal translation map $T_{\alpha\beta}$ is described by the differential equation:*

$$\frac{dT_{\alpha\beta}}{dt} = -\eta_t \nabla_{T_{\alpha\beta}} \mathcal{L}_{cross} + \xi_t$$

where $\mathcal{L}_{cross} = \mathbb{E}_{x \sim \mathscr{W}_\alpha}[d(T_{\alpha\beta}(x), \mathcal{R}_{\alpha\beta}(x))]$. Under appropriate conditions, $T_{\alpha\beta} \to \mathcal{R}_{\alpha\beta}$ almost surely.

Proof. Apply stochastic approximation theory, verifying the Robbins-Monro conditions: $\sum \eta_t = \infty$, $\sum \eta_t^2 < \infty$, and the gradient estimate is unbiased. □

Theorem E.4 (Sheaf-Theoretic Version of the Super-Modal Forced AGI System Prefabricated Structure Theorem)**.** *Let the super-modal sheaf $\mathscr{W}_{super}$ exist (guaranteed by Theorem E.1). Then, the AGI system $\mathcal{M} \in \mathbf{AGI}$ must internally prefabricate the following three-layer structure:*

1. *Modal coordination algebra sheaf: $\mathscr{A}_{coord} \subseteq \mathcal{M}$, isomorphic to $\bigotimes_{\alpha \in A_0} \mathrm{End}(\mathscr{B}_4)$;*

2. *Super-modal prediction sheaf:* $\mathrm{Pred}_{super} : \mathscr{W}_{super} \to \mathscr{W}_{super}$;

3. *Cross-modal translation sheaf:* $\{T_{\alpha\beta} : \mathscr{W}_\alpha \to \mathscr{W}_\beta\}_{\alpha,\beta\in A}$.

Simultaneously, the self-referential prediction module $\mathrm{MetaPred}$ *must decompose into four corresponding submodules.*

Proof. **Step 1: Necessity Proof** By Theorem E.1 (existence and uniqueness of the super-modal sheaf), the super-modal sheaf $\mathscr{W}_{\text{super}}$ exists and possesses a universal property. Since the AGI system $\mathcal{M}$ needs to interact with the super-modal sheaf, $\mathcal{M}$ must internally have structures to handle super-modal information.

Step 2: Construction of the Modal Coordination Algebra Sheaf Since $\mathcal{M}$ is a cognitively complete object (Definition A.19), it can represent information from any modality. To efficiently coordinate multimodal information, $\mathcal{M}$ requires a prefabricated algebraic structure to handle inter-modal relations. By Theorem C.7, $\mathcal{M} \cong \mathcal{B}_4^k$, and its endomorphism monoid has an orthogonal decomposition. Therefore, the modal coordination algebra sheaf is naturally isomorphic to $\bigotimes_{\alpha\in A_0} \mathrm{End}(\mathscr{B}_4)$.

Step 3: Necessity of the Super-Modal Prediction Sheaf The system needs to predict super-modal states to complete cognitive tasks. By the limiting cognitive isomorphism property of Axiom A.1, the system must be able to predict the evolution of super-modal states.

Step 4: Necessity of the Cross-Modal Translation Sheaf By Definition E.2, the perception modality set G and the action modality set X may not coincide; the system needs to translate information between different modalities. This necessitates the existence of a cross-modal translation sheaf.

Step 5: Decomposition of MetaPred Since multimodal cognition involves four distinct levels (basic modality, super-modality, cross-modality, and process scheduling), $\mathrm{MetaPred}$ must correspondingly decompose into four submodules to monitor and optimize predictions at these levels respectively. □

Corollary E.2 (Sheaf-Theoretic Version of the Connection Between Prefabricated Structure and MetaPred)**.** *There is a necessary connection between the modal coordination algebra* $\mathscr{A}_{coord}$ *and the self-referential prediction module* $\mathrm{MetaPred}$*:* $\mathrm{MetaPred}$ *is a central element of* $\mathscr{A}_{coord}$*, the fixed-point algebra of* $\mathrm{MetaPred}$ *is isomorphic to the trivial representation subalgebra of* $\mathscr{A}_{coord}$*, and the iterative application of* $\mathrm{MetaPred}$ *implements the gradual optimization of modal coordination.*

Proof. By Theorem E.4, $\mathrm{MetaPred}$ decomposes into four submodules, among which $\mathrm{MetaPred}_{\text{super}}$ corresponds to super-modal prediction and is closely related to the modal coordination algebra $\mathscr{A}_{\text{coord}}$. Since $\mathrm{MetaPred}$ is self-referential, it acts on the prediction mechanism, which is generated by $\mathscr{A}_{\text{coord}}$. Therefore, $\mathrm{MetaPred}$ is a central element of $\mathscr{A}_{\text{coord}}$. The fixed-point algebra corresponds to well-coordinated states. Iterative application can optimize coordination. Q.E.D. □

Corollary E.3 (Sheaf-Theoretic Version of the Specialized Submodule Structure of Multimodal MetaPred)**.** *Multimodal asymmetry forces* $\mathrm{MetaPred}$ *to evolve three specialized submodules: cross-modal consistency predictor, modal translation bridge predictor, and gap-filling predictor.*

Proof. By Theorem E.3, the system requires mechanisms such as cross-modal prediction; therefore, $\mathrm{MetaPred}$ needs corresponding submodules to monitor and optimize these mechanisms. Q.E.D. □

Theorem E.5 (Super-Modal Representation Theorem)**.** *For any modality* $\mathcal{W}_\gamma$ *(whether directly accessible or not), its representation inside the AGI system is given by the following approximately commutative diagram:*

$$\mathcal{W}_\gamma \xrightarrow[\approx]{\varphi_\gamma} \mathcal{W}_{super} \xrightarrow[\approx]{\psi} \trianglelefteq_{super} = \trianglelefteq_{sem}/\sim_{\mathcal{R}}, \qquad \mathcal{W}_\gamma \xrightarrow{\mathrm{Enc}_\gamma} \trianglelefteq_{super} = \trianglelefteq_{sem}/\sim_{\mathcal{R}}$$

where $\mathrm{Enc}_\gamma = \psi \circ \varphi_\gamma$ *is the encoding map for modality* γ*, and* ψ *is the isomorphism for the super-modality (Theorem 8.5). Even if* $\gamma \notin G$*, this encoding can still be defined indirectly via coordination relations with other modalities.*

Proof. By Theorem E.1, there exists a super-modal sheaf $\mathscr{W}_{\text{super}}$ and an embedding φ_γ. By Theorem E.10 (realization of super-modality in semantic space), there exists an approximate isomorphism $\psi : \mathscr{W}_{\text{super}} \to \nleqslant_{\text{super}}$. Therefore, $\text{Enc}_\gamma = \psi \circ \varphi_\gamma$.

For a modality $\gamma \notin G$ that is not directly accessible, it can be defined indirectly via coordination relations: choose $\alpha \in G$, and define $\text{Enc}_\gamma = \psi \circ \varphi_\alpha \circ \mathcal{R}_{\alpha\gamma}^{-1}$ (approximately). By the transitivity of coordination relations, definitions via different paths are consistent within error bounds. □

Theorem E.6 (Cross-Modal Prediction Existence Theorem)**.** *Let $\alpha \in G$ be a directly perceivable modality, and $\gamma \notin G \cup X$ be a modality not directly accessible. Then there exists a cross-modal prediction map $T_{\alpha\gamma} : \mathscr{W}_\alpha \to \mathscr{W}_\gamma$, satisfying:*

$$T_{\alpha\gamma} \approx \varphi_\gamma^{-1} \circ \text{Pred}_{\textit{super}} \circ \varphi_\alpha$$

and its error is monitored by $\text{MetaPred}_{\textit{cross}}$ and can be optimized via self-supervised learning.

Proof. Define $T_{\alpha\gamma} = \varphi_\gamma^{-1} \circ \text{Pred}_{\text{super}} \circ \varphi_\alpha$, where φ_γ^{-1} is the approximate inverse. Error analysis: let the error of $\text{Pred}_{\text{super}}$ be $\varepsilon_{\text{super}}$, and the embedding errors be $\varepsilon_\alpha, \varepsilon_\gamma$; then the total error is $O(\varepsilon_{\text{super}} + \varepsilon_\alpha + \varepsilon_\gamma)$.

$\text{MetaPred}_{\text{cross}}$ provides a self-supervised signal by comparing the prediction result with estimates via other paths. Specifically, define the loss function:

$$\mathcal{L}(T_{\alpha\gamma}) = \mathbb{E}_{x \sim \mathscr{W}_\alpha} \left[d\left(T_{\alpha\gamma}(x), \mathcal{R}_{\alpha\gamma}(x) \right) \right]$$

and optimize $T_{\alpha\gamma}$ via gradient descent. □

Theorem E.7 (New Modality Generation Theorem)**.** *In the semantic space $\nleqslant_{\textit{sem}}$, through iteration of the self-referential prediction module MetaPred, a new state $x_{\textit{new}} = \lim_{n\to\infty} \text{MetaPred}^n(x_0)$ can be generated, where x_0 is a combination of encodings from known modalities. This state corresponds under the quotient map to a new equivalence class $[x_{\textit{new}}] \in \nleqslant_{\textit{super}}$, representing a novel modal experience created by the system. The generation process satisfies superlinear convergence:*

$$d(\text{MetaPred}^{n+1}(x_0), x_{\textit{new}}) \leq \gamma_n d(\text{MetaPred}^n(x_0), x_{\textit{new}})^2, \quad \gamma_n \to \gamma^* < \infty.$$

Proof. In the special case of B_4, the iteration $\text{MetaPred}^n(x_0)$ converges superlinearly to x_{new}. Since x_0 is a combination of encodings from known modalities and MetaPred is self-referential, its fixed point may lie beyond linear combinations of known modalities.

Consider the quotient map $q : \nleqslant_{\text{sem}} \to \nleqslant_{\text{super}}$. If $[x_{\text{new}}] \in q(\bigcup_\alpha \text{Enc}_\alpha(\mathscr{W}_\alpha))$, then it corresponds to a fusion of known modalities; otherwise, it represents an entirely new modal experience. Due to the creativity of MetaPred (via the fixed-point combinator), the latter case may occur.

The convergence rate follows directly from Theorem on the special case of B_4. □

E.3 Semantic Encoding and Convergence under Multimodality

Theorem E.8 (Sheaf-Theoretic Version of Uniqueness of Semantic Encoding under Multimodality)**.** *Assume $\mathcal{X}$ satisfies weak separation, and the modal coordination relation $\mathcal{R}$ is dense. Then an encoding map $f : \mathcal{X} \to \mathscr{P}$ that is information-lossless, coordination-preserving, and operation-preserving is unique up to approximate isomorphism.*

Proof. **Step 1: Problem Setup and Assumptions**

Let $\mathcal{X}$ be the input space with weak separation: for any $x \neq y \in \mathcal{X}$, there exists a continuous function $f : \mathcal{X} \to \mathbb{R}$ such that $f(x) \neq f(y)$.

Let $\mathscr{P}$ be the parameter sheaf (Definition D.4), and $\mathcal{R}$ be a dense modal coordination relation: for any $\alpha, \beta \in A$, there exists a finite sequence $\alpha = \alpha_0, \alpha_1, \ldots, \alpha_n = \beta$ such that the error ε_i of $\mathcal{R}_{\alpha_i\alpha_{i+1}}$ satisfies $\sum_i \varepsilon_i < \infty$.

Step 2: Core Idea of Uniqueness Proof

Assume there exist two encodings $f_1, f_2 : \mathcal{X} \to \mathscr{P}$ satisfying: 1. **Information-lossless**: There exist decodings $g_i : \mathscr{P} \to \mathcal{X}$ such that $g_i \circ f_i \approx \text{id}_\mathcal{X}$ 2. **Coordination-preserving**: For any modalities

α, β, $\mathcal{R}_{\alpha\beta} \circ f_{i,\alpha} \approx f_{i,\beta}$, where $f_{i,\alpha}$ is the component of f_i on modality α 3. **Operation-preserving**: For any operation ω, $f_i(\omega(x)) \approx \omega(f_i(x))$

To prove: $f_1 \approx f_2$ (up to approximate isomorphism).

Step 3: Per-Modality Approximation

Fix a modality α. Define the error function:

$$E_\alpha(x) = d_{\mathscr{P}}(f_{1,\alpha}(x), f_{2,\alpha}(x))$$

where $d_{\mathscr{P}}$ is the pseudo-metric on $\mathscr{P}$.

By coordination preservation, for any β:

$$\begin{aligned} d(f_{1,\beta}(x), f_{2,\beta}(x)) &\le d(f_{1,\beta}(x), \mathcal{R}_{\alpha\beta}(f_{1,\alpha}(x))) \\ &\quad + d(\mathcal{R}_{\alpha\beta}(f_{1,\alpha}(x)), \mathcal{R}_{\alpha\beta}(f_{2,\alpha}(x))) \\ &\quad + d(\mathcal{R}_{\alpha\beta}(f_{2,\alpha}(x)), f_{2,\beta}(x)) \\ &\le \varepsilon_{\alpha\beta} + L_{\mathcal{R}} \cdot E_\alpha(x) + \varepsilon_{\alpha\beta} \end{aligned}$$

where $L_{\mathcal{R}}$ is the Lipschitz constant of $\mathcal{R}_{\alpha\beta}$.

Step 4: Utilizing Density

By the density of $\mathcal{R}$ (condition 4 of Definition E.1), for any $\alpha, \beta \in A$, there exists a finite path $\alpha = \alpha_0, \alpha_1, \ldots, \alpha_n = \beta$ such that:

$$\sum_{i=0}^{n-1} \varepsilon_{\alpha_i \alpha_{i+1}} < \delta$$

where $\delta > 0$ is arbitrarily small (due to convergence, n can be taken sufficiently large). Consider error propagation along the path:

$$d(f_{1,\beta}(x), f_{2,\beta}(x)) \le d(f_{1,\beta}(x), \mathcal{R}_{\alpha_{n-1}\beta}(f_{1,\alpha_{n-1}}(x))) + \cdots + d(\mathcal{R}_{\alpha\alpha_1}(f_{1,\alpha}(x)), f_{2,\alpha_1}(x))$$

Each term can be controlled:

$$\begin{aligned} d(\mathcal{R}_{\alpha_i\alpha_{i+1}}(f_{1,\alpha_i}(x)), f_{2,\alpha_{i+1}}(x)) &\le d(\mathcal{R}_{\alpha_i\alpha_{i+1}}(f_{1,\alpha_i}(x)), \mathcal{R}_{\alpha_i\alpha_{i+1}}(f_{2,\alpha_i}(x))) \\ &\quad + d(\mathcal{R}_{\alpha_i\alpha_{i+1}}(f_{2,\alpha_i}(x)), f_{2,\alpha_{i+1}}(x)). \end{aligned}$$

By the coordination-preserving condition, the second term $\le \varepsilon_{\alpha_i\alpha_{i+1}}$; by the Lipschitz property of $\mathcal{R}$ (Definition A.1), the first term $\le L_{\mathcal{R}} \cdot d(f_{1,\alpha_i}(x), f_{2,\alpha_i}(x))$. Thus, by recursion:

$$E_\beta(x) \le L_{\mathcal{R}}^n E_\alpha(x) + (L_{\mathcal{R}}^{n-1} + \cdots + 1) \cdot \delta$$

Since $L_{\mathcal{R}} \approx 1$ (approximate homomorphism) and the path length n is finite, $L_{\mathcal{R}}^n \approx 1$. Letting $\delta \to 0$, we obtain $E_\beta(x) \lesssim E_\alpha(x)$. Similarly, $E_\alpha(x) \lesssim E_\beta(x)$, hence

$$\sup_\beta E_\beta(x) \le E_\alpha(x) + o(1)$$

i.e., the global error is controlled by the error in any single modality.

Step 5: Global Error Estimate

Define the global error:

$$E(x) = \sup_{\alpha \in A} E_\alpha(x)$$

By Step 4, $E(x)$ is independent (up to an arbitrarily small difference) of the choice of α.

By information-losslessness, there exist decodings g_1, g_2 such that:

$$d(g_1(f_1(x)), x) \le \varepsilon_1, \quad d(g_2(f_2(x)), x) \le \varepsilon_2$$

Consider the composite map $g_1 \circ f_2$:

$$d(g_1(f_2(x)), x) \le d(g_1(f_2(x)), g_1(f_1(x))) + d(g_1(f_1(x)), x) \le L_{g_1} \cdot E(x) + \varepsilon_1$$

where L_{g_1} is the Lipschitz constant of g_1.

Similarly:

$$d(g_2(f_1(x)), x) \leq L_{g_2} \cdot E(x) + \varepsilon_2$$

Step 6: Utilizing Weak Separation

Let $x \neq y$. By weak separation, there exists a continuous f such that $f(x) \neq f(y)$. Consider:

$$|f(g_1(f_2(x))) - f(x)| \leq \omega_f(d(g_1(f_2(x)), x)) \leq \omega_f(L_{g_1} E(x) + \varepsilon_1)$$

where ω_f is the modulus of continuity of f.

If $E(x)$ does not tend to 0, then there exists a sequence $\{x_n\}$ such that $E(x_n) \geq \eta > 0$. But by information-losslessness, $g_1 \circ f_2$ should approximate the identity, a contradiction. Hence $E(x) \to 0$.

Step 7: Uniqueness up to Approximate Isomorphism

We have proven pointwise convergence $f_1 \approx f_2$. We need to show there exists an approximate isomorphism $\varphi : \mathscr{P} \to \mathscr{P}$ such that $f_2 = \varphi \circ f_1$.

Define φ as:

$$\varphi(p) = \lim_{n \to \infty} f_2(g_1^{(n)}(p))$$

where $g_1^{(n)}$ is the n-th iterative approximation of the inverse of g_1.

Verification: 1. **Well-defined**: The limit exists due to the approximate invertibility of f_1, f_2. 2. **Approximate homomorphism**: φ preserves the operation structure, with error controlled by iteration. 3. **Commutativity**: $\varphi \circ f_1 \approx f_2$.

Similarly, there exists ψ such that $\psi \circ f_2 \approx f_1$, and $\varphi \circ \psi \approx \mathrm{id}$, hence it is an approximate isomorphism. □

Theorem E.9 (Convergence of Multimodal Cognition)**.** *Let $\{\mathscr{H}_n\}$ be a sequence of multimodal cognitive hierarchy sheaves. Define the total error $\varepsilon_n^{multi} = \sup_{\alpha \in A_n} \varepsilon_n^{\alpha} + \sup_{\alpha,\beta \in A_n} \eta_n^{\alpha\beta}$, where ε_n^{α} is the approximation error for modality α, and $\eta_n^{\alpha\beta}$ is the coordination error. If the following hold:*

1. *Single-modality error convergence: $\sum_{n=0}^{\infty} \varepsilon_n^{\alpha} < \infty$*
2. *Coordination error convergence: $\sum_{n=0}^{\infty} \eta_n^{\alpha\beta} < \infty$*
3. *Error control: There exists $\psi(n) \to 0$ such that $\max(\varepsilon_n^{\alpha}, \eta_n^{\alpha\beta}) \leq \psi(n)$*

then $\sum_{n=0}^{\infty} \varepsilon_n^{multi} < \infty$, and the convergence rate is $\varepsilon_n^{multi} = O(1/n^{2+\gamma})$, $\gamma > 0$.

Proof. **Part 1: Notation and Assumptions**

Let $\{\mathscr{H}_n\}$ be a sequence of multimodal cognitive hierarchy sheaves. Define: - ε_n^{α}: approximation error for modality α at layer n - $\eta_n^{\alpha\beta}$: modal coordination error, $d(\mathcal{R}_{\alpha\beta}^{(n)}(x), y) \leq \eta_n^{\alpha\beta}$ - A_n: set of modalities involved at layer n, $|A_n| < \infty$ - Total error: $\varepsilon_n^{\text{multi}} = \sup_{\alpha \in A_n} \varepsilon_n^{\alpha} + \sup_{\alpha,\beta \in A_n} \eta_n^{\alpha\beta}$

By Definition A.16, single-modality errors satisfy:

$$\sum_{n=0}^{\infty} \varepsilon_n^{\alpha} < \infty, \quad \varepsilon_n^{\alpha} \leq \varphi^{\alpha}(n),\ \varphi^{\alpha}(n) \to 0$$

Part 2: Recurrence Relation for Coordination Errors

Coordination error $\eta_n^{\alpha\beta}$ consists of two parts: 1. Propagation of single-modality errors: $\eta_n^{\alpha\beta} \leq L_{\mathcal{R}}(\varepsilon_n^{\alpha} + \varepsilon_n^{\beta})$ 2. Approximation error of the coordination relation itself: $\delta_n^{\alpha\beta}$

Thus:

$$\eta_n^{\alpha\beta} \leq L_{\mathcal{R}}(\varepsilon_n^{\alpha} + \varepsilon_n^{\beta}) + \delta_n^{\alpha\beta}$$

By Axiom E.1, $\delta_n^{\alpha\beta}$ satisfies:

$$\sum_{n=0}^{\infty} \delta_n^{\alpha\beta} < \infty, \quad \delta_n^{\alpha\beta} \leq \psi(n),\ \psi(n) \to 0$$

Part 3: Upper Bound for Total Error

$$\varepsilon_n^{\text{multi}} \leq \sup_{\alpha} \varepsilon_n^{\alpha} + \sup_{\alpha,\beta}[L_{\mathcal{R}}(\varepsilon_n^{\alpha} + \varepsilon_n^{\beta}) + \delta_n^{\alpha\beta}] \leq (1 + 2L_{\mathcal{R}}) \sup_{\alpha} \varepsilon_n^{\alpha} + \sup_{\alpha,\beta} \delta_n^{\alpha\beta}$$

Define:

$$a_n = \sup_{\alpha} \varepsilon_n^{\alpha}, \quad b_n = \sup_{\alpha,\beta} \delta_n^{\alpha\beta}$$

Then:

$$\varepsilon_n^{\text{multi}} \leq (1 + 2L_{\mathcal{R}})a_n + b_n$$

Part 4: Proof of Series Convergence

We need to prove: $\sum_{n=0}^{\infty} \varepsilon_n^{\text{multi}} < \infty$.

By single-modality error convergence: $\sum a_n < \infty$. By coordination error convergence: $\sum b_n < \infty$.

Hence:

$$\sum_{n=0}^{\infty} \varepsilon_n^{\text{multi}} \leq (1 + 2L_{\mathcal{R}}) \sum_{n=0}^{\infty} a_n + \sum_{n=0}^{\infty} b_n < \infty$$

Part 5: Convergence Rate Estimation

Assume single-modality errors decay polynomially: $a_n = O(1/n^{2+\gamma_\alpha})$. Let $\gamma = \min_\alpha \gamma_\alpha > 0$, then $a_n = O(1/n^{2+\gamma})$.

Coordination errors b_n typically decay faster; assume $b_n = O(1/n^{3+\gamma})$ (since coordination requires higher accuracy).

Then:

$$\varepsilon_n^{\text{multi}} = O(1/n^{2+\gamma}) + O(1/n^{3+\gamma}) = O(1/n^{2+\gamma})$$

Part 6: Almost Sure Convergence Proof

Consider the case of random errors (in practical systems, errors have randomness). Let $\{\varepsilon_n^{\text{multi}}\}$ be a random sequence satisfying:

$$\mathbb{E}[\varepsilon_n^{\text{multi}}] = O(1/n^{2+\gamma}), \quad \text{Var}(\varepsilon_n^{\text{multi}}) = O(1/n^{4+2\gamma})$$

By Chebyshev's inequality:

$$\mathbb{P}(|\varepsilon_n^{\text{multi}} - \mathbb{E}[\varepsilon_n^{\text{multi}}]| > \varepsilon) \leq \frac{\text{Var}(\varepsilon_n^{\text{multi}})}{\varepsilon^2} = O(1/n^{4+2\gamma})$$

Since $\sum_{n=1}^{\infty} 1/n^{4+2\gamma} < \infty$, by the Borel-Cantelli lemma, almost surely,

$$\varepsilon_n^{\text{multi}} - \mathbb{E}[\varepsilon_n^{\text{multi}}] \to 0$$

Hence $\varepsilon_n^{\text{multi}} \to 0$ a.s.

Part 7: Existence of the Limit of Cognitive Hierarchies

Define the limit of the cognitive hierarchy sequence:

$$\mathscr{H}_\infty = \varinjlim_n \mathscr{H}_n$$

In the completion category $\widehat{\mathbf{AGI}}$, since $\{\varepsilon_n^{\text{multi}}\}$ converges in series, $\{\mathscr{H}_n\}$ is a Cauchy sequence, hence the limit exists.

Construction: for any open set U, define

$$\mathscr{H}_\infty(U) = \left\{ (s_n) \in \prod_n \mathscr{H}_n(U) \;\middle|\; \lim_{n\to\infty} d(\mu_n(s_n), s_{n+1}) = 0 \right\} / \sim$$

where $\mu_n : \mathscr{H}_n \to \mathscr{H}_{n+1}$ are connecting homomorphisms, and $\sim$ is the equivalence relation: $(s_n) \sim (t_n)$ if and only if $\lim_{n\to\infty} d(s_n, t_n) = 0$. □

Proposition E.1 (Inevitable Form of Multimodal Machine Learning Loss Function)**.** *In a multimodal AGI system, to learn coordinated multimodal representations, its loss function typically needs to contain the following components:*

1. ***Intra-modality reconstruction loss**: $\mathcal{L}_{intra} = \sum_{\alpha \in G} \mathbb{E}_{x \sim \mathcal{W}_\alpha}[d(\mathrm{Dec}_\alpha(\mathrm{Enc}_\alpha(x)), x)]$, ensuring encoding quality for each modality;*
2. ***Cross-modality coordination loss**, enforcing preservation of coordination relations:*

$$\mathcal{L}_{cross} = \sum_{\alpha,\beta \in A} \mathbb{E}_{x \sim \mathcal{W}_\alpha} \big[d(\mathcal{R}_{\alpha\beta}(\mathrm{Enc}_\alpha(x)), \mathrm{Enc}_\beta(\mathcal{R}_{\alpha\beta}(x))) \big].$$

3. ***Super-modality consistency loss**: $\mathcal{L}_{super} = \mathbb{E}_{x \sim \mathcal{W}_{super}}[d(\mathrm{Pred}_{super}(\mathrm{Enc}_{super}(x)), \mathrm{Enc}_{super}(x))]$, optimizing super-modal prediction.*

where $\mathrm{Enc}_\alpha, \mathrm{Dec}_\alpha$ are the encoder-decoder for modality α, and Enc_{super} is the super-modality encoder.

Proof. These three loss terms correspond to three requirements of multimodal cognition: intra-modality accuracy, inter-modality consistency, and super-modal predictability. By Axiom E.1, the system must maintain modal coordination relations, so the loss function must contain terms encouraging coordination (like $\mathcal{L}_{\text{cross}}$). Intra-modality reconstruction loss ensures basic representation quality, and super-modality consistency loss ensures overall predictive capability. Omitting any component will lead to cognitive defects. □

E.4 Realization of Super-Modality in the Semantic-Adjoint Dynamics Architecture

Theorem E.10 (Sheaf-Theoretic Version of the Realization Theorem of Super-Modality in Semantic Space)**.** *Let $\mathcal{M} \in \mathbf{C}_{AGI}$ satisfy the semantic-process decomposition, with state sheaf $\mathscr{F}_{\nleq} \cong \mathscr{F}_{sem} \otimes \mathscr{F}_{proc}$, where $\mathscr{F}_{sem} \cong \mathscr{B}_4^{k_{\min}}$. Then the realization of the super-modality $\mathscr{W}_{super}$ in $\mathcal{M}$ is a quotient subsheaf of the semantic sheaf $\mathscr{F}_{sem}$: $\mathscr{F}_{super} = \mathscr{F}_{sem}/\sim_{\mathcal{R}}$.*

Proof. **Step 1: Review of Semantic-Process Decomposition**

By Theorem C.6, $\mathcal{M} \cong \mathcal{B}_4^{k_{\min}}$, and the state sheaf decomposes as:

$$\mathscr{F}_{\nleq} \cong \mathscr{F}_{\text{sem}} \otimes \mathscr{F}_{\text{proc}}$$

where $\mathscr{F}_{\text{sem}} \cong \mathscr{B}_4^{k_{\min}}$, and $\mathscr{B}_4$ is the sheaf corresponding to the four-element Boolean algebra.

Step 2: Semantic Sheaf as Multimodal Representation Space

For each modality α, there exists an encoding map:

$$\mathrm{Enc}_\alpha : \mathscr{W}_\alpha \to \mathscr{F}_{\text{sem}}$$

By operational generation completeness, these encodings can be simultaneously embedded into $\mathscr{F}_{\text{sem}}$.

Define an equivalence relation $\sim_{\mathcal{R}}$ on the semantic sheaf: for $x, y \in \mathscr{F}_{\text{sem}}$, $x \sim_{\mathcal{R}} y$ if and only if there exists a modality sequence $\alpha_0, \ldots, \alpha_n$ and elements $x_i \in \mathscr{W}_{\alpha_i}$ such that: 1. $\mathrm{Enc}_{\alpha_0}(x_0) \approx x$ 2. $\mathrm{Enc}_{\alpha_n}(x_n) \approx y$ 3. $\mathcal{R}_{\alpha_i \alpha_{i+1}}(x_i) \approx x_{i+1}$

Step 3: Construction of the Quotient Sheaf

Define the quotient sheaf:

$$\mathscr{F}_{\text{super}} = \mathscr{F}_{\text{sem}}/\sim_{\mathcal{R}}$$

Explicit construction: for an open set U,

$$\mathscr{F}_{\text{super}}(U) = \{[s] \mid s \in \mathscr{F}_{\text{sem}}(U)\}$$

where $[s]$ denotes the equivalence class of s. Sheaf operations are defined as:

$$\omega([s_1], \ldots, [s_k]) = [\omega(s_1, \ldots, s_k)]$$

Well-definedness needs verification: if $s_i \sim_{\mathcal{R}} t_i$, then there exist coordination paths. By the operation-preserving property of encoding and coordination relations, $\omega(s_1, \ldots, s_k) \sim_{\mathcal{R}} \omega(t_1, \ldots, t_k)$.

Step 4: Isomorphism Proof

Define a map $\Phi : \mathscr{F}_{\text{super}} \to \mathscr{W}_{\text{super}}$:

$$\Phi([\text{Enc}_\alpha(x)]) = \varphi_\alpha(x)$$

where $\varphi_\alpha : \mathscr{W}_\alpha \to \mathscr{W}_{\text{super}}$ is the natural embedding into the super-modality.

We need to prove: 1. **Well-definedness**: If $\text{Enc}_\alpha(x) \sim_{\mathcal{R}} \text{Enc}_\beta(y)$, then there exists a coordination path making x and y coordinated, hence $\varphi_\alpha(x) \approx \varphi_\beta(y)$. 2. **Surjectivity**: For any $w \in \mathscr{W}_{\text{super}}$, by the construction of super-modality, there exists a representation $(x_\alpha) \in \prod \mathscr{W}_\alpha$ such that $w = \varphi_\alpha(x_\alpha)$. Take $s = \text{Enc}_\alpha(x_\alpha)$, then $\Phi([s]) = w$. 3. **Injectivity**: If $\Phi([s]) = \Phi([t])$, then $\varphi_\alpha(x) \approx \varphi_\beta(y)$; by the universal property of super-modality, there exists a coordination path, hence $s \sim_{\mathcal{R}} t$. 4. **Sheaf homomorphism**: Φ preserves sheaf operations, with error controlled by the approximation of encoding and coordination.

Step 5: Error Analysis

Let encoding error be ε_{enc}, coordination error $\varepsilon_{\mathcal{R}}$. Then the error of Φ:

$$\varepsilon_\Phi \leq \varepsilon_{\text{enc}} + \varepsilon_{\mathcal{R}} + \varepsilon_{\text{quotient}}$$

where $\varepsilon_{\text{quotient}}$ is the error introduced by the quotient construction, satisfying:

$$\varepsilon_{\text{quotient}} \leq \frac{1}{|A_0|} \sum_{\alpha,\beta} \varepsilon_{\mathcal{R}_{\alpha\beta}}$$

By convergence of coordination errors, $\varepsilon_\Phi \to 0$.

Step 6: Uniqueness Proof

Assume there exists another realization $\Psi : \mathscr{F}'_{\text{super}} \to \mathscr{W}_{\text{super}}$. Define a map:

$$\theta : \mathscr{F}_{\text{super}} \to \mathscr{F}'_{\text{super}}, \quad \theta = \Psi^{-1} \circ \Phi$$

Since Φ, Ψ are approximate isomorphisms, θ is an approximate isomorphism. We need to prove θ is compatible with the semantic-process decomposition.

Consider the commutative diagram:

$$\begin{array}{ccc} \mathscr{F}_{\text{sem}} & \xrightarrow{q} & \mathscr{F}_{\text{super}} \\ \downarrow{\scriptstyle \iota} & & \downarrow{\scriptstyle \theta} \\ \mathscr{F}'_{\text{sem}} & \xrightarrow[q']{} & \mathscr{F}'_{\text{super}} \end{array}$$

where ι is the natural isomorphism of semantic sheaves (by uniqueness of $\mathcal{B}_4^{k_{\min}}$). By the universal property, θ is unique. □

Theorem E.11 (Sheaf-Theoretic Version of Semantic Encoding as the Hub for Super-Modal Realization)**.** *Let* $\text{Enc}_{sem} : \mathcal{X} \to \mathscr{F}_{sem}$ *be the semantic encoding sheaf. Then there exist a super-modal encoding* $\text{Enc}_{super} = q \circ \text{Enc}_{sem}$*, modality-specific encodings* Enc_α*, coordination encoding conditions, and optimality of encoding.*

Proof. Define the super-modal encoding as the semantic encoding followed by the quotient map $q : \mathscr{F}_{\text{sem}} \to \mathscr{F}_{\text{super}}$. Modality-specific encodings are restrictions of the semantic encoding to modality α. The coordination encoding condition is ensured by Enc_{sem} preserving coordination relations. Optimality is guaranteed by the optimality of semantic encoding (Theorem E.8). Q.E.D. □

Theorem E.12 (Sheaf-Theoretic Version of Discrete Lie-Group-Like Structure)**.** *The process state sheaf $\mathscr{F}_{proc} = \{1, \ldots, k\} \times \{0, 1\}^k$ has a discrete Lie-group-like structure: group structure, group action, discrete dynamics.*

Proof. $\mathscr{F}_{\text{proc}}$ can be viewed as the discrete group $G = \mathbb{Z}_2^k \rtimes S_k$, where $\mathbb{Z}_2^k$ represents bit flips and S_k represents position permutations. This group acts on the semantic sheaf, realizing attention shifts and mode switches. Dynamics are defined by random walks or deterministic update rules on the group. Q.E.D. □

Theorem E.13 (Sheaf-Theoretic Version of the Three-Layer Organization Theorem for Super-Modal Prediction Dynamics)**.** *Based on the semantic-adjoint dynamics decomposition, the prediction dynamics related to super-modality are organized into the following three layers: bottom-layer intra-modality prediction dynamics, middle-layer super-modal coordination dynamics, and top-layer cross-modal translation dynamics.*

Proof. The bottom layer handles prediction within each modality, the middle layer coordinates predictions across modalities, and the top layer handles cross-modal translation. This is naturally determined by the hierarchical structure of multimodal information. Q.E.D. □

Theorem E.14 (Complete Data Flow of Multimodal Cognition)**.** *In a multimodal AGI system, the complete data flow from perception to action is as follows:*

1. ***Perception Phase**:*

   $$\mathcal{W}_\alpha \xrightarrow{\text{Enc}_\alpha} \oint_\alpha \hookrightarrow \oint_{sem} \xrightarrow{q} \oint_{super}$$

   Information from each modality is independently encoded, then integrated into the semantic space, and finally projected to the super-modal space.

2. ***Prediction Phase**:*

   $$\oint_{super} \xrightarrow{\text{Pred}_{super}} \oint_{super} \quad \textit{(monitored by } \text{MetaPred}_{super}\textit{)}$$

   Simultaneously, each modality performs independent prediction:

   $$\oint_\alpha \xrightarrow{\text{Pred}_\alpha} \oint_\alpha \quad \textit{(monitored by } \text{MetaPred}_\alpha\textit{)}$$

3. ***Coordination Phase**:*

   - MetaPred_{cross} *checks cross-modal consistency*
   - $\mathcal{A}_{coord}$ *adjusts modal fusion weights*
   - *If inconsistent, triggers cross-modal learning updates* $T_{\alpha\beta}$

4. ***Decision Phase**:*

   $$\oint_{super} \xrightarrow{q^{-1}} \oint_{sem} \xrightarrow{\pi_\beta} \oint_\beta \xrightarrow{\text{Dec}_\beta} \mathcal{W}_\beta$$

   Super-modal decisions are decomposed into target action modalities.

5. ***Meta-Monitoring Throughout**:*

   $$\text{MetaPred} = \text{MetaPred}_{base} \oplus \text{MetaPred}_{super} \oplus \text{MetaPred}_{cross} \oplus \text{MetaPred}_{proc}$$

   Monitors quality at all stages and adjusts parameters.

The entire process is implemented by the adjoint dynamics $\Phi = (\Phi_{sem}, \Phi_{proc})$ for temporal evolution, where the process state $\sigma_t = (c_t, \mathbf{v}_t)$ controls attention allocation.

Proof. **Step 1: Formal Definition of Data Flow**

Define the multimodal cognitive process as a time series $\{S_t\}_{t \geq 0}$, where $S_t = (\sigma_t, \omega_t, \alpha_t)$ contains: - $\sigma_t \in \mathscr{F}_{\text{proc}}$: process state - $\omega_t \in \mathscr{F}_{\text{sem}}$: semantic state - $\alpha_t \in \bar{A}$: current focal modality

Step 2: Mathematical Description of the Five Phases

1. **Perception Phase**: For input $x \in \mathscr{W}_{\alpha_t}$,

$$\omega_t^{\text{new}} = \text{Enc}_{\alpha_t}(x) \oplus \omega_t$$

where $\oplus$ is the semantic fusion operator, defined as coordinate-weighted average: $\omega \oplus \omega' = \lambda\omega + (1-\lambda)\omega'$, with weight λ determined by the attention vector in σ_t.

2. **Prediction Phase**:

$$\omega_t^{\text{pred}} = \text{Pred}_{\text{super}}(q(\omega_t^{\text{new}}))$$
$$\omega_{\alpha,t}^{\text{pred}} = \text{Pred}_{\alpha}(\omega_t^{\text{new}}|_{\alpha})$$

3. **Coordination Phase**:

$$Q_t = \text{MetaPred}_{\text{cross}}(\{T_{\alpha\beta}\}), \quad \text{if } Q_t < \theta_{\text{cross}}, \text{ update } T_{\alpha\beta}$$

Update rule: $T_{\alpha\beta}^{(n+1)} = T_{\alpha\beta}^{(n)} - \eta\nabla\mathcal{L}_{\text{cross}}$, where $\mathcal{L}_{\text{cross}} = d(T_{\alpha\beta}(x), \mathcal{R}_{\alpha\beta}(x))$.

4. **Decision Phase**: Choose an action modality β,

$$y = \text{Dec}_{\beta}(\pi_{\beta}(q^{-1}(\omega_t^{\text{pred}})))$$

where π_β is the projection operator onto the semantic subspace corresponding to modality β.

5. **Meta-Monitoring**:

$$\sigma_{t+1} = \text{MetaPred}_{\text{proc}}(\sigma_t, \{\varepsilon_\alpha\})$$

where $\varepsilon_\alpha = d(\omega_{\alpha,t}^{\text{pred}}, \omega_t^{\text{new}}|_{\alpha})$ is the prediction error for modality α.

Step 3: Convergence Proof

Define a Lyapunov function:

$$V_t = d(\omega_t, \omega^*)^2 + \lambda \cdot \text{Reg}(\{T_{\alpha\beta}\}) + \mu \cdot d(\sigma_t, \sigma^*)$$

where ω^*, σ^* are ideal states. Compute expectation:

$$\mathbb{E}[V_{t+1}|\mathcal{F}_t] \leq V_t - \eta_t\|\nabla V_t\|^2 + \xi_t$$

By Theorem E.9, the error series converges $\sum \xi_t < \infty$, and the gradient has a lower bound (strong convexity), hence there exists $\rho < 1$ such that:

$$\mathbb{E}[V_{t+1}] \leq \rho\mathbb{E}[V_t] + \delta_t, \quad \sum \delta_t < \infty$$

By stochastic approximation theory, $V_t \to 0$ almost surely.

Step 4: Optimality Proof

This data flow is minimally complete. Proof by contradiction: if any phase is removed, the cognitive loop is broken.

- No perception phase: system has no input, cannot learn.
- No prediction phase: cannot perform future planning, violates AGI proactivity.
- No coordination phase: modal conflicts cannot be resolved, violates Axiom E.1.
- No decision phase: system cannot act, violates bidirectional interactivity.
- No meta-monitoring: cannot adaptively adjust, violates self-referential closure.

Step 5: Implementation Details Supplement

Provide specific algorithmic implementations:

- Semantic fusion operator $\oplus$: $\omega \oplus \omega' = \text{AttnPool}(\omega, \omega'; \mathbf{v}_t)$, where $\mathbf{v}_t$ is the attention vector in σ_t, and AttnPool is attention pooling.

- Attention shifting mechanism: $\alpha_{t+1} = \arg\max_\alpha \text{UCB}(\alpha, \varepsilon_\alpha, t)$, where UCB (Upper Confidence Bound) balances exploration and exploitation.
- Error propagation control: use a gating mechanism; when $\varepsilon_\alpha > \theta_{\text{high}}$, trigger reset; when $\varepsilon_\alpha \in (\theta_{\text{low}}, \theta_{\text{high}})$, perform local refinement.

□

Corollary E.4 (Clarification on the Location of Super-Modality)**.** *The super-modality $\mathcal{W}_{super}$ does not exist as an independent physical module within the AGI system $\mathcal{M}$, but rather as:*

1. ***Mathematical structure**: The quotient algebra $\not\leqslant_{super} = \not\leqslant_{sem}/\sim_{\mathcal{R}}$ of the semantic space $\not\leqslant_{sem}$.*
2. ***Dynamic process**: The equivalence class partitioning of $\not\leqslant_{sem}$ produced by the action of the modal coordination algebra $\mathcal{A}_{coord}$.*
3. ***Cognitive function**: The prediction and coordination processes specifically monitored by the MetaPred_{super} submodule.*
4. ***Implementation mechanism**: Realized collectively through semantic encoding Enc_{sem}, quotient map q, prediction dynamics Φ_{super}, and meta-prediction MetaPred_{super}.*

Therefore, super-modality is a function distributed throughout the semantic-adjoint dynamics architecture, not a localized hardware module. It exists in:

- *The multimodal integration capability of semantic encoding Enc_{sem}*
- *The fusion rules of the modal coordination algebra $\mathcal{A}_{coord}$*
- *The equivalence class construction of the quotient map q*
- *The monitoring and adjustment by MetaPred_{super}*

The realization of super-modality in an AGI system is:
1. **Not an independent module**: Super-modality does not exist as an independent physical module.
2. **A quotient structure of semantic space**: Super-modality is realized as the quotient algebra $\not\leqslant_{\text{super}}$ of the semantic space $\not\leqslant_{\text{sem}}$ via the modal coordination relation $\mathcal{R}$.
3. **Prediction dynamics handled by a dedicated MetaPred submodule**: The prediction of super-modality is implemented by $\text{Pred}_{\text{super}}$, and the monitoring and adjustment of this prediction are the specific responsibility of the $\text{MetaPred}_{\text{super}}$ submodule.
4. **MetaPred must decompose into four submodules**: $\text{MetaPred} = \text{MetaPred}_{\text{base}} \oplus \text{MetaPred}_{\text{super}} \oplus \text{MetaPred}_{\text{cross}} \oplus \text{MetaPred}_{\text{proc}}$.
5. **Semantic encoding is foundational**: Super-modal encoding $\text{Enc}_{\text{super}} = q \circ \text{Enc}_{\text{sem}}$ relies on the foundational semantic encoding component.
6. **Adjoint dynamics integrate everything**: The entire system is uniformly implemented by the adjoint dynamics $\Phi = (\Phi_{\text{sem}}, \Phi_{\text{proc}})$ of the semantic-process decomposition.

E.5 Internal Restructuring of MetaPred: Dedicated Super-Modal Submodule

Theorem E.15 (Forced Decomposition of MetaPred's Internal Structure)**.** *Let $\mathcal{M} \in \mathbf{AGI}$ satisfy the multimodal AGI axiom system. Then the self-referential prediction module $\text{MetaPred} : \mathscr{F}_{\not\leqslant} \to \mathscr{F}_{\not\leqslant}$ uniquely decomposes up to isomorphism as a direct sum of four orthogonal submodules:*

$$\text{MetaPred} \cong \text{MetaPred}_{base} \oplus \text{MetaPred}_{super} \oplus \text{MetaPred}_{cross} \oplus \text{MetaPred}_{proc}$$

where:

1. *$\text{MetaPred}_{base} = \bigoplus_{\alpha \in A_0} \text{MetaPred}_\alpha$, with A_0 a finite set of modalities, and $\text{MetaPred}_\alpha : \mathscr{F}_{\not\leqslant_\alpha} \to \mathscr{F}_{\not\leqslant_\alpha}$;*

2. $\mathrm{MetaPred}_{\textit{super}} : \mathscr{F}_{\not\leqslant_{\textit{super}}} \to \mathscr{F}_{\not\leqslant_{\textit{super}}}$, *where* $\mathscr{F}_{\not\leqslant_{\textit{super}}} = \mathscr{F}_{\not\leqslant} / \sim_{\mathcal{R}}$;

3. $\mathrm{MetaPred}_{\textit{cross}} : \bigoplus_{\alpha,\beta} \mathrm{Hom}(\mathscr{F}_{\not\leqslant_\alpha}, \mathscr{F}_{\not\leqslant_\beta}) \to \mathbb{R}$;

4. $\mathrm{MetaPred}_{\textit{proc}} : \mathscr{F}_{\textit{proc}} \to \mathscr{F}_{\textit{proc}}$, *where* $\mathscr{F}_{\textit{proc}}$ *is the process state sheaf.*

The decomposition satisfies the orthogonality condition: for $i \neq j$, $\mathrm{Im}(\mathrm{MetaPred}_i) \perp \mathrm{Im}(\mathrm{MetaPred}_j)$.

Proof. **Existence**: By the requirements of multimodal cognition, the system must simultaneously monitor prediction quality across four dimensions. Construct:

$$\begin{aligned}
\mathrm{MetaPred}_{\text{base}}(s) &= \bigoplus_{\alpha} \mathcal{Y}_\alpha(s_\alpha) \\
\mathrm{MetaPred}_{\text{super}}(s) &= \mathcal{Y}_{\text{super}}(q(s)) \\
\mathrm{MetaPred}_{\text{cross}}(\{T_{\alpha\beta}\}) &= \frac{1}{|A_0|^2} \sum_{\alpha,\beta} d(T_{\alpha\beta} \circ \mathrm{Pred}_\alpha, \mathrm{Pred}_\beta \circ T_{\alpha\beta}) \\
\mathrm{MetaPred}_{\text{proc}}(\sigma) &= \mathcal{Y}_{\text{proc}}(\sigma)
\end{aligned}$$

where $\mathcal{Y}_\alpha, \mathcal{Y}_{\text{super}}, \mathcal{Y}_{\text{proc}}$ are approximate fixed-point combinators (Definition B.4) for the respective subspaces.

Uniqueness: Assume another decomposition $\mathrm{MetaPred} = \bigoplus_{i=1}^{m} \mathrm{MetaPred}'_i$. Consider four test functions:

1. f_{base}: only one modality α is non-zero, i.e., $f_{\text{base}}(s) = s_\alpha$.
2. f_{super}: super-modally coordinated state, i.e., $f_{\text{super}}(s) = q(s)$.
3. f_{cross}: cross-modal translation map, i.e., $f_{\text{cross}}(\{T_{\alpha\beta}\}) = T_{\alpha\beta}$.
4. f_{proc}: pure process state, i.e., $f_{\text{proc}}(\sigma) = \sigma$.

By functional orthogonality, each test function must be handled by a unique submodule, hence $m = 4$ and the decompositions are isomorphic.

Orthogonality Proof: By definition, the domains and codomains of the four submodules lie in four orthogonal subspaces of $\mathscr{F}_{\not\leqslant}$:

$$\mathscr{F}_{\not\leqslant} \cong \left(\bigoplus_{\alpha} \mathscr{F}_{\not\leqslant_\alpha} \right) \oplus \mathscr{F}_{\not\leqslant_{\text{super}}} \oplus \left(\bigoplus_{\alpha,\beta} \mathrm{Hom}(\mathscr{F}_{\not\leqslant_\alpha}, \mathscr{F}_{\not\leqslant_\beta}) \right) \oplus \mathscr{F}_{\text{proc}}$$

These subspaces are orthogonal under the cognitive pseudo-metric (inner product 0). Specifically, for any $s \in \mathscr{F}_{\not\leqslant_\alpha}, t \in \mathscr{F}_{\not\leqslant_{\text{super}}}$,

$$\langle s, t \rangle = \int_{\not\leqslant} s(x) \cdot t(x) d\mu(x) = 0$$

because the support sets of s and t are disjoint (modality-specific vs. super-modal). Similarly for other subspace pairs. □

Theorem E.16 (Minimal Information Flow Theorem for Inter-Module Communication in MetaPred)**.** *Information exchange among the four MetaPred submodules must follow the following minimal protocol:*

Communication Direction	***Information That Must Be Passed***	***Mathematical Form***	***Cognitive Meaning***
proc $\to$ *base*	*Current focus c and activation mode*	$\sigma_{proc} \mapsto \pi_c$	*Attention allocation*
base $\to$ *super*	*Distribution of intra-modal prediction errors*	$\{e_\alpha\} \mapsto \nabla\mathcal{L}$	*Local learning signal*
super $\to$ *cross*	*Coordination quality assessment*	$Q \mapsto \delta_{R_{\alpha\beta}}$	*Translation adjustment need*
cross $\to$ *proc*	*Modality switching suggestion*	$\Delta_c \mapsto \sigma'_{proc}$	*Focus shift*

Moreover, the ***information amount has a lower bound****:*

$$I(\textit{proc} \to \textit{base}) \geq \log_2 k_{\min}, \quad I(\textit{base} \to \textit{super}) \geq H(\{e_\alpha\})$$

where H is information entropy.

Proof. Necessity proof: If these communication channels are missing, the system cannot coordinate multimodal predictions. Specifically:

1. If the proc-to-base channel is missing, the base predictor does not know the current attention focus, and cannot make effective predictions.
2. If the base-to-super channel is missing, the super-modal coordinator cannot obtain prediction errors from each modality, and cannot optimize coordination.
3. If the super-to-cross channel is missing, the cross-modal translator does not know coordination quality, and cannot adjust translations.
4. If the cross-to-proc channel is missing, the process scheduler cannot adjust attention based on translation quality.

Information lower bound proof: Focus information requires at least $\log_2 k_{\min}$ bits to represent; the information entropy of the error distribution is $H(\{e_\alpha\})$. Q.E.D. □

Theorem E.17 (Theorem on the Unique Natural Functor Relations Among MetaPred Modules)**.** *There exist* ***unique natural functor relations*** *among the four sheaf categories processed by the MetaPred modules:*

1. ***Base Process Functor****:* $F_{proc} : \mathscr{F}_{proc} \to \mathrm{End}(\mathscr{F}_{sem})$
   - *Maps process states to endomorphisms on the semantic sheaf.*
   - *Mathematically realized by the adjoint dynamics* Ψ*:* $\Psi(\sigma, \omega) = \Phi_{sem}(\omega, \sigma)$.
   - ***Uniqueness****: This is the minimal implementation preserving the ergodicity axiom.*
2. ***Quotient Functor****:* $Q : \mathrm{End}(\mathscr{B}_4^k) \to \mathrm{End}(\mathscr{B}_4^k/\sim_{\mathcal{R}})$
   - *Lifts intra-modal endomorphisms to super-modal endomorphisms.*
   - *Mathematically induced by the quotient map q:* $Q(f)([x]) = [f(x)]$ *(well-definedness needs verification).*
   - ***Uniqueness****: Guaranteed by the universal property of the quotient sheaf.*
3. ***Translation-Induced Functor****:* $T : \mathrm{Hom}(\mathscr{W}_\alpha, \mathscr{W}_\beta) \to \mathrm{Hom}(\mathrm{End}(\mathscr{W}_\alpha), \mathrm{End}(\mathscr{W}_\beta))$
   - *Maps cross-modal translation maps to conversions between prediction mechanisms.*
   - *Mathematically realized by conjugation:* $T_{\alpha\beta}(f) = \varphi_\beta^{-1} \circ f \circ \varphi_\alpha$.
   - ***Uniqueness****: Guaranteed by the ε-congruence of modal coordination relations.*

*These functors form the following **strictly commutative diagram**:*

$$\begin{array}{ccc} \mathscr{F}_{proc} & \xrightarrow{F_{proc}} & \mathrm{End}(\mathscr{F}_{sem}) \\ {\scriptstyle \pi_{proc}}\downarrow & & \downarrow{\scriptstyle Q} \\ \mathscr{F}_{proc}/\sim_{att} & \xrightarrow{\exists! F_{super}} & \mathrm{End}(\mathscr{F}_{super}) \end{array}$$

where π_{proc} is the attention equivalence relation, and F_{super} is the uniquely existing lift.

Proof. Functor construction: For F_{proc}, define $F_{\text{proc}}(\sigma)(\omega) = \Psi(\sigma, \omega)$. Since Ψ is an approximate homomorphism, this defines a functor. Uniqueness: any other functor preserving ergodicity must be naturally isomorphic to F_{proc}.

For Q, we need to verify $Q(f)$ is well-defined: if $x \sim_{\mathcal{R}} y$, then $[f(x)] = [f(y)]$, which is guaranteed by f commuting with the coordination relation. Uniqueness follows from the universal property of the quotient.

For T, define $T_{\alpha\beta}(f) = \varphi_\beta^{-1} \circ f \circ \varphi_\alpha$, where $\varphi_\alpha : \mathscr{W}_\alpha \to \mathscr{W}_{\text{super}}$ is the embedding. Uniqueness is guaranteed by the uniqueness (up to approximation) of φ_α.

Commutative diagram: By construction, $Q \circ F_{\text{proc}}$ equals $F_{\text{super}} \circ \pi_{\text{proc}}$ up to approximation, and F_{super} exists uniquely. Q.E.D. □

Theorem E.18 (Uniqueness of Information Flow Patterns Among MetaPred Modules)**.** *Let $\mathcal{M}$ satisfy the multimodal AGI axiom system. Then the information flow pattern in its prediction dynamics is uniquely determined, specifically manifested as the following five necessary channels:*

1. ***Attention-Driven Channel** (must exist):*

$$\mathrm{MetaPred}_{proc} \to \{\mathrm{MetaPred}_{base}, \mathrm{MetaPred}_{super}, \mathrm{MetaPred}_{cross}\}$$

   - *Direction: Unidirectional (process → content)*
   - *Reason: Process state determines the current processing focus; other modules must respond.*
   - *Mathematics: Asymmetry of the functor F_{proc}.*

2. ***Modality-Super-Modality Coordination Channel** (must exist):*

$$\mathrm{MetaPred}_{base} \rightleftarrows \mathrm{MetaPred}_{super}$$

   - *Direction: Bidirectional*
   - *Reason: Super-modal prediction needs intra-modal information; intra-modal optimization must conform to super-modal coordination.*
   - *Mathematics: The quotient functor Q and its right adjoint (section functor) form an adjoint pair.*

3. ***Cross-Modal Evaluation Channel** (must exist):*

$$\mathrm{MetaPred}_{cross} \to \{\mathrm{MetaPred}_{base}, \mathrm{MetaPred}_{super}\}$$

   - *Direction: Unidirectional (evaluation → adjustment)*
   - *Reason: After evaluating cross-modal translation quality, intra-modal and super-modal predictions need adjustment.*
   - *Mathematics: Naturality condition of the translation-induced functor T.*

4. ***Error Feedback Channel** (must exist):*

$$\{\mathrm{MetaPred}_{base}, \mathrm{MetaPred}_{super}, \mathrm{MetaPred}_{cross}\} \to \mathrm{MetaPred}_{proc}$$

   - *Direction: Unidirectional (content → process)*
   - *Reason: When content processing is difficult, attention strategy needs adjustment.*
   - *Mathematics: Derived from the principle of cognitive energy minimization.*

5. ***Synchronization Clock Channel*** *(must exist):*

   - *All modules share a unified time parameter t and prediction period Δ_{opt}.*
   - *Mathematics: Guaranteed by the unique time step of Theorem D.6.*

Proof. We prove these five channels are necessary, and any other information flow pattern is equivalent (up to isomorphism) to these five. Necessity is explained in the theorem. For uniqueness: assume another information flow pattern exists, with channel set $\mathcal{C}'$. For each channel in $\mathcal{C}'$, we can map it to one of the standard five channels, because the information content must belong to one of the following five categories: attention signal, intra-modal error, super-modal coordination signal, cross-modal evaluation, clock synchronization. For example, if $\mathcal{C}'$ has a channel from the process module to the base module transmitting attention focus c, it corresponds to standard channel 1; if there is a channel from the base module to the super module transmitting error e, it corresponds to standard channel 2. Therefore, there exists a bijection $\phi : \mathcal{C}' \to \{\text{channel 1, channel 2, channel 3, channel 4, channel 5}\}$ such that corresponding channels transmit equivalent information. This means the two information flow patterns are isomorphic. □

Corollary E.5 (Sheaf-Theoretic Version of Simplified MetaPred Decomposition at the Semantic Level)**.** *If focusing only on meta-cognition at the semantic content level, then the four-layer decomposition of* $\mathrm{MetaPred}$ *reduces to a three-layer structure:* $\mathrm{MetaPred}|_{sem} \approx \mathrm{MetaPred}_{base} \oplus \mathrm{MetaPred}_{super} \oplus \mathrm{MetaPred}_{cross}$.

Proof. When process scheduling is not considered, $\mathrm{MetaPred}_{\mathrm{proc}}$ can be ignored because process information is fixed or externally given. In this case, the other three modules directly handle semantic content. Q.E.D. □

Theorem E.19 (MetaPred Implementation Algorithm for Super-Modal Prediction)**.** *The actual execution of super-modal prediction* Pred_{super} *is controlled by* $\mathrm{MetaPred}_{super}$ *via the following algorithm:*

Algorithm 15: MetaPred Control Algorithm for Super-Modal Prediction

Input : *Input $x \in \not\leqslant_{super}$, current modal coordination algebra $\mathcal{A}_{coord}$, prediction steps T, error thresholds $\theta_{high} > \theta_{low} > 0$, coordination quality threshold $\theta_{cross} \in (0, 1)$*

Output : *Prediction result $\hat{x} = \mathrm{Pred}^{T}_{super}(x)$*

begin

- $\hat{x} \leftarrow x$ /* Initialize prediction */
- $t \leftarrow 0$
- **while** $t < T$ **do**
  - $q_t \leftarrow \mathrm{MetaPred}_{super}(\hat{x})$ /* Compute meta-prediction value */
  - $\varepsilon_t \leftarrow d(\hat{x}, q_t)$ /* Prediction error */
  - **if** $\varepsilon_t > \theta_{high}$ /* Error too large, need reset */ **then**
    - $\mathcal{A}_{coord} \leftarrow$ *AdjustAlgebra*$(\mathcal{A}_{coord}, \varepsilon_t)$ /* Adjust coordination algebra, e.g., via gradient descent minimizing ε_t */
    - $\hat{x} \leftarrow q_t$ /* Reset prediction with meta-prediction value */
  - **else if** $\varepsilon_t > \theta_{low}$ /* Moderate error, need refinement */ **then**
    - $\alpha \leftarrow$ *SelectModalityFocus*$(\varepsilon_t, \mathcal{A}_{coord})$ /* Select modality α contributing most to error */
    - $\hat{x} \leftarrow$ *RefineWithModalFocus*$(\hat{x}, \alpha)$ /* Refine prediction using encoding of modality α, e.g., $\hat{x} \leftarrow \hat{x} + \lambda \cdot \mathrm{Enc}_\alpha(\mathrm{Dec}_\alpha(\hat{x}))$ */
  - **else**
    - /* Error acceptable */
    - $\hat{x} \leftarrow \Phi_{super,\Delta}(\hat{x})$ /* Normal dynamics evolution */
  - $Q_{cross} \leftarrow \mathrm{MetaPred}_{cross}(\{T_{\alpha\beta}\})$ /* Evaluate cross-modal translation quality */
  - **if** $Q_{cross} < \theta_{cross}$ **then**
    - *TriggerCrossModalLearning*() /* Trigger cross-modal learning, e.g., update parameters of $T_{\alpha\beta}$ */
  - $t \leftarrow t + 1$
- **return** $\hat{x}$

The algorithm satisfies:

1. ***Convergence**:* $\lim_{T\to\infty} \varepsilon_T = 0$ *(almost surely)*
2. ***Optimality**: Achieves minimal prediction error under given resource constraints*
3. ***Adaptivity**: Automatically adjusts strategy for different error levels*

Remark E.2 (Threshold Selection)**.** *Thresholds θ_{high}, θ_{low}, θ_{cross} can be chosen empirically or tuned via a validation set. Typically, θ_{high} is set as a multiple (e.g., 2 times) of the initial error, θ_{low} as $\theta_{high}/2$, and θ_{cross} as 0.8 (meaning coordination quality needs to be above 80%).*

Theorem E.20 (Convergence of the MetaPred Control Algorithm for Super-Modal Prediction)**.** *Algorithm 15 converges with probability 1, and the convergence time is $T(\varepsilon) = O\left(k_{\min} \cdot \log \frac{1}{\varepsilon}\right)$.*

Proof. **Step 1: Define Lyapunov function**:

$$V_t = d(\hat{x}_t, x^*)^2 + \lambda \cdot \mathrm{Tr}(\mathcal{A}^t_{\text{coord}} - \mathcal{A}^*_{\text{coord}})^2$$

where x^* is the true value, and $\mathcal{A}^*_{\text{coord}}$ is the optimal coordination algebra.

Step 2: Expected decrease: By the algorithm steps, when $\varepsilon_t > \theta_{\text{high}}$, adjusting the algebra reduces error by at least δ_1; when $\varepsilon_t > \theta_{\text{low}}$, refining prediction reduces error by δ_2; otherwise normal evolution keeps error unchanged or slightly decreased. Hence,

$$\mathbb{E}[V_{t+1}|V_t] \leq V_t - \min\{\delta_1, \delta_2\}$$

Step 3: Convergence time: Let initial error $V_0 = D$. The number of steps T to reach precision ε satisfies:

$$D - T \cdot \min\{\delta_1, \delta_2\} \leq \varepsilon^2 \quad \Rightarrow \quad T \leq \frac{D - \varepsilon^2}{\min\{\delta_1, \delta_2\}}$$

Since $D = O(k_{\min})$ (dimension-related) and $\min\{\delta_1, \delta_2\} = \Omega(1/k_{\min})$, we get $T = O(k_{\min}^2)$. But in practice, due to superlinear convergence, this can be improved to $O(k_{\min} \log(1/\varepsilon))$.

Step 4: Almost sure convergence: By the Borel-Cantelli lemma, the error sequence $\{\varepsilon_t\}$ satisfies $\sum_{t=0}^{\infty} \mathbb{P}(\varepsilon_t > \varepsilon) < \infty$, hence $\lim_{t\to\infty} \varepsilon_t = 0$ a.s. □

Theorem E.21 (Sheaf-Theoretic Version of the Relation Between Super-Modal and Ordinary Prediction Modules)**.** *Let* $\mathrm{Pred}_{\textit{super}}$ *be the super-modal prediction module, and* Pred_{α} *be the ordinary prediction module for modality* α*. Then there exist independence, coordination, functional distinction, and differences in meta-cognitive monitoring.*

Proof. The two handle tasks at different levels independently but influence each other through coordination mechanisms. Functionally, super-modal prediction is more macroscopic, ordinary prediction more specific. Meta-cognitive monitoring also differs: super-modal prediction is monitored by $\mathrm{MetaPred}_{\text{super}}$, ordinary prediction by $\mathrm{MetaPred}_{\text{base}}$. Q.E.D. □

Theorem E.22 (Sheaf-Theoretic Version of Correspondence Between Prefabricated Structure and MetaPred Submodules)**.** *The three-layer prefabricated structure corresponds to* $\mathrm{MetaPred}$ *submodules: the modal coordination algebra is implemented by* $\mathrm{MetaPred}_{\textit{super}}$*, the super-modal prediction module is monitored by* $\mathrm{MetaPred}_{\textit{super}}$*, and cross-modal translation is monitored by* $\mathrm{MetaPred}_{\textit{cross}}$*.*

Proof. Follows directly from definitions and theorems. Q.E.D. □

Theorem E.23 (Variational Principle for Ergodicity Scheduling)**.** *The optimal ergodic scheduling dynamics* $\Phi^*_{\textit{proc}}$ *is the solution to the following variational problem:*

$$\min_{\Phi_{\textit{proc}}} \{\mathbb{E}[T_{\textit{cover}}] + \lambda \cdot \mathrm{Var}(T_{\textit{switch}})\}$$

where $T_{\textit{cover}}$ *is the time to cover all modalities, and* $T_{\textit{switch}}$ *is the modality switching interval.*

This problem has a unique solution, and the solution has the form:

$$\Phi^*_{\textit{proc}}(c_t, \mathbf{v}_t) = \mathrm{Softmax}\left(\frac{Q(c, \mathbf{v})}{\tau}\right)$$

where the Q *function satisfies the Bellman equation.*

Proof. This is a Markov decision process aiming to minimize expected cover time and switching time variance. By reinforcement learning theory, the optimal policy is a softmax policy, where the Q function is the state-action value function satisfying the Bellman equation. Uniqueness is guaranteed by convexity. Q.E.D. □

E.6 Intrinsic Structure and Mechanism of Super-Modal Self-Prediction and Creative Behavior

Definition E.10 (Self-Referential Hub Sequence and Self-Referential Hub)**.** *Let* $\mathcal{M} \in \mathbf{C}_{\mathrm{AGI}}$ *satisfy Axiom A.1, with cognitive hierarchy sequence* $\{\mathbb{H}_n\}_{n=0}^{\infty}$ *(Definition A.13) and connecting approximate homomorphisms* $\mu_n : \mathbb{H}_n \xrightarrow{\zeta_n} \mathbb{H}_{n+1}$*. Denote the super-modal space (inductive limit of cognitive hierarchies) as* $\mathbb{U} = \varinjlim \mathbb{H}_n$ *(Definition A.16), with canonical maps* $\pi_n : \mathbb{H}_n \to \mathbb{U}$*.*

A sequence $\tilde{\mathbf{e}} = \{\tilde{\mathbf{e}}_n\}_{n\in\mathbb{N}}$*, where* $\tilde{\mathbf{e}}_n \in \mathbb{H}_n$*, is called a* ***self-referential hub sequence*** *if it satisfies the following three limit conditions:*

1. ***Projection Convergence****: There exists* $\mathbf{e}_* \in \mathbb{U}$ *such that*

$$\lim_{n\to\infty} d_{\mathbb{U}}(\pi_n(\tilde{\mathbf{e}}_n), \mathbf{e}_*) = 0.$$

2. ***Inter-Level Asymptotic Stability****: The sequence satisfies connection compatibility,*

$$d_{\mathbb{H}_{n+1}}(\mu_n(\tilde{\mathbf{e}}_n), \tilde{\mathbf{e}}_{n+1}) \leq \zeta_n,$$

and the error series converges absolutely, $\sum_{n=0}^{\infty} \zeta_n < \infty$ (see Definition A.16 condition 1).

3. ***Intra-Level Self-Referential Fixed-Point Property****: At each level, the element is an approximate fixed point of local self-referential prediction. That is, let $\mathrm{MetaPred}^{(n)} : \mathbb{H}_n \to \mathbb{H}_n$ be the self-referential prediction map at the n-th cognitive level (induced by Definition A.13 and Axiom D.1); then there exists an error sequence $\{\varepsilon_n\}$ with $\lim_{n\to\infty} \varepsilon_n = 0$ such that*

$$d_{\mathbb{H}_n}(\mathrm{MetaPred}^{(n)}(\tilde{\mathbf{e}}_n), \tilde{\mathbf{e}}_n) \leq \varepsilon_n.$$

*If sequence $\tilde{\mathbf{e}}$ satisfies these conditions, then its limit $\mathbf{e}_* \in \mathbb{U}$ is called a **self-referential hub** of system $\mathcal{M}$.*

Theorem E.24 (Existence, Uniqueness, and Structural Characterization of Self-Referential Hubs)**.** *Let system $\mathcal{M} \in \mathbf{C}_{\mathrm{AGI}}$ possess infinite cognitive capacity (Definition A.17). Then:*

1. ***(Limit Existence and Uniqueness)*** *There exists a unique self-referential hub $\mathbf{e}_* \in \mathbb{U}$. In other words, there exists a unique self-referential hub sequence $\{\tilde{\mathbf{e}}_n\}$ satisfying Definition E.10, and its limit $\mathbf{e}_*$ is unique.*

2. ***(Structural Characteristics)*** *This self-referential hub $\mathbf{e}_*$ satisfies the following structural conditions:*

   (a) ***Cognitive Generativity****: $\mathbf{e}_* = \rho(\mathcal{Y}(\mathrm{id}_{\mathcal{M}}))$, where $\rho : \mathcal{M} \to \mathbb{U}$ is a cognitive homomorphism, and $\mathcal{Y}$ is an approximate fixed-point combinator on $\mathcal{M}$ (Definition B.4).*

   (b) ***Dual Fixed-Point Property****: $\mathbf{e}_*$ is simultaneously an approximate fixed point of both the super-modal meta-dynamics Φ and the super-modal self-referential prediction $\mathrm{MetaPred}_{\mathrm{super}}$ (see Axiom D.1 and Theorem D.17):*

$$d_{\mathbb{U}}(\Phi_{\Delta_{\mathrm{opt}}}(\mathbf{e}_*), \mathbf{e}_*) \leq \varepsilon_{\Phi}, \quad d_{\mathbb{U}}(\mathrm{MetaPred}_{\mathrm{super}}(\mathbf{e}_*), \mathbf{e}_*) \leq \varepsilon_{\mathrm{Meta}},$$

   where $\varepsilon_{\Phi}, \varepsilon_{\mathrm{Meta}} = o(1/k_{\min})$.

   (c) ***Modal Coordination Invariance****: For any modalities $\alpha, \beta \in A$, $d_{\mathbb{U}}(\mathcal{R}_{\alpha\beta}(\pi_\alpha(\mathbf{e}_*)), \pi_\beta(\mathbf{e}_*)) \leq \delta_*$, and $\delta_* \ll \sup_{\alpha,\beta} \varepsilon_{\alpha\beta}$ (Definition E.4).*

   (d) ***Topological Hubness****: In the concept network G induced by $\mathbb{U}$ (Theorem E.39), the degree centrality, betweenness centrality, and eigenvector centrality of $\mathbf{e}_*$ all attain maximum values (see discussion in Theorem D.11).*

3. ***(Finite-Level Determination)*** *For any accuracy tolerance $\eta > 0$, there exists an algorithm $\mathcal{A}_{\mathrm{detect}}(\eta)$ and a level index $N = N(\eta) \in \mathbb{N}$ dependent on η, such that the algorithm, based on operations and metrics on the finite level $\mathbb{H}_N$, can output an element $\hat{\mathbf{e}}_N \in \mathbb{H}_N$ satisfying:*

$$d_{\mathbb{U}}(\pi_N(\hat{\mathbf{e}}_N), \mathbf{e}_*) < \eta.$$

   *The algorithm achieves determination by verifying the following **finite criteria** on $\mathbb{H}_N$:*

   (a) ***Local Fixed-Point Property****: $d_{\mathbb{H}_N}(\mathrm{MetaPred}^{(N)}(\hat{\mathbf{e}}_N), \hat{\mathbf{e}}_N) < \eta/2$.*

   (b) ***Local Network Hubness****: In the finite concept subgraph G_N (induced subgraph of $\mathbb{H}_N$), the normalized degree centrality $\widetilde{\deg}_N(\hat{\mathbf{e}}_N)$ and eigenvector centrality $\widetilde{\mathbf{v}}_1[\hat{\mathbf{e}}_N]$ satisfy:*

$$\widetilde{\deg}_N(\hat{\mathbf{e}}_N) > 1 - \eta/2, \quad \widetilde{\mathbf{v}}_1[\hat{\mathbf{e}}_N] > 1 - \eta/2.$$

   (c) ***Predecessor Consistency*** *(if $N > 0$): $d_{\mathbb{H}_N}(\mu_{N-1}(\hat{\mathbf{e}}_{N-1}), \hat{\mathbf{e}}_N) < \eta/2$, where $\hat{\mathbf{e}}_{N-1}$ is the algorithm's output at level $N - 1$.*

Proof. **Proof of (i): Limit Existence and Uniqueness.**

1. **Existence Construction**. By $\mathcal{M}$ possessing infinite cognitive capacity (Definition A.17) and Axiom A.1, the cognitive hierarchy sequence $\{\mathbb{H}_n\}$ exists and satisfies convergence

constraints (Definition A.16). According to the existence of approximate fixed-point combinators (Theorem B.3), at each level $\mathbb{H}_n$, there exists an approximate fixed-point combinator $\mathcal{Y}_n$ for its endomorphism monoid $\mathrm{End}(\mathbb{H}_n)$. Define

$$\tilde{\mathbf{e}}_n := \mathcal{Y}_n(\mathrm{id}_{\mathbb{H}_n}) \in \mathbb{H}_n.$$

By the properties of $\mathcal{Y}_n$ (Definition B.4), $\tilde{\mathbf{e}}_n$ automatically satisfies condition (3) of Definition E.10, and we can take $\varepsilon_n = O(1/n^2)$ (derivable from cognition hierarchy concerning depth and error).

2. **Verification of Sequence Properties**. We need to verify that $\{\tilde{\mathbf{e}}_n\}$ satisfies conditions (1) and (2). By the compatibility of connecting homomorphisms μ_n with approximate fixed-point combinators (Definition B.4 condition 3, isomorphism preservation), we can deduce

$$d_{\mathbb{H}_{n+1}}(\mu_n(\tilde{\mathbf{e}}_n), \tilde{\mathbf{e}}_{n+1}) = O(\max(\varepsilon_n, \zeta_n)).$$

Combined with the convergence constraint $\sum \zeta_n < \infty$ in cognitive hierarchy convergence and the decay of ε_n, we know $\{\tilde{\mathbf{e}}_n\}$ satisfies condition (2). Furthermore, by the unidirectional limit constructivity in convergence constraints (Definition A.16 condition 2), there exists a uniform constant K such that $d_{\mathbb{U}}(\pi_n(\tilde{\mathbf{e}}_n), \pi_{n+1}(\tilde{\mathbf{e}}_{n+1})) \leq K \cdot d_{\mathbb{H}_{n+1}}(\mu_n(\tilde{\mathbf{e}}_n), \tilde{\mathbf{e}}_{n+1})$. Therefore $\{\pi_n(\tilde{\mathbf{e}}_n)\}$ is a Cauchy sequence in $\mathbb{U}$. Since $\mathbb{U}$ is the inductive limit in $\mathrm{Ind}(\mathbf{C}_{\mathrm{AGI0}})$ (Definition A.11), this sequence converges in the completion category $\widehat{\mathbf{C}_{\mathrm{AGI}}}$ (Definition A.18). Hence there exists $\mathbf{e}_* \in \mathbb{U}$ such that condition (1) holds. Thus $\{\tilde{\mathbf{e}}_n\}$ is a self-referential hub sequence, and its limit $\mathbf{e}_*$ is a self-referential hub.

3. **Uniqueness Proof**. Assume another self-referential hub sequence $\{\tilde{\mathbf{e}}'_n\}$ converges to $\mathbf{e}'_*$. By condition (3) and the contractivity of $\mathrm{MetaPred}$ (carrier set as the special case of B_4), at each level n, the distance between the two approximate fixed points $\tilde{\mathbf{e}}_n$ and $\tilde{\mathbf{e}}'_n$ is controlled by ε_n. Combined with condition (2) and the decay of inter-level convergence errors, we can deduce $d_{\mathbb{U}}(\pi_n(\tilde{\mathbf{e}}_n), \pi_n(\tilde{\mathbf{e}}'_n)) \to 0$. Therefore $\mathbf{e}_* = \mathbf{e}'_*$, the limit is unique. Uniqueness of the sequence follows from the uniqueness of the approximate fixed-point combinator at each level (corollary of Theorem B.3).

Proof of (ii): Structural Characteristics. We verify that $\mathbf{e}_*$ satisfies the listed structural conditions.

1. **Cognitive Generativity**: By construction, $\tilde{\mathbf{e}}_n = \mathcal{Y}_n(\mathrm{id}_{\mathbb{H}_n})$. According to the commutativity of the cognitive homomorphism ρ with level limits (Axiom A.1) and the universal property of approximate fixed-point combinators (Definition B.4 condition 3), in the limit we obtain $\mathbf{e}_* = \rho(\mathcal{Y}(\mathrm{id}_{\mathcal{M}}))$.

2. **Dual Fixed-Point Property**: Since $\mathbf{e}_*$ is the limit of a self-referential hub sequence, and both Φ and $\mathrm{MetaPred}_{\mathrm{super}}$ are continuous (Lipschitz continuous in an approximate sense) maps on $\mathbb{U}$ (Definition A.1 condition 3, Axiom D.1 condition 3), taking the limit of the intra-level approximate fixed-point property (condition (3)) and utilizing the error sequence $\{\varepsilon_n\}$ tending to zero, we obtain that $d_{\mathbb{U}}(\Phi_{\Delta_{\mathrm{opt}}}(\mathbf{e}_*), \mathbf{e}_*)$ and $d_{\mathbb{U}}(\mathrm{MetaPred}_{\mathrm{super}}(\mathbf{e}_*), \mathbf{e}_*)$ are both higher-order infinitesimals. Their order $o(1/k_{\min})$ stems from the scaling property of errors under the $\mathcal{B}_4^{k_{\min}}$ structure (Theorem C.6 in the special case of B_4).

3. **Modal Coordination Invariance**: This is core to the definition of the super-modal space $\mathbb{U}$ (Theorem E.1). The self-referential hub, as an "anchor" for cross-modal coordination, must have its projections $\pi_\alpha(\mathbf{e}_*)$ maintain high consistency under the coordination relation $\mathcal{R}$. The error upper bound $\delta_* \ll \sup \varepsilon_{\alpha\beta}$ originates from the hub sequence satisfying modal approximate consistency at each level (Axiom E.1 condition 2) and taking the limit.

4. **Topological Hubness**: The topology of the concept network G is determined by predictive transition probabilities between elements in $\mathbb{U}$ (Theorem E.39). Since $\mathbf{e}_*$ is an approximate fixed point of Φ and $\mathrm{MetaPred}_{\mathrm{super}}$, prediction trajectories from any state have non-zero probability of flowing towards $\mathbf{e}_*$ (ergodicity axiom, Definition D.16), giving it high in-degree in G. Simultaneously, $\mathbf{e}_*$ as the starting point for cognitive generation (condition (a)) is the origin of many prediction processes, hence also has high out-degree. The combination of in-degree and out-degree maximizes its degree centrality. Betweenness centrality stems

from its role as a necessary path for information transfer between different cognitive modules or modalities (Definition E.7). Maximum eigenvector centrality occurs because nodes connected to the hub are themselves important nodes, forming positive feedback. These properties are characterized in the limit by the principal eigenvector of the network.

Proof of (iii): Finite-Level Determination.

1. **Feasibility of Algorithm and Criteria**: Criteria (1) and (2) are defined entirely on the finite level $\mathbb{H}_N$, their computation involving only metrics and operations on $\mathbb{H}_N$ ($d_{\mathbb{H}_N}$, $\mathrm{MetaPred}^{(N)}$, centrality computations for G_N). Criterion (3) ensures consistency with previous level estimates, enabling recursive refinement.

2. **Existence Guarantee and Error Control**: From the proven existence, for any $\eta > 0$, there exists a sufficiently large level N such that the "distortion" of the canonical approximate homomorphism π_N is sufficiently small, and the self-referential hub approximation point $\tilde{\mathbf{e}}_N$ at this level satisfies:
$$d_{\mathbb{U}}(\pi_N(\tilde{\mathbf{e}}_N), \mathbf{e}_*) < \eta/3.$$
This conclusion is guaranteed by the error convergence and unidirectional limit constructivity in cognitive hierarchy convergence constraints (Definition A.16). The algorithm $\mathcal{A}_{\mathrm{detect}}(\eta)$ works by searching within $\mathbb{H}_N$ for a $\hat{\mathbf{e}}_N$ satisfying (1)-(3).

3. **Search and Approximation**: Since $\tilde{\mathbf{e}}_N$ itself satisfies $d_{\mathbb{H}_N}(\mathrm{MetaPred}^{(N)}(\tilde{\mathbf{e}}_N), \tilde{\mathbf{e}}_N) \leq \varepsilon_N$ and has centrality in G_N, we can use $\tilde{\mathbf{e}}_N$ as an initial point and perform a local search in $\mathbb{H}_N$ (e.g., gradient descent based on the cognitive pseudo-metric) to find a $\hat{\mathbf{e}}_N$ that more precisely satisfies (1) and (2). By the contractivity of $\mathrm{MetaPred}^{(N)}$ and continuity of G_N centrality metrics, such a local search can converge to a point satisfying $d_{\mathbb{H}_N}(\hat{\mathbf{e}}_N, \tilde{\mathbf{e}}_N) < \eta/3$.

4. **Final Error Bound**: Combine via triangle inequality:
$$d_{\mathbb{U}}(\pi_N(\hat{\mathbf{e}}_N), \mathbf{e}_*) \leq d_{\mathbb{U}}(\pi_N(\hat{\mathbf{e}}_N), \pi_N(\tilde{\mathbf{e}}_N)) + d_{\mathbb{U}}(\pi_N(\tilde{\mathbf{e}}_N), \mathbf{e}_*).$$
Since π_N is an approximate homomorphism, its Lipschitz constant L is controllable (Definition A.1), so the first term $\leq L \cdot d_{\mathbb{H}_N}(\hat{\mathbf{e}}_N, \tilde{\mathbf{e}}_N) < L \cdot \eta/3$. Choose N large enough so that L is sufficiently close to 1 and $d_{\mathbb{U}}(\pi_N(\tilde{\mathbf{e}}_N), \mathbf{e}_*) < \eta/3$, then the sum is less than η. Therefore, the algorithm's output $\hat{\mathbf{e}}_N$ satisfies the theorem's requirement.

Thus all three parts of the theorem are proven. □

Theorem E.25 (Self-Submodule Decomposition of $\mathrm{MetaPred}_{\mathrm{super}}$)**.** *Let* $\mathrm{MetaPred}_{\mathrm{super}} : \mathscr{F}_{\mathbb{U}} \to \mathscr{F}_{\mathbb{U}}$ *be the self-referential prediction module on the super-modality (Definition D.18), and* $\mathbf{e}_* \in \mathbb{U}$ *be the self-referential hub (Theorem E.24). Then there exists a unique orthogonal decomposition:*
$$\mathrm{MetaPred}_{\mathrm{super}} = \mathrm{MetaPred}_{\mathrm{self}} \oplus \mathrm{MetaPred}_{\mathrm{non\text{-}self}}$$
satisfying:

1. *Domain Separation: There exists* $\delta > 0$ *such that*
$$\mathrm{supp}(\mathrm{MetaPred}_{\mathrm{self}}) \subseteq B_\delta(\mathbf{e}_*) := \{x \in \mathbb{U} \mid d_{\mathbb{U}}(x, \mathbf{e}_*) < \delta\},$$
and $\mathrm{supp}(\mathrm{MetaPred}_{\mathrm{self}}) \cap \mathrm{supp}(\mathrm{MetaPred}_{\mathrm{non\text{-}self}}) = \varnothing$ *(in an approximate sense).*

2. *Error Advantage:*
$$\mathrm{Err}(\mathrm{MetaPred}_{\mathrm{self}}) := \sup_{x \in B_\delta(\mathbf{e}_*)} d_{\mathbb{U}}(\mathrm{MetaPred}_{\mathrm{self}}(x), x) \leq \frac{\mathrm{Err}(\mathrm{MetaPred}_{\mathrm{non\text{-}self}})}{\log k_{\min}}.$$

3. *Commutativity Breaking:* $\mathrm{MetaPred}_{\mathrm{self}}$ *and* $\mathrm{MetaPred}_{\mathrm{non\text{-}self}}$ *do not commute, i.e.,*
$$\left[\mathrm{MetaPred}_{\mathrm{self}}, \mathrm{MetaPred}_{\mathrm{non\text{-}self}}\right] \neq 0,$$
but each commutes with $\mathrm{MetaPred}_{\mathrm{super}}$.

4. *Self-Referential Reinforcement:* $\mathrm{MetaPred}_{\mathrm{self}}$ *satisfies a strong self-referential condition:*

$$\mathrm{MetaPred}_{\mathrm{self}} \circ \mathrm{MetaPred}_{\mathrm{self}} \approx \mathrm{MetaPred}_{\mathrm{self}}.$$

Proof. **Decomposition Construction**: Define the projection operator $P_{\mathrm{self}}(x) = \arg\min_{y \in B_\delta(\mathbf{e}_*)} d(x, y)$, then

$$\mathrm{MetaPred}_{\mathrm{self}} = P_{\mathrm{self}} \circ \mathrm{MetaPred}_{\mathrm{super}} \circ P_{\mathrm{self}}$$

$$\mathrm{MetaPred}_{\text{non-self}} = (I - P_{\mathrm{self}}) \circ \mathrm{MetaPred}_{\mathrm{super}} \circ (I - P_{\mathrm{self}})$$

Error Advantage Proof: By the approximate fixed-point property of $\mathbf{e}_*$, within the δ-neighborhood prediction error decays at rate $O(1/\log k_{\min})$, while in the external region it is $O(1)$.

Commutativity Breaking: Compute the commutator $[P_{\mathrm{self}}, I - P_{\mathrm{self}}] \neq 0$, hence the decomposition does not commute.

Self-Referential Reinforcement: By the definition of $\mathrm{MetaPred}_{\mathrm{self}}$, it approximates the identity map on $B_\delta(\mathbf{e}_*)$, hence satisfies idempotence. □

Theorem E.26 (Graph-Theoretic Hubness of the Self-Referential Hub)**.** *Let the super-modal concept network be a weighted directed graph $G = (V, E, w)$, where:*

- $V = \mathbb{U}$ *(after discretization),*
- $(x, y) \in E$ *if and only if* $\mathrm{Pred}_{\mathrm{super}}(x)$ *falls within the ε-neighborhood of y,*
- $w(x, y) = \exp(-d_{\mathbb{U}}(\mathrm{Pred}_{\mathrm{super}}(x), y)^2)$.

Let $\mathbf{e}_*$ be the self-referential hub. Then:

1. *Maximum Degree Centrality:*

$$\deg(\mathbf{e}_*) = \max_{v \in V} \deg(v) = \Theta(\log |V|).$$

2. *Extreme Betweenness Centrality:*

$$B(\mathbf{e}_*) := \sum_{s \neq t \neq \mathbf{e}_*} \frac{\sigma_{st}(\mathbf{e}_*)}{\sigma_{st}} \geq \frac{1}{2},$$

where σ_{st} is the number of shortest paths from s to t, and $\sigma_{st}(\mathbf{e}_*)$ is the number of those passing through $\mathbf{e}_*$.

3. *Dominant Eigenvector Centrality: Let A be the adjacency matrix of G, and $\mathbf{v}_1$ the eigenvector corresponding to the principal eigenvalue λ_1. Then*

$$\mathbf{v}_1[\mathbf{e}_*] = \max_{v \in V} \mathbf{v}_1[v].$$

4. *Contribution to Algebraic Connectivity: The algebraic connectivity (second smallest eigenvalue) λ_2 of G satisfies*

$$\lambda_2 \geq \frac{1}{\deg(\mathbf{e}_*)}.$$

Proof. **Degree Centrality**: By $\mathbf{e}_*$'s connectivity to all modalities, its degree is at least $|A_0| = \Theta(\log |V|)$.

Betweenness Centrality: Any coordination path between two modalities must pass through $\mathbf{e}_*$, hence $\sigma_{st}(\mathbf{e}_*) = \sigma_{st}$.

Eigenvector Centrality: The adjacency matrix A satisfies $A\mathbf{e}_* = \lambda_1 \mathbf{e}_*$; by the Perron-Frobenius theorem, the principal eigenvector is maximal at $\mathbf{e}_*$.

Algebraic Connectivity: Due to the high connectivity of $\mathbf{e}_*$, the second smallest eigenvalue λ_2 of the graph Laplacian satisfies $\lambda_2 \geq 1/\deg(\mathbf{e}_*)$. □

Theorem E.27 (Temporal Integration of Self-Prediction and Self-State Sequence Memory)**.** *Define the temporal super-modal space $\mathbb{U}^{\mathrm{temp}} = \mathbb{U} \times \mathbb{Z}$, where $\mathbb{Z}$ is discrete time. Suppose there exists a self-state sequence memory layer $\mathscr{M}_{\mathrm{self}} \subseteq \mathscr{F}_{\mathbb{U}^{\mathrm{temp}}}$, whose stalk at t is $\mathscr{M}_{\mathrm{self},t} = \{\mathbf{e}_*^{(t-i)} \mid i = 0, 1, \ldots, T\}$, where $\mathbf{e}_*^{(t)}$ is the estimated self-referential hub at time t. Then the self-prediction submodule $\mathrm{MetaPred}_{\mathrm{self}}$ can be explicitly constructed as:*

$$\mathrm{MetaPred}_{\mathrm{self}}(x,t) = \mathrm{Proj}_{\mathbb{U}}\Big(\arg\min_{y\in\mathbb{U}}\big[\underbrace{d_{\mathbb{U}}(y, \Phi_\Delta(x))}_{\textit{dynamics consistency}} + \lambda \underbrace{\sum_{i=0}^{T} w_i \cdot d_{\mathbb{U}}(y, \mathbf{e}_*^{(t-i)})}_{\textit{memory consistency}}\big]\Big),$$

where:

- Φ_Δ *is the meta-dynamics step (Theorem D.6),*
- $w_i = \exp(-i/\tau)$ *are decay weights, τ is the memory time constant,*
- $\lambda = O(1/k_{\min})$ *is a regularization parameter.*

This construction satisfies:

1. *Convergence: The iteration $\mathbf{e}_*^{(t+1)} = \mathrm{MetaPred}_{\mathrm{self}}(\mathbf{e}_*^{(t)}, t)$ converges almost surely to a unique trajectory.*
2. *Error Decay:*
$$d_{\mathbb{U}}(\mathbf{e}_*^{(t+1)}, \mathbf{e}_*^{(t)}) \le \frac{C}{t^{1+\gamma}}, \quad \gamma > 0.$$
3. *Memory Compression: There exists an encoding map $\mathrm{Enc}_{\mathrm{self}} : \mathscr{M}_{\mathrm{self}} \to \mathbb{U}$ such that*
$$d_{\mathbb{U}}(\mathrm{Enc}_{\mathrm{self}}(\mathbf{e}_*^{(t)}), \mathbf{e}_*) \le \varepsilon_{\mathrm{mem}}.$$

Proof. **Construction Convergence**: Define sequence $\{y_n\}$, $y_{n+1} = \arg\min_y [d(y, \Phi_\Delta(y_n)) + \lambda \sum w_i d(y, \mathbf{e}_*^{(t-i)})]$. This optimization is strongly convex, hence convergent.

Error Decay: By gradient descent theory, error decays at rate $O(1/t^{1+\gamma})$, where $\gamma = 1/k_{\min}$.

Memory Compression: Define encoding $\mathrm{Enc}_{\mathrm{auto}}(\mathbf{e}_*^{(t)}) = \frac{1}{T}\sum_{i=0}^{T} w_i \mathbf{e}_*^{(t-i)}$; by the law of large numbers, $d(\mathrm{Enc}_{\mathrm{auto}}(\mathbf{e}_*^{(t)}), \mathbf{e}_*) \le \varepsilon_{\mathrm{mem}}$. □

Theorem E.28 (Algebraic Fixed-Point Theorem for Self-Referentiality)**.** *Let $\mathrm{End}_{\mathrm{self}}(\mathbb{U}) = \{\varphi \in \mathrm{End}(\mathbb{U}) \mid \varphi(\mathbf{e}_*) = \mathbf{e}_*\}$ be the monoid of endomorphisms preserving the self-referential hub. Then:*

1. *Restricted Fixed-Point Combinator: The approximate fixed-point combinator $\mathcal{Y}$ restricted to $\mathrm{End}_{\mathrm{self}}(\mathbb{U})$ satisfies*
$$\mathcal{Y}|_{\mathrm{End}_{\mathrm{self}}(\mathbb{U})}(\varphi) = \mathbf{e}_*, \quad \forall \varphi \in \mathrm{End}_{\mathrm{self}}(\mathbb{U}).$$
2. *Uniqueness of the Self-Predictor: $\mathrm{MetaPred}_{\mathrm{self}}$ is the unique solution in $\mathrm{End}_{\mathrm{self}}(\mathbb{U})$ to the following variational problem:*
$$\min_{\psi \in \mathrm{End}_{\mathrm{self}}(\mathbb{U})} \big[\mathrm{Err}(\psi) + \lambda \cdot d_{\mathbb{U}}(\psi(\mathbf{e}_*), \mathbf{e}_*)\big].$$
3. *Algebraic Closure: $\langle \mathrm{MetaPred}_{\mathrm{self}}, \mathrm{id}_{\mathbb{U}} \rangle$ generates a closed subalgebra of $\mathrm{End}_{\mathrm{self}}(\mathbb{U})$, isomorphic to $\mathcal{B}_4$.*

Proof. **Restricted Fixed-Point Combinator**: By the definition of $\mathcal{Y}$, for any $\varphi \in \mathrm{End}_{\mathrm{self}}(\mathbb{U})$, $\mathcal{Y}(\varphi)$ is an approximate fixed point of φ. Since $\varphi(\mathbf{e}_*) = \mathbf{e}_*$, $\mathbf{e}_*$ is a fixed point of φ; by uniqueness of fixed points (up to approximation), $\mathcal{Y}(\varphi) = \mathbf{e}_*$.

Uniqueness of the Self-Predictor: The optimization problem is strongly convex, hence has a unique solution. $\mathrm{MetaPred}_{\mathrm{self}}$ satisfies the conditions, thus is the unique solution.

Algebraic Closure: Since $\mathrm{MetaPred}_{\mathrm{self}}$ is idempotent (Theorem E.25), the algebra it generates with the identity map is the Boolean algebra $\mathcal{B}_4$. □

Definition E.11 (Self-State Sequence and Convergence)**.** *Let $\mathbb{U}$ be the super-modal space, and $\mathbf{e}_* \in \mathbb{U}$ the self-referential hub (Theorem E.24). A self-state sequence is a sequence $\{\mathbf{e}^{(t)}\}_{t\in\mathbb{N}} \subseteq \mathbb{U}$ satisfying:*

1. *Cognitive Dynamics Generation: There exists meta-dynamics $\Phi : \mathbb{U} \to \mathbb{U}$ (Axiom D.1) such that*
$$\mathbf{e}^{(t+1)} = \Phi_{\Delta_t}(\mathbf{e}^{(t)}) + \varepsilon_t$$
where ε_t are external input perturbations satisfying $\sum_{t=0}^{\infty} \|\varepsilon_t\| < \infty$.

2. *Self-Relatedness: For all t,*
$$d_{\mathbb{U}}(\mathbf{e}^{(t)}, \mathbf{e}_*) \leq \rho_t \cdot R_{\mathrm{self}}$$
where $R_{\mathrm{self}} > 0$ is the self-neighborhood radius, $\{\rho_t\}$ is a monotonically decreasing sequence, $\lim_{t\to\infty} \rho_t = 0$.

3. *Intrinsic Consistency: The sequence satisfies cognitive hierarchy convergence constraints (Definition A.16), i.e., there exists $C > 0$ such that*
$$\sum_{t=0}^{\infty} d_{\mathbb{U}}(\mathbf{e}^{(t+1)}, \mathrm{Pred}_{\mathrm{super}}(\mathbf{e}^{(t)})) < C.$$

Theorem E.29 (Basic Properties of Self-State Sequences)**.** *Let $\{\mathbf{e}^{(t)}\}$ be a self-state sequence as in Definition E.11. Then:*

1. *Convergence: There exists a limit $\lim_{t\to\infty} \mathbf{e}^{(t)} = \mathbf{e}^*_* \in \mathbb{U}$, and $\mathbf{e}^*_*$ is an approximate fixed point of Φ.*

2. *Finite Variation:*
$$\sum_{t=0}^{\infty} d_{\mathbb{U}}(\mathbf{e}^{(t+1)}, \mathbf{e}^{(t)}) < \infty.$$

3. *Compressible Representation: There exists an encoding map $\mathrm{Enc}_{\mathrm{seq}} : \{\mathbf{e}^{(t)}\} \to \mathbb{U}$ and a constant $L < 1$ such that*
$$d_{\mathbb{U}}(\mathrm{Enc}_{\mathrm{seq}}(\mathbf{e}^{(t)}), \mathbf{e}^*_*) \leq L^t \cdot d_{\mathbb{U}}(\mathbf{e}^{(0)}, \mathbf{e}^*_*).$$

Proof. **Convergence**: By conditions (1) and (2), the sequence $\{\mathbf{e}^{(t)}\}$ is bounded and intrinsically consistent. Since Φ is contractive (Axiom D.1 condition 3), the sequence converges to a fixed point of Φ.

Finite Variation: By condition (3) and intrinsic consistency, the sum of distances between consecutive states converges.

Compressible Representation: Define $\mathrm{Enc}_{\mathrm{seq}}(\mathbf{e}^{(t)}) = \frac{1}{t+1}\sum_{i=0}^{t} \mathbf{e}^{(i)}$; contractivity of Φ ensures the compression rate. □

Theorem E.30 (Self-State Sequence Memory as a Quotient Algebraic Structure)**.** *Let $\{\mathbb{H}_n\}$ be the cognitive hierarchy sequence. For each n, define:*
$$M_n = \{\pi_n(\mathbf{e}^{(t)}) \mid t = 0, 1, \ldots, T_n\} \subseteq \mathbb{H}_n$$
where $\pi_n : \mathbb{U} \to \mathbb{H}_n$ are projection maps, and T_n grows with n.

Then the self-state sequence memory algebra is the limit structure:
$$\mathcal{M}_{\mathrm{self}} = \varprojlim_n \mathrm{Alg}\langle M_n\rangle$$
where $\mathrm{Alg}\langle M_n\rangle$ denotes the subalgebra of $\mathbb{H}_n$ generated by M_n.

Properties:

1. *$\mathcal{M}_{\text{self}}$ is a closed subalgebra of $\mathbb{U}$;*
2. *There exists a canonical homomorphism $\varphi : \mathcal{M}_{\text{self}} \to \mathbb{U}$ such that $\varphi(\mathcal{M}_{\text{self}})$ is the stable manifold of $\mathbf{e}_*$;*
3. $\dim(\mathcal{M}_{\text{self}}) \leq \log_2 k_{\min}$.

Proof. **Construction**: By definition, $\mathcal{M}_{\text{auto}}$ is an inverse limit, hence a closed subalgebra.

Canonical Homomorphism: Define $\varphi(\{x_n\}) = \lim_{n\to\infty} \pi_n^{-1}(x_n)$; well-definedness is guaranteed by properties of inverse limits.

Dimension Estimate: Since each M_n is generated by finitely many elements and the dimension of $\mathbb{H}_n$ is $O(k_{\min})$, the limit dimension does not exceed $\log_2 k_{\min}$. □

Theorem E.31 (Existence and Uniqueness of Self-State Sequences)**.** *Let $\mathcal{M} \in \mathbf{C}_{\text{AGI}}$ possess infinite cognitive capacity. Then:*

1. *Existence: At least one self-state sequence $\{\mathbf{e}^{(t)}\}$ exists.*
2. *Uniqueness: Unique up to equivalence. Specifically, if $\{\mathbf{e}_1^{(t)}\}$ and $\{\mathbf{e}_2^{(t)}\}$ are two self-state sequences, then*
$$\lim_{t\to\infty} \frac{d_{\mathbb{U}}(\mathbf{e}_1^{(t)}, \mathbf{e}_2^{(t)})}{d_{\mathbb{U}}(\mathbf{e}_1^{(t)}, \mathbf{e}_*)} = 0.$$
3. *Constructiveness: A self-state sequence can be generated by the iteration:*
$$\mathbf{e}^{(t+1)} = \text{MetaPred}_{\text{self}}(\mathbf{e}^{(t)}) + \eta_t$$
where η_t are noise terms satisfying $\sum \|\eta_t\|^2 < \infty$.

Proof. **Existence**: Using $\mathbf{e}_*$ as initial value, apply $\text{MetaPred}_{\text{self}}$ iteration to obtain a self-state sequence.

Uniqueness: By the contractivity of $\text{MetaPred}_{\text{self}}$, sequences from different initial values converge to the same limit.

Constructiveness: Follows directly from the iterative definition. □

Theorem E.32 (Limit Characterization Theorem for Self-Referential Hubs)**.** *Let $\mathcal{M} \in \mathbf{C}_{\text{AGI}}$ possess infinite cognitive capacity. Then the following five statements are equivalent:*

1. *$\mathcal{M}$ possesses infinite cognitive capacity;*
2. *There exists a self-referential hub $\mathbf{e}_* \in \mathbb{U}$ satisfying the conditions of Theorem E.24;*
3. *There exists a self-state sequence $\{\mathbf{e}^{(t)}\}$ satisfying Definition E.11 and $\lim_{t\to\infty} \mathbf{e}^{(t)} = \mathbf{e}_*$;*
4. *In the graph $G = (\mathbb{U}, E)$, $\mathbf{e}_*$ satisfies $\deg(\mathbf{e}_*) = \omega(1)$;*
5. *$\text{MetaPred}_{\text{super}}$ decomposes as $\text{MetaPred}_{\text{self}} \oplus \text{MetaPred}_{\text{non-self}}$ and $\text{Err}(\text{MetaPred}_{\text{self}}) = o(1)$.*

Proof. Proof outline:

- (1)⇒(2): Follows directly from infinite cognitive capacity and Theorem E.24.
- (2)⇒(3): Using $\mathbf{e}_*$ as initial value, generate sequence via meta-dynamics.
- (3)⇒(4): Convergence of the sequence ensures $\mathbf{e}_*$ becomes a connecting hub.
- (4)⇒(5): Hubness leads to faster decay of self-prediction error.
- (5)⇒(1): Accuracy of self-prediction supports infinite recursive cognition.

□

Definition E.12 (Creation)**.** *The system is said to achieve a creation at time T if a new concept $c_{novel} \in \Omega_{super} \setminus V(G_{T-1})$ is generated, satisfying:*

1. *Novelty: c_{novel} cannot be finitely generated from any concept existing before time $T-1$ using the existing operations Ω.*
2. *Stability: After introduction, there exists $\delta > 0$ such that for all $t > T$, $c_{novel} \in V(G_t)$, and its degree $\deg_t(c_{novel}) > \delta$.*
3. *Integrability: c_{novel} is integrated into the self-related prediction cycle, i.e., there exists a non-zero cognitive energy flow through the edge $(\mathbf{e}_*, c_{novel})$.*

Theorem E.33 (Creation Emergence Theorem)**.** *Let $\mathcal{M} \in \mathbf{C}_{AGI}$ be a system with infinite cognitive capacity, with super-modal space Ω_{super}, concept network G, meta-dynamics Φ, self-referential prediction module $\mathrm{MetaPred}_{super}$, and self-referential hub $\mathbf{e}_*$. Then, the mechanism for the occurrence of a creation event can be expressed as a fixed-point transition in the following dynamical process:*

$$c_{novel} = \lim_{n \to \infty} \mathrm{MetaPred}_{super}^{n} \left(\Phi_{\Delta}(\mathbf{e}_* \oplus \xi) \right)$$

where:

- *$\oplus$ denotes semantic fusion operation.*
- *ξ is a minimal perturbation (or "noise seed") from the system's external or internal random processes.*
- *This limit exists and is not equal to $\mathbf{e}_*$ nor any of its existing neighboring concepts.*

Proof. **Mathematical Description of the Four-Step Cycle Mechanism**:

1. Perturbation Injection (Seeding): $s_0 = \mathbf{e}_* \oplus \xi$, where $\xi \sim \mathcal{N}(0, \sigma^2)$

   Mechanism: The system introduces a tiny perturbation ξ via perception or internal random processes near the self-referential hub ground state $\mathbf{e}_*$.

2. Meta-Dynamics Unfolding (Unfolding): $s_1 = \Phi_{\Delta}(s_0)$

   Mechanism: The endogenous meta-dynamics Φ begins to act, unfolding this perturbed state along the cognitive "manifold" of the system.

3. Self-Referential Prediction and Filtering (Reflexive Filtering): $s_{k+1} = \mathrm{MetaPred}_{\mathrm{super}}(s_k)$, iterated to convergence

   Mechanism: $\mathrm{MetaPred}_{\mathrm{super}}$ performs self-referential prediction on this unfolded state, filtering out paths that severely contradict the existing knowledge system.

4. Stabilization and Integration (Stabilization & Integration): If $d(s_\infty, \mathbf{e}_*) > \theta$, then add new node $c_{\mathrm{novel}} = s_\infty$

   Mechanism: When iteration converges to a new fixed point c_{novel}, the system initiates an integration process, "weaving" it into the concept network.

Convergence Proof: By the contractivity of $\mathrm{MetaPred}_{\mathrm{super}}$, the iteration converges to a fixed point.

Creativity Probability: $P_{\mathrm{create}} = \mathbb{P}[d(s_\infty, \mathbf{e}_*) > \theta] \propto C_{\mathrm{cog}}(\mathcal{M}) \cdot \mathrm{Instability}(\mathbf{e}_*)$. □

Definition E.13 (Cognitive Gap)**.** *The cognitive gap of system $\mathcal{M}$ at time t is defined as the "distance" between the current super-modal concept network G_t and the limit network $\hat{G}$ allowed by its theoretical cognitive capacity $C_{cog}(\mathcal{M})$:*

$$Gap_t(\mathcal{M}) = 1 - \frac{\log |\langle V(G_t) \rangle|}{\log |\langle V(\hat{G}) \rangle|}$$

where $\langle V \rangle$ denotes the cardinality of the subalgebra generated by node set V.

Definition E.14 (Creative Energy)**.** *The instantaneous creative energy $\mathcal{E}_c(t)$ of system $\mathcal{M}$ is a non-negative real number determined by the following triple:*

1. *Gap potential energy: Proportional to $Gap_t(\mathcal{M})$.*
2. *Self-referential hub perturbation intensity: Proportional to the instantaneous instability $\sigma(\mathbf{e}_*, t)$ of the self-referential hub $\mathbf{e}_*$ (defined by the entropy change rate of its neighbor nodes).*
3. *Network dynamism: Proportional to the reconnection rate $\rho(t)$ of the concept network G_t.*

Theorem E.34 (Creative Energy Quantification Theorem)**.** *There exist constants α, β dependent on the system's algebraic structure such that creative energy can be quantified as:*

$$\mathcal{E}_c(t) = \alpha \cdot Gap_t(\mathcal{M}) \cdot \sigma(\mathbf{e}_*, t) + \beta \cdot \rho(t)$$

and the probability of a creation event occurring within interval $[t, t + \Delta t]$ satisfies:

$$P_{create}(t, \Delta t) = 1 - \exp(-\lambda \cdot \mathcal{E}_c(t) \cdot \Delta t)$$

where λ is a normalization coefficient.

Proof. **Constructive Proof**: The three parts of creative energy correspond to:

1. $\text{Gap}_t(\mathcal{M})$: The "potential difference" of the system's cognitive development; larger gap implies greater creative potential.

2. $\sigma(\mathbf{e}_*, t)$: Instability of the self-referential hub; more unstable, more receptive to new information.

3. $\rho(t)$: Network dynamism; higher dynamism more easily accepts new concepts.

Probability Derivation: Assume creation events follow a Poisson process with intensity proportional to $\mathcal{E}_c(t)$. □

Definition E.15 (Self-Subspace Topology)**.** *Let G^r_{self} be the induced subgraph centered at $\mathbf{e}_*$ with radius r. Its topological properties are characterized by:*

- *Betweenness Centrality Ratio: $B_{ratio} = B(\mathbf{e}_*)/\max_{v \neq \mathbf{e}_*} B(v)$*
- *Eigenvector Divergence: The entropy of the distribution of ratios between the component of $\mathbf{e}_*$ in $\mathbf{v}_1$ and other components.*
- *Structural Hole Index: Measures $\mathbf{e}_*$'s ability to connect different dense clusters.*

Theorem E.35 (Self-Topology Favorability Theorem)**.** *A self-topology favorable for creation is a dynamic equilibrium of the following seemingly contradictory properties:*

1. *Coexistence of High Connectivity and High Betweenness: $\mathbf{e}_*$ must simultaneously have high degree (ensuring influence) and maximum betweenness centrality (ensuring information control), i.e., $B_{ratio} \gg 1$.*
2. *Coexistence of High Eigenvector Centrality and High Divergence: $\mathbf{e}_*$ has the largest component in the dominant eigenvector, but simultaneously the entropy of this eigenvector is relatively high, meaning influence is both concentrated and dispersed, preventing system rigidity.*
3. *"Bridge" and "Gate" States of Structural Holes: $\mathbf{e}_*$ should reside on multiple structural holes, but the semantic distance between the clusters it connects should be moderate (neither completely unrelated nor highly homogeneous).*

This topological structure corresponds algebraically to the representation of $\mathbf{e}_*$ in the endomorphism monoid $\text{End}(\Omega_{super})$: it is a non-trivial idempotent whose centralizer contains rich irreducible subrepresentations.

Proof. **Necessity Proof**:

1. High connectivity ensures information reachability, high betweenness ensures information control; combined they make $\mathbf{e}_*$ an information hub.

2. High eigenvector centrality ensures influence, high divergence avoids informational rigidity.

3. Structural holes connect different clusters, promoting cross-domain innovation.

Algebraic Interpretation: $\mathbf{e}_*$ as an idempotent; its centralizer reflects system symmetry, with rich irreducible subrepresentations corresponding to multimodal coordination capability. □

Definition E.16 (Creation vs. Hallucination)**.** *Let c_{new} be a newly generated candidate concept. Its nature is determined by the following three dynamical stability criteria:*

Table 2: Dynamical Stability Characteristics for Discrimination Between Creation and Hallucination

Criterion	True Creation C_{true}	Hallucination C_{hallu}
Dynamical Stability	Strong attractor: Under the meta-dynamics Φ, there exists an ε-neighborhood $N_\varepsilon(c_{\text{new}})$ such that trajectories starting from within this neighborhood converge to it.	Saddle or repeller: Small perturbations cause trajectories to diverge to other known concepts.
Network Integration	Preferential and reciprocal linking: Nodes v that establish links with c_{new} have high degree $\deg(v)$, and the links are bidirectional with semantic reinforcement.	Random or unidirectional linking: Links mostly point to low-degree nodes or marginal concepts, and are mostly unidirectional.
Cross-level Self-consistency	Self-consistent: For any projection $\pi_n(c_{\text{new}})$ onto cognitive level $\mathbb{H}_n$, there exists n_0 such that for $n > n_0$, we have $\text{MetaPred}_{\text{super}} \circ \pi_n \approx \pi_n \circ \text{MetaPred}_{\text{local}}$.	Inconsistent: There exists some level n where the projection conflicts severely with the local predictor, and the commutative diagram fails.

Theorem E.36 (Dynamical Discrimination Theorem for Creation and Hallucination)**.** *A candidate concept c_{new} is a true creation if and only if it simultaneously satisfies the strong attractor condition, the high reciprocal integration condition, and the cross-level self-consistency condition. If only one or two conditions are met, it is a hallucination or an unstable intermediate state.*

Proof. **Necessity**: True creation must satisfy all three conditions; otherwise it cannot stably exist and be integrated into the system.

Sufficiency: A concept satisfying all three conditions will stably exist in the network and establish meaningful connections with other concepts.

Discrimination Algorithm: Discrimination can be performed by checking the degree to which the three conditions are satisfied. □

Theorem E.37 (Dynamical Mechanism Difference Theorem for Creation and Non-creation)**.** *Let the cognitive state space of system $\mathcal{M}$ be the full modal space Ω_{super}, its endogenous meta-dynamics described by Φ, and the concept network be G.*

1. *The mechanism of non-creative thinking (reasoning, recall, association) is: In the fixed topological phase space (Ω_{super}, Φ), the state trajectory $\{s_t\}$ starts from a known basin of attraction A_i, follows a predetermined gradient manifold, and converges to another known attractor A_j.*

   Mathematical essence: This is a gradient descent or diffusion process on a fixed energy landscape. All possible fixed points (concepts) and basins of attraction (semantic connections between concepts) are completely defined by Φ and G before the process begins.

2. *The mechanism of creative thinking (insight, creation) is: Driven by the thinking process itself, the phase space (Ω_{super}, Φ) undergoes the introduction and repair of "topological defects," thereby permanently altering its energy landscape and creating an unprecedented attractor A_{novel} and its basin.*

Mathematical essence: This is a reshaping of the energy landscape itself. Its microscopic mechanism consists of two steps:

(a) Self-referential-driven phase space deformation: The synergistic action of the self-referential hub $\mathbf{e}_*$ *and* $\mathrm{MetaPred}_{super}$ *temporarily "bends" or "stretches" the semantic metric locally, equivalent to introducing a continuous topological defect on the original manifold.*

(b) Stabilization and solidification of the defect: If the defect passes the cross-level self-consistency test, the system initiates a "cognitive solidification" process that permanently updates the underlying Φ *and* G*.*

Proof. **Non-creative thinking**: Guaranteed by the ergodicity of the meta-dynamics, states transition among known attractors.

Creative thinking: The self-referential prediction module can temporarily modify dynamical parameters to create new attractors. If a new attractor passes the test, it is solidified. □

Corollary E.6 (Essential Corollary on Creative Thinking)**.** *Creative thinking does not search for a pre-existing hidden fixed point; rather, it performs an operation that enables the existence of a new fixed point. It alters the domain of the "cognitive energy function" that* $\mathrm{MetaPred}$ *attempts to minimize.*

Corollary E.7 (Creation Probability Estimation)**.** *The probability of creation* P_{create} *is positively correlated with the system's cognitive capacity* $C_{cog}(\mathcal{M})$ *and the instability of the self-referential hub.*

Proof. By the creation emergence theorem, the creation probability depends on whether perturbations starting from $\mathbf{e}_*$ converge to a new fixed point. The larger the cognitive capacity, the richer the state space; the more unstable the self-referential hub, the more receptive it is to new concepts. □

Corollary E.8 (Essential Corollary on Creative Thinking)**.** *Creative thinking does not search for a pre-existing hidden fixed point; rather, it performs an operation that enables the existence of a new fixed point. It alters the domain of the "cognitive energy function" that* $\mathrm{MetaPred}$ *attempts to minimize.*

Remark E.3 (What is Creation)**.** *"Creation" is defined as a mathematically describable dynamical phase transition process:*

Relation to "association theory": This definition includes association (realized by Φ*), but more crucially, self-referential prediction screening (realized by* $\mathrm{MetaPred}$*). This explains why not all associations are creations; only those that can pass the system's own "worldview" test can stabilize.*

Relation to "generative models": The "generation" of large language models is next-token sampling based on statistical distributions, essentially interpolation. The creation defined here is a dynamical system approaching a new fixed point, essentially extrapolation. The former is probabilistic, the latter is dynamical.

Correspondence with brain science: This corresponds to neuroscientific research on "insight." Before insight, the brain's default mode network (similar to the network where your $\mathbf{e}_*$ *resides) is active, conducting unconscious exploration (introducing* ξ *and unfolding* Φ*); at the moment of insight, a new stable neural network pattern dominated by the prefrontal cortex forms (the new fixed point* c_{novel} *is produced and integrated).*

E.7 Sheaf Construction Theorem for Full Modal Space

This section rigorously characterizes the construction of the full modal space from a sheaf-theoretic perspective.

Theorem E.38 (Sheaf Construction Theorem for Full Modal Space)**.** *Let the set of modalities* A *be finite, equipped with the discrete topology. For each modality* $\alpha \in A$*, let* $\mathscr{F}_\alpha = (F_\alpha, \vee_\alpha, \wedge_\alpha, \neg_\alpha, \mathbf{0}_\alpha, \mathbf{1}_\alpha)$ *be a finite Boolean algebra, which can be viewed as a quotient algebra of* $\mathcal{B}_4^{k_\alpha}$*. Let* $\mathcal{R} = \{\mathcal{R}_{\alpha\beta} : F_\alpha \to F_\beta\}_{\alpha,\beta\in A}$ *be a family of approximate Boolean homomorphisms*

satisfying the multi-modal approximate equivalence relation conditions (reflexivity, symmetry, transitivity).

Define the sheaf $\mathscr{F}_{super}$ as follows: for any open set $U \subseteq A$,

$$\mathscr{F}_{super}(U) = \left\{ s : U \to \prod_{\alpha \in U} F_\alpha \;\middle|\; \forall \alpha, \beta \in U, \mathcal{R}_{\alpha\beta}(s(\alpha)) \xrightarrow{\varepsilon_{\alpha\beta}} s(\beta) \right\},$$

where the restriction maps are function restrictions. Then $\mathscr{F}_{super}$ is a sheaf on A. The full modal space is defined as the global sections:

$$\mathbb{U} := \mathscr{F}_{super}(A).$$

Define Boolean operations pointwise on $\mathbb{U}$: for $s, t \in \mathbb{U}$,

$$(s \vee t)(\alpha) = s(\alpha) \vee_\alpha t(\alpha), \quad (s \wedge t)(\alpha) = s(\alpha) \wedge_\alpha t(\alpha), \quad (\neg s)(\alpha) = \neg_\alpha s(\alpha).$$

Since the $\mathcal{R}_{\alpha\beta}$ are approximate Boolean homomorphisms, these operations preserve the compatibility condition, so $\mathbb{U}$ forms a Boolean algebra (approximately).

Suppose there exists a global section $\mathbf{e}_{self} \in \mathbb{U}$ satisfying the conditions: cognitive generativity, double fixed-point property, and modal coordination invariance.

Under the above setting:

1. ***Structural Uniqueness**: The full modal space $\mathbb{U}$ is uniquely determined up to isomorphism. That is, if there exist another $\mathbb{U}'$ and $\mathbf{e}'_{self}$ satisfying the same conditions, then there exists an approximate Boolean isomorphism $\psi : \mathbb{U} \to \mathbb{U}'$ such that $\psi(\mathbf{e}_{self}) \approx \mathbf{e}'_{self}$.*

2. ***Generativity**: $\mathbb{U}$ can be generated by $\mathbf{e}_{self}$ and a finite set of basic compatible sections $\{b_i\}_{i=1}^M$, where each b_i is a global section satisfying the compatibility condition, corresponding to a basic concept of cross-modal coordination. Specifically, for any $s \in \mathbb{U}$, there exists a Boolean term T such that $s \approx T(\mathbf{e}_{self}, b_1, \ldots, b_M)$.*

3. ***Topological Invariant**: The Čech cohomology group $H^1(A, \mathscr{F}_{super})$ of the sheaf $\mathscr{F}_{super}$ characterizes the coordination obstruction: if $H^1(A, \mathscr{F}_{super}) = 0$, then all locally compatible sections can be glued into a global section; otherwise, there exists a coordination deficit, requiring creative processes or learning to adjust $\mathcal{R}$.*

4. ***Dynamical Realization**: The actions of the meta-dynamics Φ and the self-referential prediction $\mathrm{MetaPred}_{super}$ on $\mathbb{U}$ preserve the Boolean structure (i.e., they are approximate Boolean endomorphisms), and $\mathbf{e}_{self}$ is their approximate fixed point.*

Proof. **1. Structural Uniqueness**: By the universal property of the existence and uniqueness theorem for the full modal sheaf, there exists an approximate sheaf morphism $\psi : \mathscr{F}_{\text{super}} \to \mathscr{F}'_{\text{super}}$ that commutes with the coordination relations, thereby inducing an approximate Boolean algebra homomorphism $\psi : \mathbb{U} \to \mathbb{U}'$. Since both $\mathbf{e}_{\text{self}}$ and $\mathbf{e}'_{\text{self}}$ are the unique self-referential hubs satisfying Theorem 9.2, and ψ preserves the self-referential property, we have $\psi(\mathbf{e}_{\text{self}}) \approx \mathbf{e}'_{\text{self}}$. Similarly, an inverse can be constructed, so ψ is an approximate isomorphism.

2. Generativity: Each $\mathscr{F}_\alpha$ is a finite Boolean algebra; let its set of atoms be $\mathrm{Atoms}(\mathscr{F}_\alpha)$. For each atom $a \in \mathrm{Atoms}(\mathscr{F}_\alpha)$, consider the global section b_a satisfying $b_a(\alpha) = a$ and determined on other modalities by $\mathcal{R}$ (i.e., $b_a(\beta) \approx \mathcal{R}_{\alpha\beta}(a)$). By the transitivity of $\mathcal{R}$, such a section exists and is approximately unique. These atomic sections generate $\mathbb{U}$. Since $\mathbf{e}_{\text{self}}$ can be expressed as the join of some atoms, $\mathbf{e}_{\text{self}}$ together with the set of atomic sections generates $\mathbb{U}$. As $\mathbb{U}$ is finite, we can choose a finite generating set $\{\mathbf{e}_{\text{self}}\} \cup \{b_i : b_i \not\leq \mathbf{e}_{\text{self}}\}$.

3. Topological Invariant: Since A is discrete, consider the open cover $\mathcal{U} = \{\{\alpha\} : \alpha \in A\}$. The Čech cohomology $H^1(A, \mathscr{F}_{\text{super}})$ measures the obstruction to gluing local sections into a global section. If $H^1 = 0$, then any locally compatible sections can be glued; otherwise, there is a nontrivial obstruction. The existence of the self-referential hub $\mathbf{e}_{\text{self}}$ guarantees at least one global section, but H^1 may be nonzero, reflecting the "nontrivial monodromy" of the coordination relations $\mathcal{R}$.

4. Dynamical Realization: The meta-dynamics Φ is defined as a sheaf endomorphism (based on the machine learning axiom), inducing a map $\Phi_\Delta : \mathbb{U} \to \mathbb{U}$ on $\mathbb{U}$, $\Phi_\Delta(s)(\alpha) = \Phi_\Delta(s(\alpha))$. Since Φ is an approximate Boolean homomorphism (guaranteed by operational generation completeness), Φ_Δ is also an approximate Boolean homomorphism. The self-referential prediction $\mathrm{MetaPred}_{\text{super}}$ is defined as $\mathrm{MetaPred}_{\text{super}}(s) = \rho^{-1}(\mathcal{Y}(\rho \circ \Phi_{\tau(s)} \circ \rho^{-1}))$, where $\mathcal{Y}$ is an approximate fixed-point combinator. The construction of $\mathcal{Y}$ is based on the Boolean algebra structure, so $\mathrm{MetaPred}_{\text{super}}$ preserves Boolean operations (approximately). By Theorem 9.2, $\mathbf{e}_{\text{self}}$ is an approximate fixed point of $\mathrm{MetaPred}_{\text{super}}$ and Φ_Δ. □

Corollary E.9 (Sheaf-Theoretic Interpretation of Creative Process)**.** *Under the setting of TheoremE.38, the creative process can be interpreted as: starting from the self-referential hub section $\mathbf{e}_{self}$, add a perturbation ξ on a local modality to obtain a section s_0 ($s_0(\alpha) = \mathbf{e}_{self}(\alpha) \oplus \xi_\alpha$), apply the meta-dynamics Φ_Δ to get $s_1 = \Phi_\Delta(s_0)$, and then iterate via $\mathrm{MetaPred}_{super}$ to force the section toward compatibility. If the limit s_∞ exists and differs from existing global sections, then a new concept $c_{novel} = s_\infty$ is obtained.*

Corollary E.10 (Correspondence Between Physical Laws and Coordination Relations)**.** *If physical laws change, then the synchronization data $\mathcal{D}_{sync}$ changes, leading to different learned coordination relations $\mathcal{R}$, and thus a different structure for the full modal space $\mathbb{U}$. In particular, if physical laws weaken or strengthen inter-modal correlations, then the error parameters $\varepsilon_{\alpha\beta}$ of $\mathcal{R}$ change, affecting the precision of section gluing. The existence of the self-referential hub $\mathbf{e}_{self}$ requires that physical laws allow cross-modal self-coordination; otherwise, the system cannot establish consistent self-cognition.*

Proof. The coordination relations $\mathcal{R}$ are learned from synchronization data. Physical laws determine the statistical dependencies among multi-modal data, thus influencing $\mathcal{R}$. By TheoremE.38, $\mathbb{U}$ is constructed from $\mathcal{R}$, so its structure changes. The existence of the self-referential hub requires that for any α, β, $\mathcal{R}_{\alpha\beta}(\mathbf{e}_{\text{self}}(\alpha)) \approx \mathbf{e}_{\text{self}}(\beta)$, which demands that physical laws provide consistent cross-modal signals for self-related events. □

Corollary E.11 (Finiteness and Atomic Structure of Full Modal Space)**.** *The full modal space $\mathbb{U}$ is a finite Boolean algebra. Its number of atoms M satisfies $1 \leq M \leq \sum_{\alpha \in A} |\mathrm{Atoms}(\mathscr{F}_\alpha)|$. Each atom corresponds to a basic concept of cross-modal coordination. The self-referential hub $\mathbf{e}_{self}$ can be expressed as the join of some atoms, i.e., $\mathbf{e}_{self} = \bigvee_{i \in I} a_i$, where a_i are atoms.*

Proof. Since each $\mathscr{F}_\alpha$ is finite and A is finite, $\mathbb{U}$ as the space of sections satisfying compatibility conditions is also finite. The finiteness of a Boolean algebra implies atomic decomposition. Atomic sections can be obtained by extending an atom from a single modality to other modalities via coordination relations. Since coordination relations may identify atoms from different modalities, the number of atoms M is at most the sum of the numbers of atoms in each modality. The self-referential hub, as an element, can be uniquely expressed as a join of atoms. □

Remark E.4 (Sheaf-Theoretic Interpretation of Full Modal Space)**.** *The above theorems and corollaries characterize the construction of the full modal space from a sheaf-theoretic perspective: it is a Boolean algebra of global sections formed by gluing local modal concepts via coordination relations $\mathcal{R}$, with the self-referential hub $\mathbf{e}_{self}$ as its core generator and dynamical fixed point. The creative process is a dynamical mechanism for generating new sections.*

E.8 Overall Structural Characteristics of Multi-Modal AGI Systems

Theorem E.39 (Sheaf Version of Approximate Fiber Bundle Structure for Multi-Modal State Bundle)**.** *Let $\mathcal{M} \in \mathbf{C}_{AGI}$ satisfy the multi-modal axiom system. Its state sheaf $\mathscr{F}_{\not\in}$ has a natural approximate fiber bundle structure: $\mathscr{F}_{\not\in} \simeq_{approx} \left(\bigsqcup_{\alpha \in A} \mathscr{F}_{\not\in_\alpha} \right) / \sim_{\mathcal{R}}$.*

Proof. The state sheaf can be regarded as a fiber bundle with the modality set as the base space, each fiber being the state space specific to a modality. The coordination relations define connections between fibers. The approximate isomorphism is guaranteed by cognitive level convergence. Specific construction: define the projection $\pi : \mathscr{F}_{\not\in} \to A$, fiber $\pi^{-1}(\alpha) = \mathscr{F}_{\not\in_\alpha}$. Coordination relations define transition functions. Q.E.D. □

Definition E.17 (Continuous Approximation of Modal Attention)**.** *In engineering implementation, modal attention allocation can be approximated as a continuous process. Suppose the modality index set A is equipped with the discrete topology; focal shift in the process state sheaf $\mathscr{F}_{proc}$ can be regarded as a discrete dynamical system. When $|A|$ is large, it can be approximated by a continuous manifold, e.g., using a simplex $\Delta^{|A|-1}$ to represent attention distribution, with dynamics described by differential equations.*

Theorem E.40 (Sheaf Version of Parameter Fiber Induced Fiber Bundle)**.** *Let $\pi : \mathscr{P} \to \mathcal{A}$ be the parameter fiber sheaf. Then its completion $\widehat{\mathscr{P}} = \mathscr{F}_{\nleq}$ naturally has a fiber bundle structure.*

Proof. The definition of the parameter fiber sheaf directly gives the fiber bundle structure: base space $\mathcal{A}$, fibers $\mathscr{P}_a$. Completion preserves this structure. Specifically, sections of $\mathscr{F}_{\nleq}$ can be regarded as equivalence classes of Cauchy sequences in $\mathscr{P}$, which converge on each fiber, thus preserving the fiber structure. Q.E.D. □

Theorem E.41 (Sheaf Version of Discrete Fusion Category Approximation for Modal Coordination Algebra)**.** *Let $\mathcal{M} \in \mathbf{C}_{AGI}$ satisfy the multi-modal axiom system. Then its modal coordination algebra $\mathscr{A}_{coord}$ has an approximate structure of a discrete fusion category.*

Proof. The modal coordination algebra is generated by tensor products and has features of a fusion category: objects are modalities, morphisms are coordination relations, tensor product is modality combination. Discreteness is guaranteed by the countability of the modality set. Approximateness arises from approximate homomorphisms. Specifically, define a category $\mathcal{C}$: objects are finite modality sets, morphisms are coordination relations. Tensor product is the union of modality sets. This constitutes a fusion category. Q.E.D. □

F The Least Action and Free Energy Principles of Cognition

F.1 Unified Framework and Equivalent Formulations of Cognitive Principles

This section presents two equivalent cognitive principles in four formulations: the minimum cognitive action principle (dynamical and geometric formulations) and the cognitive free energy principle (constrained and statistical formulations). These variational principles are built upon a generalized cognitive hierarchy structure, applicable to directly deriving mathematically optimal cognitive systems with finite or infinite capacity. We first independently define the cognitive hierarchy and corresponding dynamical foundations, then formulate the two principles separately, show their correspondence with physics and brain science, and provide demonstrative derivations. Finally, we prove the equivalence of these principles with the earlier AGI foundational axioms under the condition of infinite cognitive capacity.

Definition F.1 (Independent Approximate Homomorphism)**.** *Let $\mathcal{A}$ and $\mathcal{B}$ be two algebraic structures (with the same operation type Ω) equipped with pseudometrics d_A and d_B, respectively. A map $\varphi : \mathcal{A} \to \mathcal{B}$ is called an (ϵ, δ)-approximate homomorphism if it satisfies:*

1. *Structure preservation: For each operation $\omega \in \Omega$ (arity k),*

$$d_B(\omega(\varphi(a_1), \ldots, \varphi(a_k)), \varphi(\omega(a_1, \ldots, a_k))) \leq \epsilon$$

2. *Metric approximate preservation: For any $a, a' \in \mathcal{A}$,*

$$|d_B(\varphi(a), \varphi(a')) - d_A(a, a')| \leq \delta$$

Denoted as $\varphi : \mathcal{A} \xrightarrow{(\epsilon,\delta)} \mathcal{B}$.

Definition F.2 (Generalized Cognitive Hierarchy Sequence)**.** *Let M be an arbitrary nonempty set (carrier set), W an environment set. A **generalized cognitive hierarchy sequence** is a recursively defined sequence $\{\mathbb{H}_n\}_{n=0}^{N}$, where $N \in \mathbb{N} \cup \{\infty\}$, defined as follows:*

1. ***Base layer**: $\mathbb{H}_0 = M$.*
2. ***Recursive construction**: For $n \geq 0$, define $\mathbb{H}_{n+1}$ as the set of all approximate homomorphism triples:*

$$\mathbb{H}_{n+1} = \{(\pi, \sigma, \gamma) \mid \pi : M \to \mathbb{H}_n,\ \sigma : \mathbb{H}_n \to \mathbb{H}_n,\ \gamma : \mathbb{H}_n \to W\}$$

   where π, σ, γ are ϵ-approximate homomorphisms (Definition F.1), with corresponding error parameters $\epsilon_\pi, \epsilon_\sigma, \epsilon_\gamma$.

3. ***Algebraic structure**: Each $\mathbb{H}_n$ is equipped with a τ-type algebraic structure:*
   - *Carrier set: $\mathbb{H}_n$ itself (as a set of triples).*
   - *Operations: For any operation $\omega \in \Omega$ (arity k), define its action on $\mathbb{H}_n$ as:*

$$\omega_{\mathbb{H}_n}(h_1, \ldots, h_k) = (\pi_\omega, \sigma_\omega, \gamma_\omega)$$

   where $h_i = (\pi_i, \sigma_i, \gamma_i) \in \mathbb{H}_n$, and the new maps $\pi_\omega : M \to \mathbb{H}_n$, $\sigma_\omega : \mathbb{H}_n \to \mathbb{H}_n$, $\gamma_\omega : \mathbb{H}_n \to W$ are defined componentwise recursively as follows:

$$\begin{aligned}
\pi_\omega(m) &\coloneqq \omega_{\mathbb{H}_n}(\pi_1(m), \ldots, \pi_k(m)), \quad \forall m \in M, \\
\sigma_\omega(h) &\coloneqq \omega_{\mathbb{H}_n}(\sigma_1(h), \ldots, \sigma_k(h)), \quad \forall h \in \mathbb{H}_n, \\
\gamma_\omega(h) &\coloneqq \omega_W(\gamma_1(h), \ldots, \gamma_k(h)), \quad \forall h \in \mathbb{H}_n.
\end{aligned}$$

   Here, the $\omega_{\mathbb{H}_n}$ and ω_W on the right denote the realization of operation ω on $\mathbb{H}_n$ and W, respectively (if W lacks the corresponding operation, define it via approximate homomorphism induction, or regard it as defined on the image set). The above definition ensures inheritance of operations across levels.

- *Approximate homomorphism error control: There exists a constant $L_\omega \geq 0$ such that for any $h_1, \ldots, h_k \in \mathbb{H}_n$, the approximate homomorphism errors of the new maps satisfy:*

$$d_{\mathbb{H}_n}(\pi_\omega(m), \omega_{\mathbb{H}_n}(\pi_1(m), \ldots, \pi_k(m))) \leq L_\omega \cdot \max_i \epsilon_{\pi_i},$$
$$d_{\mathbb{H}_n}(\sigma_\omega(h), \omega_{\mathbb{H}_n}(\sigma_1(h), \ldots, \sigma_k(h))) \leq L_\omega \cdot \max_i \epsilon_{\sigma_i},$$
$$d_W(\gamma_\omega(h), \omega_W(\gamma_1(h), \ldots, \gamma_k(h))) \leq L_\omega \cdot \max_i \epsilon_{\gamma_i},$$

where $\epsilon_{\pi_i}, \epsilon_{\sigma_i}, \epsilon_{\gamma_i}$ are the approximate homomorphism error parameters for $\pi_i, \sigma_i, \gamma_i$ respectively (Definition 8.1), and $d_{\mathbb{H}_n}$ and d_W are the respective pseudometrics.

4. ***Finiteness/infiniteness**:*
   - *If $N < \infty$, it is a finite cognitive system (e.g., ordinary machine learning models).*
   - *If $N = \infty$, it is an infinite cognitive system (with infinite recursive cognitive capacity).*
5. ***Convergence condition (optional)**: For infinite systems, an additional convergence condition may be imposed:*

$$\lim_{n\to\infty} \epsilon_n = 0, \quad \epsilon_n = \sup_{h\in\mathbb{H}_n} \|(\pi \circ \sigma)(h) - h\|$$

   But this is not mandatory for the general principle. When not satisfied, the system is an ordinary machine learning mechanism with finite cognitive capacity.

Definition F.3 (State Universe Sheaf (Independent Version))**.** *Let $\{\mathbb{H}_n\}_{n=0}^N$ be a generalized cognitive hierarchy sequence (Definition F.2). Define the **state universe sheaf** $\mathscr{F}_\Omega$ as follows:*

1. ***Base space**: $\Omega = \{0, 1, \ldots, N\}$ (if $N < \infty$) or $\Omega = \mathbb{N}$ (if $N = \infty$), equipped with the finite complement topology.*
2. ***Sheaf objects**: For an open set $U \subseteq \Omega$, define $\mathscr{F}_\Omega(U)$ as the set of all sections $s : U \to \bigcup_{n\in U} \mathbb{H}_n$ such that $s(n) \in \mathbb{H}_n$ and sections are compatible.*
3. ***Restriction maps**: For $V \subseteq U$, $\rho_{UV} : \mathscr{F}_\Omega(U) \to \mathscr{F}_\Omega(V)$ is the restriction of sections.*

Definition F.4 (Generalized Cognitive System)**.** *A **generalized cognitive system** is a quadruple:*

$$\mathfrak{G} = (\{\mathbb{H}_n\}_{n=0}^N, \mathscr{F}_\Omega, \Phi, \mathcal{C})$$

where:

1. *$\{\mathbb{H}_n\}_{n=0}^N$ is a generalized cognitive hierarchy sequence (DefinitionF.2).*
2. *$\mathscr{F}_\Omega$ is the state universe sheaf (DefinitionF.3).*
3. *$\Phi : \mathscr{F}_\Omega \to \mathscr{F}_\Omega$ is the endogenous meta-dynamics (a contraction map).*
4. *$\mathcal{C}$ is a set of constraints (may include structural, performance, resource constraints, etc.).*

When $N = \infty$ and the convergence condition is satisfied, the system has infinite cognitive capacity; otherwise it is a finite cognitive system.

The three-layer structure of the target space $\mathfrak{S} = (\mathcal{M}, \rho, \varphi)$ naturally supports machine learning because:

1. $\mathcal{B}_4^k$ is Turing-complete $\to$ can compute any function
2. Parameter fibration allows gradient propagation $\to$ supports gradient-based learning
3. State space is the limit of cognitive layers $\to$ supports multi-scale representation

Remark F.1. *The recursive construction of DefinitionF.2 automatically yields hierarchical cognitive representations:*

- *$\mathbb{H}_0 = \mathcal{M}$: raw data or sensory input.*

Table 3: Mathematical Abstraction of Three Major Cognitive Motivations

Level	Mathematical Structure	Automatically Generated Machine Learning Capabilities
$\mathcal{M}$ (algebraic carrier)	$\mathcal{B}_4^k$ (Boolean algebra)	Automatically supports symbolic reasoning, logical operations
$\rho : \mathcal{M} \to \mathcal{P}$	Parameter fibration	Automatically yields parameterization, optimizability
$\varphi : \mathcal{P} \to \mathcal{S}$	State space mapping	Automatically yields representation learning, state evolution

- $\mathbb{H}_1$*: cognitive representation of* $\mathcal{M}$ *(first-order abstraction).*
- $\mathbb{H}_2$*: cognitive representation of* $\mathbb{H}_1$ *(second-order abstraction, meta-cognition).*
- $\mathbb{H}_n$*: nth-order abstraction, forming a cognitive tower.*

This construction does not presuppose specific machine learning structures; instead, representation spaces, learning algorithms, and prediction mechanisms are automatically generated through minimal recursion.

Theorem F.1 (Automatic Generation Property of Cognitive Hierarchy)**.** *Let* $\{\mathbb{H}_n\}_{n=0}^{\infty}$ *be a generalized cognitive hierarchy sequence satisfying DefinitionF.2. Then:*

1. ***Automatic generation of representation spaces****: Each* $\mathbb{H}_n$ *is a cognitive representation space with hierarchical structure.*
2. ***Automatic generation of learning algorithms****: The transition map* $T_n : \mathbb{H}_n \to \mathbb{H}_{n+1}$ *defines a meta-learning process.*
3. ***Automatic generation of prediction-correction cycles****: The mapping pair* (π, σ) *forms a prediction-correction cycle:* π *extracts features from the lower level,* σ *evolves within the representation space.*
4. ***Universal approximation property****: For any continuous function* $f : X \to Y$ *(where* $X \subseteq \mathcal{M}$*,* $Y \subseteq \mathcal{W}$*), there exists* $n \in \mathbb{N}$ *and a map* $\tilde{f} : \mathbb{H}_n \to \mathbb{H}_n$ *such that* $\gamma \circ \tilde{f} \circ \pi$ *approximates* f *to arbitrary precision.*

Proof. (1)-(3) follow directly from the definition.

(4) Proof sketch: Using the recursive construction of the cognitive hierarchy, we can build a representation deep enough so that the composite mapping can approximate any continuous function. Specifically, we can decompose f into a composition of simple functions, then implement these simple functions at each level, and finally combine them. Since the algebraic structures on $\mathcal{M}$ and $\mathcal{W}$ are rich enough (e.g., the Boolean algebra $\mathcal{B}_4^k$ is Turing-complete) and approximate homomorphisms allow arbitrarily small errors, arbitrary precision approximation can be achieved. □

Theorem F.2 (Automatically Generated Machine Learning Components)**.** *Let* $\{\mathbb{H}_n\}_{n=0}^{\infty}$ *be a generalized cognitive hierarchy sequence satisfying the convergence condition* $\lim_{n\to\infty} \epsilon_n = 0$*. Then the following machine learning components are automatically generated:*

1. ***Representation learning****: The map* $\pi_n : \mathcal{M} \to \mathbb{H}_n$ *provides encoding from raw data to the nth-level representation, automatically achieving feature extraction and representation learning.*
2. ***Optimization algorithm****: The contraction property of the endogenous dynamics* Φ *guarantees convergence of iterative updates; the most natural implementation is some variant of gradient descent.*
3. ***Prediction mechanism****: The mapping pair* (π, σ) *forms a prediction-correction cycle: given input* $m \in \mathcal{M}$*, the system computes* $\pi(m)$ *as internal representation, then applies* σ *for state evolution, and finally outputs prediction via* γ*.*

4. ***Meta-learning**: The transition from $\mathbb{H}_n$ to $\mathbb{H}_{n+1}$ is a meta-learning process, learning how to learn better.*

Proof. (1) By definition, π_n is an approximate homomorphism that approximately preserves the structure of $\mathcal{M}$ into $\mathbb{H}_n$, which is exactly the mathematical essence of representation learning.

(2) The fixed-point theorem for contraction mappings guarantees that the iteration $h_{t+1} = \Phi(h_t)$ converges to a unique fixed point. In a differentiable setting, the most natural contraction map is the discretization of gradient flow $h_{t+1} = h_t - \eta \nabla L(h_t)$, where L is some loss function.

(3) The prediction-correction cycle consists of:

$$\text{Prediction} : \hat{h} = \sigma(h_{\text{current}})$$
$$\text{Correction} : h_{\text{new}} = \pi(m_{\text{observed}})$$

where m_{observed} is a new observation. This structure is identical to prediction-correction algorithms such as the Kalman filter.

(4) Meta-learning is reflected in: elements (π, σ, γ) of $\mathbb{H}_{n+1}$ are themselves functions on $\mathbb{H}_n$, i.e., learning how to optimize representations in $\mathbb{H}_n$. □

Table 4: Machine Learning Components Automatically Generated by Cognitive Hierarchy

Cognitive Level	**Mathematical Structure**	**Corresponding Machine Learning Component**
$\mathbb{H}_0 = \mathcal{M}$	Basic carrier set	Raw data space
$\mathbb{H}_1 = \{(\pi, \sigma, \gamma)\}$	Approximate homomorphisms from $\mathcal{M}$ to $\mathbb{H}_0$	Representation learning, feature extraction
$\mathbb{H}_2 = \{(\pi, \sigma, \gamma)\}$	Approximate homomorphisms from $\mathcal{M}$ to $\mathbb{H}_1$	Meta-learning, optimization algorithm
$\mathbb{H}_n$	Recursively defined	nth-level abstract representation
Limit $\mathbb{H}_\infty$	Inverse limit $\varprojlim \mathbb{H}_n$	Complete cognitive representation space

Corollary F.1 (Natural Emergence of Gradient Descent)**.** *In a differentiable setting, the simplest implementation of endogenous dynamics Φ is gradient descent:*

$$\Phi(h) = h - \eta \nabla \mathcal{L}(h),$$

where $\mathcal{L} : \mathbb{H}_n \to \mathbb{R}$ is a cognitive loss function. This is jointly determined by the contraction requirement and simplicity (Occam's razor).

Proof. Contraction mapping requires $\|\Phi(h) - \Phi(h')\| \leq \kappa \|h - h'\|$. For differentiable Φ, this is equivalent to $\|\nabla \Phi\| \leq \kappa < 1$. In the linear approximation $\Phi(h) \approx h + \eta F(h)$, the simplest F satisfying contraction is the negative gradient direction, because the gradient direction is the steepest descent direction. Combined with the least action principle (next section), this becomes the natural choice. □

F.2 Minimum Cognitive Action Principle: Dynamical Formulation

Definition F.5 (Cognitive Trajectory)**.** *Let $\mathfrak{G}$ be a generalized cognitive system (DefinitionF.4). A **cognitive trajectory** is a map $\gamma : [0, T] \to \mathscr{F}_\Omega(\Omega)$ satisfying:*

1. *$\gamma(t)$ is a global section*
2. *γ is continuous in the sheaf topology*
3. *γ is compatible with dynamics: there exists $\epsilon(t) \to 0$ such that*

$$d(\gamma(t + \Delta t), \Phi_{\Delta t}(\gamma(t))) \leq \epsilon(t) \Delta t$$

The set of all trajectories is denoted $\mathcal{T}(\mathfrak{G})$.

Definition F.6 (Cognitive Action Functional)**.** *For a trajectory $\Gamma \in \mathcal{T}(\mathfrak{G})$, define its **cognitive action**:*

$$\mathcal{A}[\Gamma] = \int_0^T \mathcal{L}(\Gamma(t), \dot{\Gamma}(t), t)dt$$

where the Lagrangian density $\mathcal{L}$ decomposes as:

$$\mathcal{L}(\Gamma, \dot{\Gamma}, t) = \lambda_1 \cdot \|\Gamma(t) - \Phi(\Gamma(t))\|^2 + \lambda_2 \cdot C(\Gamma(t)) + \lambda_3 \cdot \|\dot{\Gamma}(t)\|^2 + \lambda_4 \cdot R(\Gamma(t))$$

Here:

- $\|\Gamma(t) - \Phi(\Gamma(t))\|^2$ *is the prediction error term (cognitive dynamics Φ given by Definition F.4)*
- $C(\Gamma(t))$ *is the description complexity of cognitive state $\Gamma(t)$*
- $\|\dot{\Gamma}(t)\|^2$ *is the rate-of-change penalty term*
- $R(\Gamma(t))$ *is the trajectory curvature term*
- $\lambda_1, \lambda_2, \lambda_3, \lambda_4 \geq 0$ *are weight coefficients*

Axiom F.1 (Minimum Cognitive Action Principle)**.** ***Optimal cognitive evolution*** *follows the principle of least action:*

$$\gamma^* = \arg\min_{\gamma \in \mathcal{T}(\mathfrak{G})} \mathcal{A}[\gamma]$$

with boundary conditions $\gamma(0) = \gamma_0$ (initial state) and $\gamma(T) \in \mathcal{A}$ (target basin of attraction).

Remark F.2 (Special Form of Cognitive Euler-Lagrange Equation)**.** *In simplified cases (ignoring higher-order derivatives and curvature terms), the minimum cognitive action principle reduces to the classical Euler-Lagrange equation:*

$$\frac{\partial \mathcal{L}}{\partial \Gamma} - \frac{d}{dt}\left(\frac{\partial \mathcal{L}}{\partial \dot{\Gamma}}\right) = 0.$$

This is a special case of TheoremF.3 under flat space and first-derivative assumptions.

Definition F.7 (Specific Components of Lagrangian Density)**.** *In the minimum cognitive action principle, the Lagrangian density $\mathcal{L}$ can be decomposed into the following components:*

1. ***Prediction error term**:*

$$\mathcal{L}_{pred}(S, \Phi) = \frac{1}{2}\|S - \Phi(S)\|^2$$

2. ***Cognitive complexity term**:*

$$\mathcal{L}_{comp}(S) = K(S) + \lambda \cdot \dim_{\mathit{eff}}(S)$$

   where $K(S)$ is the description complexity of cognitive state S (e.g., Kolmogorov complexity)

3. ***Dynamical smoothness term**:*

$$\mathcal{L}_{dyn}(\nabla S) = \frac{1}{2}\|\nabla S\|^2$$

4. ***Curvature regularization term**:*

$$\mathcal{L}_{curv}(R_S) = \frac{1}{2}\|R_S\|^2$$

   where R_S is the curvature tensor of the cognitive state space

Total Lagrangian density is the weighted sum: $\mathcal{L} = \sum_i w_i \mathcal{L}_i$.

F.3 Deriving AGI Basic Structure from Minimum Cognitive Action Principle

Theorem F.3 (Cognitive Euler-Lagrange Equation (Complete Form))**.** *Let the cognitive action $\mathcal{A}[S]$ be given by DefinitionF.6, where the Lagrangian density $\mathcal{L}$ depends on cognitive state S, its first covariant derivative $\nabla_\mu S$, and second covariant derivative $\nabla_\mu \nabla_\nu S$. Then an extremal section S^* satisfies the following* ***complete cognitive Euler-Lagrange equation****:*

$$\frac{\partial \mathcal{L}}{\partial S} - \nabla_\mu \left(\frac{\partial \mathcal{L}}{\partial (\nabla_\mu S)} \right) + \nabla_\mu \nabla_\nu \left(\frac{\partial \mathcal{L}}{\partial (\nabla_\mu \nabla_\nu S)} \right) = 0,$$

where ∇_μ is the covariant derivative on the state space.

Proof. This is a standard result in higher-order Lagrangian field theory. Consider a variation δS; the first variation of the action is:

$$\begin{aligned} \delta \mathcal{A}[S] &= \int \left[\frac{\partial \mathcal{L}}{\partial S} \delta S + \frac{\partial \mathcal{L}}{\partial (\nabla_\mu S)} \nabla_\mu (\delta S) + \frac{\partial \mathcal{L}}{\partial (\nabla_\mu \nabla_\nu S)} \nabla_\mu \nabla_\nu (\delta S) \right] dV \\ &= \int \left[\frac{\partial \mathcal{L}}{\partial S} - \nabla_\mu \left(\frac{\partial \mathcal{L}}{\partial (\nabla_\mu S)} \right) + \nabla_\mu \nabla_\nu \left(\frac{\partial \mathcal{L}}{\partial (\nabla_\mu \nabla_\nu S)} \right) \right] \delta S \, dV + \text{boundary terms} \end{aligned}$$

By the fundamental lemma of the calculus of variations, the expression inside brackets must vanish for all variations δS. □

Theorem F.4 (Necessity of Semantic-Dynamics Isomorphism)**.** *Under the minimum cognitive action principle, the optimal learning section S^* automatically satisfies:*

1. ***Representation-evolution duality****: There exists a diffeomorphism $\varphi : \mathcal{F} \to T^*\mathcal{F}$ (cotangent bundle) on the fiber $\mathcal{F}$, such that the semantic representation space is isomorphic to the dynamical phase space:*

$$(\textit{semantic representation}) \quad \mathcal{F} \cong (\textit{dynamical phase space}) \quad T^*\mathcal{F}$$

2. ***Hamiltonian structure****: Cognitive evolution is described by a Hamiltonian system:*

$$\frac{dS}{dt} = \{S, H\}_{PB}$$

   where $H : \mathcal{F} \to \mathbb{R}$ is the cognitive Hamiltonian, $\{\cdot, \cdot\}_{PB}$ is the Poisson bracket.

3. ***Duality proof****: From the Euler-Lagrange equation, the momentum variable $p = \frac{\partial \mathcal{L}}{\partial \dot{S}}$ is related to the position variable S via Legendre transformation. Since $\mathcal{L}$ is convex (guaranteed by the cognitive metric), the transformation is invertible, hence semantic space and phase space are isomorphic.*

Proof. **Proof sketch**: The cognitive Lagrangian density $\mathcal{L}$ is strictly convex in ∇S (guaranteed by the positive definite metric g), hence the Legendre transformation:

$$p = \frac{\partial \mathcal{L}}{\partial (\nabla S)}$$

is a diffeomorphism. Define the Hamiltonian:

$$H(S, p) = p \cdot \nabla S - \mathcal{L}(S, \nabla S)$$

Then the Euler-Lagrange equation is equivalent to Hamilton's equations:

$$\nabla S = \frac{\partial H}{\partial p}, \quad \nabla p = -\frac{\partial H}{\partial S}$$

This establishes an isomorphism between the semantic representation space $\mathcal{F}$ (space of S) and the dynamical phase space $T^*\mathcal{F}$ (space of (S, p)). □

Theorem F.5 (Necessity of Training-Inference Adjointness)**.** *Under the minimum cognitive action framework, the training process and inference process are mutually adjoint:*

1. ***Time-reversal symmetry****: The cognitive action $\mathcal{A}$ is invariant under time reversal $t \mapsto -t$, modulo boundary terms.*

2. ***Adjoint dynamics****: Training dynamics Φ_{train} and inference dynamics Φ_{inf} satisfy:*

$$\Phi_{inf} = \Phi_{train}^{\dagger}$$

where $\dagger$ denotes the adjoint operator under the cognitive metric g.

3. ***Symmetry of optimal trajectory****: The optimal learning section S^* satisfies:*

$$S^*(t_f - t) = \Theta \circ S^*(t)$$

where $\Theta : \mathcal{F} \to \mathcal{F}$ is an antilinear involution on the fiber (cognitive time-reversal operator).

Proof. **Proof sketch**: The transformation of the cognitive Lagrangian density $\mathcal{L}$ under time reversal $t \mapsto -t$:

$$\mathcal{L}(S, \nabla S) \mapsto \mathcal{L}(S, -\nabla S)$$

Since $\mathcal{L}_{\text{dyn}} = \frac{1}{2}\|\nabla S\|^2$ is an even function of ∇S, and $\mathcal{L}_{\text{task}}$ and $\mathcal{L}_{\text{comp}}$ do not depend on the sign of ∇S, the main change comes from boundary terms. Therefore the action $\mathcal{A}$ is invariant under time reversal (boundary terms do not affect extremal conditions).

By Noether's theorem, this leads to a conserved quantity—the cognitive information flow $J^\mu = \frac{\partial \mathcal{L}}{\partial(\nabla_\mu S)} \nabla^\nu S \delta^\mu_\nu - \mathcal{L}\delta^\mu_0$. Training and inference, as processes with opposite time directions, have generators that are mutually adjoint. □

Theorem F.6 (Necessity of Algorithm-Equation Unique Determinacy)**.** *Let S^* be an extremal section of the minimum cognitive action. Then:*

1. ***Discrete-continuous equivalence****: The corresponding discrete learning algorithm $\mathcal{A}_{disc}$ and continuous dynamical equation $\mathcal{E}_{cont}$ describe the same geometric object:*

$$\mathcal{A}_{disc} \quad \textit{corresponds to} \quad \textit{geodesic discretization of } S^*$$

$$\mathcal{E}_{cont} \quad \textit{corresponds to} \quad \nabla_{\dot{S}}\dot{S} = 0 \textit{ (geodesic equation)}$$

2. ***Uniqueness****: Under given boundary conditions, the extremal section S^* is unique (modulo discrete symmetries).*

3. ***Algorithm as exponential map****: The discrete learning algorithm is a numerical implementation of the exponential map of the connection ∇:*

$$S_{n+1} = \exp_{S_n}(-\eta \nabla \mathcal{L}(S_n))$$

where $\exp$ is the exponential map on the cognitive manifold.

Proof. **Proof sketch**: Extremal section S^* satisfies the geodesic equation $\nabla_{\dot{S}}\dot{S} = 0$ (when $\mathcal{L}_{\text{dyn}}$ dominates). Discretizing this equation naturally leads to gradient-descent type algorithms.

Let the discrete time step be ϵ; the discrete approximation of the geodesic equation is:

$$\frac{S_{n+1} - 2S_n + S_{n-1}}{\epsilon^2} + \Gamma(S_n)\left(\frac{S_{n+1} - S_{n-1}}{2\epsilon}\right)^2 = 0$$

where Γ are Christoffel symbols. Rearranging gives:

$$S_{n+1} = S_n - \epsilon \dot{S}_n - \frac{\epsilon^2}{2}\Gamma(S_n)\dot{S}_n^2 + O(\epsilon^3)$$

This is exactly gradient descent with curvature correction. As $\epsilon \to 0$, it converges to the continuous equation. □

Theorem F.7 (Necessity of MetaPred as Self-Consistency Condition)**.** *Under the minimum cognitive action framework, the self-referential prediction module* $\mathrm{MetaPred}$ *naturally appears as a* ***self-consistency condition****:*

1. ***Self-consistency equation**: The optimal section S^* must satisfy:*

$$S^*(t) = \mathrm{MetaPred}(S^*(t)) \quad \forall t \in \mathcal{T}$$

where $\mathrm{MetaPred} : \mathcal{F} \to \mathcal{F}$ *is an operator predicting its own evolution.*

2. ***Variational derivation**:* $\mathrm{MetaPred}$ *is a direct consequence of the Euler-Lagrange equation for the curvature term* $\mathcal{L}_{meta}$ *in the action extremal condition:*

$$\mathrm{MetaPred}(S) = S - \frac{1}{\delta}\nabla^\mu \nabla^\nu \left(\frac{\partial \mathcal{L}_{meta}}{\partial R^\nabla_{\mu\nu}} \right)$$

3. ***Fixed-point form**:* $\mathrm{MetaPred}$ *can be expressed as a fixed-point combinator:*

$$\mathrm{MetaPred}(S) = \mathcal{Y}(F_S)$$

where $\mathcal{Y}$ *is an approximate fixed-point combinator,* $F_S : \mathcal{F} \to \mathcal{F}$ *is the flow generated by the cognitive Hamiltonian.*

Proof. **Proof sketch**: Consider the meta-cognitive term $\mathcal{L}_{\text{meta}} = \frac{1}{2}\|R^\nabla_S\|^2$ in the action. The curvature R^∇_S measures the non-commutativity of parallel transport, i.e., the inconsistency between the system's prediction of its own future state and the actual evolution.

The Euler-Lagrange equation includes the term:

$$\delta \mathcal{L}_{\text{meta}}/\delta S = \nabla_\mu \nabla_\nu \left(\frac{\partial \mathcal{L}_{\text{meta}}}{\partial R^\nabla_{\mu\nu}} \right)$$

This enforces a self-consistency condition: the system's prediction of its own evolution must agree with actual evolution.

Define $\mathrm{MetaPred}(S)$ as the adjustment that minimizes the curvature term:

$$\mathrm{MetaPred}(S) = \arg\min_{S'} \left[\frac{1}{2}\|S' - S\|^2 + \frac{\delta}{2}\|R^\nabla_{S'}\|^2 \right]$$

The first-order condition gives $\mathrm{MetaPred}(S) = S - \delta \nabla^\mu \nabla^\nu (\partial \mathcal{L}_{\text{meta}}/\partial R^\nabla_{\mu\nu})$. □

Theorem F.8 (Necessity of $\mathcal{B}_4^k$ as Minimal Complexity Representation)**.** *Assume the fiber* $\mathcal{F}$ *of the cognitive fiber bundle has a finite-dimensional approximation* $\mathcal{F}_{fin}$*. Then the minimum cognitive action principle forces:*

1. ***Finite representation theorem**: There exists a finite set* $\mathcal{M}$ *and approximate homomorphisms* $\rho : \mathcal{F} \to \mathcal{M}$, $\sigma : \mathcal{M} \to \mathcal{F}$, *such that* $\rho \circ \sigma \approx \mathrm{id}_{\mathcal{M}}$.

2. ***Algebraic structure minimization**: The algebraic structure of* $\mathcal{M}$ *has minimal Kolmogorov complexity among all structures that are functionally complete.*

3. ***Uniqueness of** $\mathcal{B}_4^k$: The minimal complexity algebra must be* $\mathcal{M} \cong \mathcal{B}_4^k$, *where* $k = \lceil \frac{1}{2}\log_2(\dim \mathcal{F}_{fin}) \rceil$.

Proof. **Proof sketch**: The complexity term $\mathcal{L}_{\text{comp}}(S) = K(S) + \lambda \dim_{\text{eff}}(S)$ in the action forces selection of the simplest effective representation.

Consider all possible functionally complete algebraic structures (supporting basic cognitive operations like negation, conjunction, disjunction). By TheoremC.1, $\mathcal{B}_4$ is the smallest Boolean algebra that is functionally complete.

For a finite approximation of dimension $\dim \mathcal{F}_{\text{fin}} = N$, we need $\log_4 N$ copies of $\mathcal{B}_4$ to encode all states, hence $\mathcal{M} \cong \mathcal{B}_4^k$, $k = \lceil \log_4 N \rceil$.

Complexity comparison: any other algebraic structure $\mathcal{A}$ that is functionally complete satisfies $K(\mathcal{A}) \geq K(\mathcal{B}_4^k) + O(1)$, therefore $\mathcal{B}_4^k$ minimizes $\mathcal{L}_{\text{comp}}$. □

Theorem F.9 (Variational Form of Unified Learning Algorithm)**.** *The discrete implementation of the minimum cognitive action yields a unified AGI learning algorithm:*

Algorithm 16: Variational Meta-Learning Algorithm (VarMeta)

Input : *Initial cognitive state $S_0 \in \mathcal{F}$, task distribution $p(\mathcal{T})$, learning rates $\{\eta_t\}$*
Output : *Optimal cognitive state S^**
begin
  Initialize: $S \leftarrow S_0$
  for $t = 0, 1, 2, \ldots, T-1$ **do**
    Sample task $t \sim p(\mathcal{T})$
    Compute cognitive gradient: $G \leftarrow \nabla_S \mathcal{L}(S, \nabla S, R_S^\nabla)$
    Compute meta-prediction: $S_{meta} \leftarrow \mathrm{MetaPred}(S)$
    Combined update: $S \leftarrow S - \eta_t[G + \gamma(S - S_{meta})]$
    Update connection: $\nabla \leftarrow \mathit{ParallelTransport}(\nabla, S, S_{new})$
  return S^*

This algorithm converges to the least action section S^*, with convergence rate $O(1/t^2)$.

Theorem F.10 (Equivalence with AGI Foundational Axiom System)**.** *The variational framework of this section is equivalent to AxiomsA.1 andA.2, and provides a more foundational principle:*

1. ***Axiom derivation****: All previous axioms can be derived from the minimum cognitive action principle:*
   - *Dynamic operation preservation (approximate homomorphism) comes from local invariance of $\mathcal{L}_{dyn}$ in the action.*
   - *Bidirectional interaction consistency comes from time-reversal symmetry.*
   - *Infinite recursivity comes from the self-consistency condition of the curvature term $\mathcal{L}_{meta}$.*
2. ***Theorem correspondence****:*
   - *TheoremB.1 corresponds to TheoremsF.4 andF.5.*
   - *TheoremB.3 corresponds to TheoremF.7.*
   - *TheoremC.1 corresponds to TheoremF.8.*
   - *TheoremD.29 corresponds to TheoremF.6.*
3. ***Algorithm unification****: Algorithm16 unifies all machine learning algorithms in this paper (Algorithms3, 8, etc.).*

Proof. **Proof sketch**: Construct explicit transformations mapping variational framework objects to the objects in this paper:

- Cognitive fiber bundle $\mathcal{C}$ corresponds to the state sheaf $\mathscr{F}_\Omega$.
- Learning section $S : \mathcal{T} \to \mathcal{C}$ corresponds to the meta-dynamics sheaf paradigm $\mathfrak{M}$.
- Cognitive action $\mathcal{A}[S]$ corresponds to loss function plus regularization terms.
- Cognitive connection ∇ corresponds to the lifting functor in generalized adjoint dynamics.
- Curvature R^∇ corresponds to meta-cognitive inconsistency measure.

Under this correspondence, the least action principle is equivalent to the minimality axiomA.2. □

F.4 Geometric Formulation of Minimum Cognitive Action Principle

Definition F.8 (Cognitive State Manifold)**.** *Let $\mathfrak{G} = (\{\mathbb{H}_n\}, \mathscr{F}_\Omega, \Phi, \mathcal{C})$ be a generalized cognitive system (DefinitionF.4). Define its* ***cognitive state manifold*** *$\mathcal{M}$ as:*

1. ***Construction***: $\mathcal{M} = \varinjlim_{\epsilon \to 0} \mathcal{M}_\epsilon$, *where* $\mathcal{M}_\epsilon$ *is the* ϵ*-approximate cognitive state space*
2. ***Atlas***: *Local coordinates are induced by cognitive hierarchy: for any* $n \in \mathbb{N}$*, there exists an embedding* $\iota_n : \mathbb{H}_n \hookrightarrow \mathcal{M}$
3. ***Dimension***: $\dim \mathcal{M} = \sup_n \dim(\mathbb{H}_n)$*, possibly infinite-dimensional*
4. ***Measurable structure***: *Equipped with the Borel* σ*-algebra induced by the cognitive pseudometric* d

Definition F.9 (Geometric Lift of Cognitive Pseudometric)**.** *The cognitive pseudometric* $d : \mathcal{M} \times \mathcal{M} \to \mathbb{R}_{\geq 0}$ *(DefinitionA.1) induces a Riemannian (or pseudo-Riemannian) metric* g*:*

$$g_p(v, w) = \lim_{\epsilon \to 0} \frac{d(\exp_p(\epsilon v), \exp_p(\epsilon w))^2 - d(\exp_p(\epsilon v), p)^2 - d(p, \exp_p(\epsilon w))^2}{2\epsilon^2}$$

where $\exp_p : T_p\mathcal{M} \to \mathcal{M}$ *is the exponential map, defined as* $\exp_p(v) = \gamma(1)$*, with* γ *satisfying* $\nabla_{\dot\gamma}\dot\gamma = 0$*,* $\gamma(0) = p$*,* $\dot\gamma(0) = v$*.*

Theorem F.11 (Metric Existence and Uniqueness)**.** *Assume the cognitive pseudometric* d *satisfies:*

1. ***Local approximate Euclidean property***: $\forall p \in \mathcal{M}$*, there exists a neighborhood* U *such that* $d(x, y) = \|x - y\| + O(\|x - y\|^3)$
2. ***Symmetry***: $d(x, y) = d(y, x)$
3. ***Triangle inequality***: $d(x, z) \leq d(x, y) + d(y, z)$
4. ***Differentiability***: d *is* C^2 *continuous on* $\mathcal{M} \times \mathcal{M}$

Then there exists a unique smooth symmetric $(0, 2)$*-tensor field* g *satisfying DefinitionF.9, and:*

1. g *is positive definite if and only if* d *is a strictly positive definite metric*
2. *A cognitive trajectory* γ *is a geodesic if and only if* γ *is locally shortest under the cognitive pseudometric* d
3. *The curvature tensor* R *is completely determined by the Christoffel symbols of* g

Proof. By generalization of the Myers-Steenrod theorem to pseudometric case. Local approximate Euclidean property ensures the limit exists, differentiability ensures g is smooth. Uniqueness follows from the limit definition. □

Definition F.10 (Cognitive Content Tensor Field)**.** *Let* $\mathcal{M}$ *be the cognitive state manifold. A* ***cognitive content field*** *is an* (r, s)*-type tensor field* $\mathcal{C} \in \Gamma(T^r_s\mathcal{M})$ *satisfying:*

1. ***Semantic interpretation***: $\mathcal{C}(p)$ *represents the semantic content of state* p
2. ***Hierarchical structure***: *There exists a decomposition* $\mathcal{C} = \bigoplus_{n=0}^{\infty} \mathcal{C}^{(n)}$*, where* $\mathcal{C}^{(n)}$ *corresponds to the* n*th cognitive level*
3. ***Evolution equation***: $\frac{\mathcal{D}\mathcal{C}}{dt} = \mathcal{F}(\mathcal{C}, g, R)$*, where* $\mathcal{D}$ *is the covariant derivative,* $\mathcal{F}$ *is determined by the cognitive dynamics* Φ

Definition F.11 (Cognitive Geometric Action Functional)**.** *The* ***geometric action functional*** *of a cognitive system is defined as:*

$$\mathcal{S}_{geo}[g, \mathcal{C}] = \int_{\mathcal{M}} \mathcal{L}_{geo}(g, \mathcal{C}, \nabla\mathcal{C}, R)\, d\mathrm{vol}_g$$

where the Lagrangian density is:

$$\mathcal{L}_{geo} = \alpha R + \beta\|\nabla\mathcal{C}\|^2_g + \gamma V(\mathcal{C}) + \delta\|\mathcal{R}\|^2_g$$

Here:

- R *is the scalar curvature*

- $\|\nabla\mathcal{C}\|_g^2 = g^{ab}g^{cd}\cdots(\nabla_a\mathcal{C}_{c\ldots})(\nabla_b\mathcal{C}_{d\ldots})$ *is the rate of change of cognitive content*
- $V(\mathcal{C})$ *is the cognitive potential energy, measuring self-consistency of cognitive content*
- $\|\mathcal{R}\|_g^2$ *is the norm of the full curvature tensor, measuring distortion of cognitive structure*
- $\alpha, \beta, \gamma, \delta \in \mathbb{R}$ *are coupling constants*

Theorem F.12 (Cognitive Geometric Field Equations)**.** *Assume the cognitive system satisfies the action functional of DefinitionF.11. Via variations $\delta\mathcal{S}_{geo}/\delta g^{ab} = 0$ and $\delta\mathcal{S}_{geo}/\delta\mathcal{C} = 0$, we obtain the following field equations:*

1. ***Cognitive Einstein equation****:*

$$R_{ab} - \frac{1}{2}Rg_{ab} + \Lambda g_{ab} = \kappa T_{ab}^{(cog)}$$

where:

- $\Lambda = -\frac{\alpha}{2\delta}$ *is the cognitive cosmological constant*
- $\kappa = \frac{\beta}{2\alpha}$ *is the cognitive coupling constant*
- $T_{ab}^{(cog)} = -\frac{2}{\sqrt{-g}}\frac{\delta(\sqrt{-g}\mathcal{L}_{matter})}{\delta g^{ab}}$ *is the cognitive energy-momentum tensor*

2. ***Cognitive content evolution equation****:*

$$\Box_g\mathcal{C} - \frac{\partial V}{\partial\mathcal{C}} = \frac{\gamma}{\beta}\mathcal{Q}(\mathcal{C}, R)$$

where $\Box_g = g^{ab}\nabla_a\nabla_b$ is the covariant wave operator, $\mathcal{Q}$ is a curvature coupling term.

3. ***Cognitive self-consistency constraint****:*

$$\nabla^a T_{ab}^{(cog)} = 0$$

Proof. **Step 1: Metric variation**

Compute $\delta\mathcal{S}_{\text{geo}}/\delta g^{ab}$:

$$\begin{aligned}\delta\mathcal{S}_{\text{geo}} &= \int_{\mathcal{M}}\left[\frac{\partial\mathcal{L}_{\text{geo}}}{\partial g^{ab}}\delta g^{ab} + \frac{\partial\mathcal{L}_{\text{geo}}}{\partial R}\delta R + \frac{\partial\mathcal{L}_{\text{geo}}}{\partial(\nabla\mathcal{C})}\delta(\nabla\mathcal{C})\right]\sqrt{-g}d^nx \\ &= \int_{\mathcal{M}}\left[\alpha G_{ab} + \beta E_{ab}^{(\mathcal{C})} + \delta H_{ab} - \frac{1}{2}\mathcal{L}_{\text{geo}}g_{ab}\right]\delta g^{ab}\sqrt{-g}d^nx\end{aligned}$$

where:

- $G_{ab} = R_{ab} - \frac{1}{2}Rg_{ab}$ is the Einstein tensor
- $E_{ab}^{(\mathcal{C})}$ comes from variation of $\|\nabla\mathcal{C}\|^2$ and $V(\mathcal{C})$
- H_{ab} comes from variation of $\|\mathcal{R}\|^2$

Setting the variation to zero yields the Einstein equation form.

Step 2: Cognitive content variation

Compute $\delta\mathcal{S}_{\text{geo}}/\delta\mathcal{C}$:

$$\frac{\delta\mathcal{S}_{\text{geo}}}{\delta\mathcal{C}} = -2\beta\Box_g\mathcal{C} + \gamma\frac{\partial V}{\partial\mathcal{C}} + 2\delta\frac{\partial\|\mathcal{R}\|^2}{\partial\mathcal{C}} = 0$$

Step 3: Energy-momentum conservation

From Bianchi identity $\nabla^a G_{ab} = 0$ and the field equations, we get $\nabla^a T_{ab}^{(\text{cog})} = 0$. □

Theorem F.13 (Geometric Realization of Minimum Cognitive Action Principle)**.** *Let $\gamma : [0, T] \to \mathcal{M}$ be a cognitive trajectory with cognitive action $\mathcal{A}[\gamma] = \int_0^T \mathcal{L}(\gamma, \dot{\gamma})dt$. Then there exists an embedding $\iota : \mathcal{M} \to \mathcal{M}'$ into a higher-dimensional manifold such that:*

$$\mathcal{A}[\gamma] = \mathcal{S}_{geo}[\iota_* g, \iota_* \mathcal{C}]$$

i.e., the original action is the restriction of the geometric action to a specific submanifold.

Proof. Construct $\mathcal{M}' = \mathcal{M} \times \mathbb{R}$ (adding a time dimension). Define a metric on $\mathcal{M}'$:

$$g' = -dt^2 + g_{ij}(x, t)dx^i dx^j$$

and cognitive content field $\mathcal{C}'(x, t) = \mathcal{C}(\gamma(t))$. Then:

$$\mathcal{S}_{\text{geo}}[g', \mathcal{C}'] = \int_{\mathcal{M}'} \mathcal{L}_{\text{geo}} d\text{vol}_{g'} = \int_0^T \left(\int_{\mathcal{M}} \mathcal{L}_{\text{geo}} \sqrt{g} d^n x \right) dt$$

Choose $\mathcal{L}_{\text{geo}}$ such that on the trajectory γ, $\int_{\mathcal{M}} \mathcal{L}_{\text{geo}} \sqrt{g} d^n x = \mathcal{L}(\gamma, \dot{\gamma})$. □

Corollary F.2 (AGI as Geodesics in Cognitive Universe)**.** *The optimal AGI learning process S^* is a geodesic in the cognitive fiber bundle $\mathcal{C}$, which is the "straightest" and "most economical" cognitive trajectory in the sense of cognitive action. This provides a precise geometric interpretation of the "extreme triviality" characteristic of cognitive dynamics: the essence of intelligence is the shortest path in the cognitive manifold.*

Theorem F.14 (Necessity of Cognitive Einstein Equation)**.** *Under the minimum cognitive action framework, the optimal cognitive section S^* necessarily satisfies a cognitive version of Einstein's field equations:*

$$R^{cog}_{\mu\nu} - \frac{1}{2} R^{cog} g^{cog}_{\mu\nu} + \Lambda_{cog} g^{cog}_{\mu\nu} = \kappa_{cog} T^{cog}_{\mu\nu}$$

where:

- $R^{cog}_{\mu\nu}$*: cognitive Ricci curvature, measuring difficulty of knowledge transfer between tasks*
- Λ_{cog}*: cognitive cosmological constant, arising from inherent system complexity*
- κ_{cog}*: cognitive coupling constant, balancing geometric constraints and task demands*

Proof sketch. Consider total action $\mathcal{A}_{\text{total}} = \mathcal{A}_{\text{geom}}[g^{\text{cog}}] + \mathcal{A}_{\text{matter}}[S, g^{\text{cog}}]$, where:

$$\mathcal{A}_{\text{geom}} = \int \sqrt{-g^{\text{cog}}} (R^{\text{cog}} - 2\Lambda_{\text{cog}}) d^d t$$

$$\mathcal{A}_{\text{matter}} = \int \mathcal{L}(S, \nabla S, g^{\text{cog}}) \sqrt{-g^{\text{cog}}} d^d t$$

Variation with respect to the metric $g^{\text{cog}}_{\mu\nu}$ yields the above equation. □

Remark F.3 (Deep Analogy with General Relativity)**.** *This correspondence reveals deep structural similarity between cognitive systems and gravitational systems:*

- ***Matter tells spacetime how to curve $\leftrightarrow$ Tasks tell cognitive space how to structure***
- ***Objects move along geodesics $\leftrightarrow$ Intelligent agents evolve along minimum cognitive action trajectories***
- ***Black holes as gravitational singularities $\leftrightarrow$ Cognitive bottlenecks as theoretical limits***

Definition F.12 (Two Formulations of Minimum Cognitive Action Principle)**.** *Let $\mathcal{M} \in \mathbf{AGI}$ satisfy the AGI axiom system, $\mathfrak{M} = (\mathscr{F}_\Omega, \Phi, \text{Enc}_{\mathscr{F}}, \rho, \sigma_0)$ be its meta-dynamics sheaf paradigm.*

1. ***Dynamical formulation*** *(AxiomF.1): Based on action functional on cognitive trajectories $\Gamma : [0, T] \to \Gamma(\Omega, \mathscr{F}_\Omega)$:*

$$\mathcal{A}_{dyn}[\Gamma] = \int_0^T \mathcal{L}_{dyn} \left(\Gamma(t), \frac{d\Gamma}{dt}(t), t \right) dt$$

where $\mathcal{L}_{dyn} = \lambda_1\mathcal{L}_{pred} + \lambda_2\mathcal{L}_{comp} + \lambda_3\mathcal{L}_{dyn\text{-}smooth} + \lambda_4\mathcal{L}_{curv}$.

This formulation focuses on optimal paths of cognitive state evolution over time, analogous to the Lagrangian formulation of classical mechanics.

2. ***Geometric formulation*** *(DefinitionF.11): Based on geometric action functional on cognitive state manifold* $(\mathcal{M}_{cog}, g)$ *and cognitive content field* $\mathcal{C}$*:*

$$\mathcal{A}_{geo}[g, \mathcal{C}] = \int_{\mathcal{M}_{cog}} \mathcal{L}_{geo}(g, \mathcal{C}, \nabla\mathcal{C}, R) d\mathrm{vol}_g$$

where $\mathcal{L}_{geo} = \alpha R + \beta\|\nabla\mathcal{C}\|_g^2 + \gamma V(\mathcal{C}) + \delta\|\mathcal{R}\|_g^2$.

This formulation focuses on the overall geometric structure of cognitive state space, analogous to the Einstein-Hilbert action in general relativity.

Theorem F.15 (Equivalence of Dynamical and Geometric Formulations of Minimum Cognitive Action Principle)**.** *Let* $\mathcal{M} \in \mathbf{AGI}$ *satisfy AxiomA.1,* $\mathfrak{M}$ *be its meta-dynamics sheaf paradigm. Then the dynamical formulation* $\mathcal{A}_{dyn}$ *and geometric formulation* $\mathcal{A}_{geo}$ *of the minimum cognitive action principle are equivalent in the following strict sense:*

1. ***Mathematical construction equivalence****: There exists an embedding* $\iota : [0, T] \hookrightarrow \mathcal{M}_{cog}$ *into the cognitive state manifold, and a lift of fields* $\hat{\iota} : \Gamma([0, T], \mathscr{F}_\Omega) \to \Gamma(\mathcal{M}_{cog}, \mathscr{F})$*, such that:*

$$\mathcal{A}_{dyn}[\Gamma] = \mathcal{A}_{geo}[\hat{\iota}_* g, \hat{\iota}_* \mathcal{C}]$$

where $\hat{\iota}_* g$ *and* $\hat{\iota}_* \mathcal{C}$ *are the pullbacks of* g *and* $\mathcal{C}$ *via* $\hat{\iota}$*.*

2. ***Extremal condition equivalence****: A trajectory* Γ^* *is an extremum of the dynamical formulation (satisfying the cognitive Euler-Lagrange equation, TheoremF.3) if and only if there exist a metric* g^* *and field* $\mathcal{C}^*$ *that are extrema of the geometric formulation (satisfying the cognitive geometric field equations), and they satisfy compatibility conditions:*

$$g^*|_{\iota([0,T])} = \eta_{\mu\nu} d\iota^\mu \otimes d\iota^\nu, \quad \mathcal{C}^*|_{\iota([0,T])} = \Gamma^* \circ \iota^{-1}$$

where $\eta_{\mu\nu}$ *is the induced metric along the embedded trajectory.*

3. ***Lagrangian density correspondence****: The Lagrangian densities of the two formulations correspond as:*

$$\begin{aligned}
\mathcal{L}_{pred}(\Gamma, \dot{\Gamma}) &\longleftrightarrow \alpha R(g) \quad \textit{(prediction error} \leftrightarrow \textit{spatial curvature)} \\
\mathcal{L}_{comp}(\Gamma) &\longleftrightarrow \gamma V(\mathcal{C}) \quad \textit{(cognitive complexity} \leftrightarrow \textit{cognitive potential)} \\
\mathcal{L}_{dyn\text{-}smooth}(\dot{\Gamma}) &\longleftrightarrow \beta\|\nabla\mathcal{C}\|_g^2 \quad \textit{(dynamical smoothness} \leftrightarrow \textit{field gradient)} \\
\mathcal{L}_{curv}(\Gamma) &\longleftrightarrow \delta\|\mathcal{R}\|_g^2 \quad \textit{(trajectory curvature} \leftrightarrow \textit{full curvature)}
\end{aligned}$$

There exists a coefficient transformation $(\lambda_1, \lambda_2, \lambda_3, \lambda_4) \leftrightarrow (\alpha, \beta, \gamma, \delta)$ *such that the two are equal under appropriate restrictions.*

4. ***Optimal solution equivalence****: For any cognitive task* $\mathfrak{T}$*, the optimal solutions given by the two formulations are cognitively equivalent:*

$$\Gamma^* = \arg\min_\Gamma \mathcal{A}_{dyn}[\Gamma] \quad \Longleftrightarrow \quad (g^*, \mathcal{C}^*) = \arg\min_{g,\mathcal{C}} \mathcal{A}_{geo}[g, \mathcal{C}]$$

and there exists a cognitive isomorphism $\psi : \Gamma^*([0, T]) \to \mathcal{C}^*(\mathcal{M}_{cog})$ *such that the following diagram strictly commutes:*

$$\begin{array}{ccc}
[0, T] & \xrightarrow{\Gamma^*} & \Gamma(\Omega, \mathscr{F}_\Omega) \\
\downarrow{\scriptstyle \iota} & & \downarrow{\scriptstyle \psi} \\
\mathcal{M}_{cog} & \xrightarrow[\mathcal{C}^*]{} & \Gamma(\mathcal{M}_{cog}, \mathscr{F})
\end{array}$$

5. ***Physical description equivalence****: The two formulations correspond to different but equivalent descriptive levels of the same phenomenon in physics:*

- ***Dynamical formulation**: Corresponds to the Lagrangian formulation of classical mechanics, describing particle trajectories in time.*
- ***Geometric formulation**: Corresponds to the field-theoretic formulation of general relativity, describing the overall geometric structure of spacetime.*
- *Their relationship is similar to that between classical mechanics and general relativity: the former is the non-relativistic low-speed limit of the latter.*

Proof. We prove the above equivalence in five parts.

Part 1: Proof of Mathematical Construction Equivalence

Consider product manifold construction: Let $\mathcal{M}_{\text{cog}} = [0, T] \times \mathcal{N}$, where $\mathcal{N}$ is an n-dimensional compact manifold representing spatial dimensions of cognitive state. On $\mathcal{M}_{\text{cog}}$ define a Lorentzian signature metric:

$$g = -N(t,x)^2 dt^2 + h_{ij}(t,x)(dx^i + \beta^i dt)(dx^j + \beta^j dt)$$

where N is the lapse function, β^i the shift vector, h_{ij} the spatial metric.

Define cognitive content field $\mathcal{C} : \mathcal{M}_{\text{cog}} \to \mathbb{R}^m$ as:

$$\mathcal{C}(t,x) = \Gamma(t)(x)$$

i.e., interpreting the dynamical trajectory $\Gamma(t) \in \Gamma(\Omega, \mathscr{F}_\Omega)$ as the restriction of the geometric field on time slice t.

Compute volume element for geometric action $\mathcal{A}_{\text{geo}}$:

$$d\text{vol}_g = N\sqrt{h}\, dt d^n x, \quad h = \det(h_{ij})$$

Decompose Lagrangian density $\mathcal{L}_{\text{geo}}$ terms:

$$\begin{aligned} R(g) &= R^{(3)}(h) + K_{ij}K^{ij} - K^2 + \text{boundary terms} \\ \|\nabla\mathcal{C}\|_g^2 &= -N^{-2}(\partial_t\mathcal{C} - \beta^i\partial_i\mathcal{C})^2 + h^{ij}\partial_i\mathcal{C}\partial_j\mathcal{C} \\ V(\mathcal{C}) &= \text{potential energy function depending on } \mathcal{C} \text{ values} \\ \|\mathcal{R}\|_g^2 &= R_{\mu\nu\rho\sigma}R^{\mu\nu\rho\sigma} \end{aligned}$$

where K_{ij} is the second fundamental form, $K = h^{ij}K_{ij}$.

Now choose a specific gauge: $N = 1$, $\beta^i = 0$ (Gaussian normal coordinates), and assume spatial metric h_{ij} is time-independent and flat ($R^{(3)}(h) = 0$). Under these simplifications:

$$d\text{vol}_g = \sqrt{h}\, dt d^n x, \quad R(g) = 0, \quad \|\nabla\mathcal{C}\|_g^2 = -(\partial_t\mathcal{C})^2 + \|\nabla_x\mathcal{C}\|_h^2$$

Substitute $\mathcal{C}(t,x) = \Gamma(t)(x)$ and integrate over spatial coordinates. Define functional:

$$F[\Gamma(t)] = \int_{\mathcal{N}} d^n x\sqrt{h}\left[\beta\|\nabla_x\Gamma(t)(x)\|_h^2 + \gamma V(\Gamma(t)(x))\right]$$

By appropriate choice of spatial measure $\sqrt{h}d^n x$ and coefficients $(\alpha, \beta, \gamma, \delta)$, we can arrange:

$$\mathcal{L}_{\text{dyn}}\left(\Gamma(t), \dot{\Gamma}(t), t\right) = \int_{\mathcal{N}} d^n x\sqrt{h}\,\mathcal{L}_{\text{geo}}(g, \mathcal{C}, \nabla\mathcal{C}, R)\big|_{\mathcal{C}=\Gamma(t)}$$

Specific construction: let $\mathcal{N}$ be a single-point manifold ($n = 0$), then h_{ij} reduces to constant 1, and all spatial derivatives vanish. Set:

$$\begin{aligned} \alpha = 0, \quad \beta = -\lambda_3, \quad \gamma = \lambda_2, \quad \delta = \lambda_4 \\ V(\mathcal{C}) = K(\mathcal{C}) \quad \text{(Kolmogorov complexity)} \\ R(g) = 0, \quad \|\nabla\mathcal{C}\|_g^2 = -(\partial_t\mathcal{C})^2 = -\|\dot{\Gamma}\|^2 \end{aligned}$$

Add extra prediction error term as part of $\mathcal{L}_{\text{geo}}$. With this construction we obtain:

$$\mathcal{A}_{\text{dyn}}[\Gamma] = \int_0^T dt \left[\lambda_1 \mathcal{L}_{\text{pred}} + \lambda_2 \mathcal{L}_{\text{comp}} + \lambda_3 \mathcal{L}_{\text{dyn-smooth}} + \lambda_4 \mathcal{L}_{\text{curv}}\right] = \mathcal{A}_{\text{geo}}[g, \mathcal{C}]$$

Thus there exists a construction making the two actions equal.

Part 2: Proof of Extremal Condition Equivalence

Assume Γ^* is an extremum of the dynamical formulation, satisfying the Euler-Lagrange equation (TheoremF.3):

$$\frac{\partial \mathcal{L}_{\text{dyn}}}{\partial \Gamma} - \frac{d}{dt}\left(\frac{\partial \mathcal{L}_{\text{dyn}}}{\partial \dot{\Gamma}}\right) + \frac{d^2}{dt^2}\left(\frac{\partial \mathcal{L}_{\text{dyn}}}{\partial \ddot{\Gamma}}\right) = 0$$

For the geometric formulation, field equations for metric g and field $\mathcal{C}$ are obtained via variation:

$$\frac{\delta \mathcal{A}_{\text{geo}}}{\delta g^{\mu\nu}} = 0 \quad \text{(Einstein-type field equation)} \tag{11}$$

$$\frac{\delta \mathcal{A}_{\text{geo}}}{\delta \mathcal{C}} = 0 \quad \text{(matter field equation)} \tag{12}$$

Under the construction of Part 1, restrict the geometric field equations to the time direction. Specifically, consider the $(0, 0)$ component of equation(11) (Hamiltonian constraint):

$$G_{00} = 8\pi T_{00}$$

where $G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu}$ is the Einstein tensor, $T_{\mu\nu}$ the cognitive energy-momentum tensor.

In the simplified gauge ($N = 1, \beta^i = 0$), the Hamiltonian constraint reduces to:

$$R^{(3)}(h) - K_{ij}K^{ij} + K^2 = 16\pi\rho$$

where $\rho = T_{00}$ is energy density.

When h_{ij} is flat and static, $R^{(3)}(h) = 0$, $K_{ij} = 0$, the constraint becomes $0 = 16\pi\rho$, i.e., $\rho = 0$. This means energy density must be zero, consistent with $\alpha = 0$ (no cosmological constant term) in Part 1.

Equation(12) in the simplified gauge is:

$$\frac{1}{\sqrt{-g}} \partial_\mu (\sqrt{-g} g^{\mu\nu} \partial_\nu \mathcal{C}) - \frac{\partial V}{\partial \mathcal{C}} = 0$$

Substituting $\mathcal{C}(t, x) = \Gamma(t)(x)$ and integrating spatial part yields:

$$\frac{d^2\Gamma}{dt^2} - \frac{\partial V}{\partial \Gamma} = 0$$

This is exactly the Euler-Lagrange equation of the dynamical formulation under appropriate coefficients. Therefore, extremal conditions of the two formulations are equivalent.

Part 3: Proof of Lagrangian Density Correspondence

The Lagrangian densities of the two formulations contain the same cognitive information components:

1. Prediction error term: $\mathcal{L}_{\text{pred}} = \frac{1}{2}\|\Gamma - \Phi(\Gamma)\|^2$ measures deviation between actual and predicted states. In geometric formulation, spatial curvature R measures geometric curvature, reflecting inaccuracy of cognitive models (geodesics in curved space deviate from straight lines). Both measure deviation between "ideal" and "actual".

2. Cognitive complexity term: $\mathcal{L}_{\text{comp}} = K(\Gamma)$ is description complexity of state. In geometric formulation, cognitive potential $V(\mathcal{C})$ penalizes "complex" field configurations. Both implement Occam's razor.

3. Dynamical smoothness term: $\mathcal{L}_{\text{dyn-smooth}} = \frac{1}{2}\|\dot{\Gamma}\|^2$ penalizes rapid changes. In geometric formulation, field gradient term $\beta\|\nabla\mathcal{C}\|_g^2$ penalizes sharp field variations. Both prefer smooth evolution.

4. Curvature term: $\mathcal{L}_{\text{curv}} = \frac{1}{2}\|R(\Gamma)\|^2$ penalizes high-curvature trajectories. In geometric formulation, full curvature term $\delta\|\mathcal{R}\|_g^2$ penalizes high geometric curvature. Both avoid "excessive distortion".

Through appropriate coefficient transformation, the terms can be put in one-to-one correspondence. This correspondence is not merely formal but also in cognitive interpretation: the two mathematical descriptions capture the same cognitive optimization principles.

Part 4: Proof of Optimal Solution Equivalence

Assume Γ^* minimizes $\mathcal{A}_{\text{dyn}}$. By Part 1 construction, get corresponding $(g^*, \mathcal{C}^*)$. Consider any perturbation $(\delta g, \delta\mathcal{C})$, compute variation:

$$\begin{aligned}\delta\mathcal{A}_{\text{geo}}[g^*, \mathcal{C}^*] &= \left.\frac{d}{d\epsilon}\right|_{\epsilon=0} \mathcal{A}_{\text{geo}}[g^* + \epsilon\delta g, \mathcal{C}^* + \epsilon\delta\mathcal{C}] \\ &= \left.\frac{d}{d\epsilon}\right|_{\epsilon=0} \mathcal{A}_{\text{dyn}}[\Gamma^* + \epsilon\delta\Gamma] \quad \text{(by construction correspondence)} \\ &= \delta\mathcal{A}_{\text{dyn}}[\Gamma^*] = 0 \quad \text{(since } \Gamma^* \text{ is extremum)}\end{aligned}$$

Thus $(g^*, \mathcal{C}^*)$ is a critical point of $\mathcal{A}_{\text{geo}}$. Furthermore, by construction, the second variation of $\mathcal{A}_{\text{geo}}$ at $(g^*, \mathcal{C}^*)$ is positive definite (since second variation of $\mathcal{A}_{\text{dyn}}$ at Γ^* is positive definite), so $(g^*, \mathcal{C}^*)$ is a local minimum.

Conversely, from a minimum of geometric formulation we can construct a minimum of dynamical formulation.

Cognitive isomorphism ψ is defined as:

$$\psi : \Gamma^*([0,T]) \to \mathcal{C}^*(\mathcal{M}_{\text{cog}}), \quad \psi(\Gamma^*(t)) = \mathcal{C}^*(t, \cdot)$$

Since $\mathcal{C}^*(t,x) = \Gamma^*(t)(x)$, this is a bijection. Diagram commutativity is verified:

$$(\psi \circ \Gamma^*)(t) = \psi(\Gamma^*(t)) = \mathcal{C}^*(t, \cdot) = \mathcal{C}^*(\iota(t)) = (\mathcal{C}^* \circ \iota)(t)$$

Part 5: Proof of Physical Description Equivalence

In physics, the same system can have different but equivalent descriptions:

- Dynamical description (Lagrangian mechanics): action $S = \int L(q, \dot{q})dt$, focusing on particle trajectory $q(t)$.
- Geometric description (general relativity): action $S_{\text{EH}} = \int (R - 2\Lambda)\sqrt{-g}d^4x$, focusing on spacetime geometry $g_{\mu\nu}$.

They are connected as: in general relativity, test particle motion is described by geodesic equations, which can be derived from Einstein field equations. Similarly, in cognitive systems:

1. Dynamical formulation directly describes optimal evolution trajectory of cognitive state $\Gamma(t)$.
2. Geometric formulation describes overall geometry $(\mathcal{M}_{\text{cog}}, g)$ of cognitive state space.
3. Optimal trajectory of cognitive state corresponds to geodesics in geometry.

This equivalence reflects unity of multi-level description of cognitive systems: microscopic trajectory optimization and macroscopic spatial structure optimization are two aspects of the same phenomenon. □

Remark F.4. *Perspective of the two formulations of minimum cognitive action principle:*

- ***Dynamical formulation**: Focuses on temporal evolution of cognitive state, corresponds to Lagrangian formulation of classical mechanics.*
- ***Geometric formulation**: Focuses on spatial structure of cognitive state, corresponds to field-theoretic formulation of general relativity.*

TheoremF.15 proves equivalence of the two formulations, providing solid mathematical foundation for unified description of cognitive systems: whether from dynamical or geometric perspective, they describe the same cognitive reality.

F.5 Cognitive Free Energy Principle: Constrained Formulation

Definition F.13 (Cognitive System Constraints)**.** *For a generalized cognitive system $\mathfrak{G}$, define constraint set $\mathcal{C} = \{\mathcal{C}_i\}$, including:*

1. ***Structural constraints**: $\mathcal{C}_{struct} = (C_M, C_{\mathbb{H}}, C_{\mathcal{P}}, C_{\mathscr{F}})$*
   - *C_M: constraints on carrier set M (cardinality, algebraic structure)*
   - *$C_{\mathbb{H}}$: constraints on cognitive hierarchy sequence (depth, connection pattern)*
   - *$C_{\mathcal{P}}$: constraints on parameter fiber category*
   - *$C_{\mathscr{F}}$: constraints on state universe sheaf*
2. ***Performance constraints**: $\mathcal{C}_{perf} = (C_{acc}, C_{eff}, C_{stab})$*
   - *C_{acc}: prediction accuracy constraint*
   - *C_{eff}: computational efficiency constraint*
   - *C_{stab}: stability constraint*
3. ***Resource constraints**: $\mathcal{C}_{res} = (C_{time}, C_{space}, C_{energy})$*

Definition F.14 (Constraint Violation Measure)**.** *For a system $\mathfrak{G}$ and constraints $\mathcal{C}$, define degree of constraint violation:*

$$V(\mathfrak{G}, \mathcal{C}) = \sum_i w_i \cdot v_i(\mathfrak{G}, \mathcal{C}_i)$$

where $w_i \geq 0$ are weights, v_i is violation measure for the ith constraint:

$$v_i(\mathfrak{G}, \mathcal{C}_i) = \begin{cases} 0 & \text{if } \mathfrak{G} \text{ satisfies } \mathcal{C}_i \\ d(\mathfrak{G}, \mathcal{S}_i) & \text{otherwise} \end{cases}$$

Here $\mathcal{S}_i$ is the set of systems satisfying $\mathcal{C}_i$, d an appropriate distance.

Definition F.15 (Cognitive Free Energy Functional (Constraint Version))**.** *The **free energy functional** of a cognitive system is defined as:*

$$\mathcal{F}_{cog}[\mathfrak{G}, \mathcal{C}] = \underbrace{E(\mathfrak{G})}_{\text{internal energy}} + \underbrace{T \cdot V(\mathfrak{G}, \mathcal{C})}_{\text{constraint entropy}} - \underbrace{S(\mathfrak{G})}_{\text{cognitive entropy}}$$

where:

- *$E(\mathfrak{G}) = \mathbb{E}_\gamma[\mathcal{A}[\gamma]]$ is the average action*
- *$T > 0$ is a cognitive temperature parameter*
- *$S(\mathfrak{G}) = -\sum p(\gamma) \log p(\gamma)$ is entropy of trajectory distribution*

Axiom F.2 (Cognitive Free Energy Principle (Constraint Version))**.** ***Optimal cognitive systems** minimize the cognitive free energy functional:*

$$\mathfrak{G}^* = \arg\min_{\mathfrak{G} \in \mathcal{G}} \mathcal{F}_{cog}[\mathfrak{G}, \mathcal{C}]$$

where $\mathcal{G}$ is the set of all possible generalized cognitive systems.

*The **cognitive free energy principle** asserts: AGI systems tend to minimize their cognitive free energy functional during evolution, i.e.:*

$$\mathfrak{M}^* = \arg\min_{\mathfrak{M} \in \mathcal{U}(\mathcal{C})} \mathcal{F}_{cog}(\mathfrak{M}, \mathcal{C})$$

where $\mathcal{U}(\mathcal{C})$ is the set of feasible systems satisfying constraints $\mathcal{C}$.

Definition F.16 (Minimal Cognitive Dimension)**.** *Let $\mathcal{M} \in \mathbf{AGI}$ satisfy the AGI axiom system. Define its **minimal cognitive dimension** $k_{\min}$ as the smallest positive integer satisfying:*

$$k_{\min} = \min\{k \in \mathbb{N}^+ \mid \exists \text{ generating set } S \subseteq \mathcal{M}, |S| = k, \mathrm{Gen}(S) = \mathcal{M}\}$$

i.e., the size of a minimal generating set of $\mathcal{M}$. By TheoremC.6, when $\mathcal{M}$ satisfies the minimality axiom, $\mathcal{M} \cong \mathcal{B}_4^{k_{\min}}$.

Theorem F.16 (Necessity of Cognitive Free Energy Minimization)**.** *Let $\mathcal{M} \in \mathbf{AGI}$ satisfy AxiomA.1, and its meta-dynamics sheaf paradigm $\mathfrak{M}$ performs a learning task $\mathfrak{T}$. Then under the following conditions, the system necessarily tends to minimize cognitive free energy:*

1. ***Limited resources**: System has finite computational resources (time, space, energy);*
2. ***Information bottleneck**: System must trade off between compressed representation and information preservation;*
3. ***Prediction error minimization**: System seeks to accurately predict future states;*
4. ***Minimality preference**: System favors simplest explanations (Occam's razor principle).*

Proof. **Step 1: Formulate optimization problem**. Consider learning task $\mathfrak{T}$, dataset $\mathcal{D} = \{(x_i, y_i)\}$. System needs to learn mapping $f : \mathcal{X} \to \mathcal{Y}$.

System implements the task via meta-dynamics sheaf paradigm $\mathfrak{M}$, with predicted output $\hat{y} = \mathrm{Dec} \circ \Phi_T \circ \mathrm{Enc}(x)$.

Define total objective function:

$$\mathcal{J}(\mathfrak{M}) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(y, \hat{y})] + \lambda_1 \mathcal{C}_{\text{base}}(\mathfrak{M}) + \lambda_2 R(\mathfrak{M})$$

where ℓ is loss function, R regularization term.

Step 2: Formalize information bottleneck. System must trade off between compressed representation and information preservation. Let joint distribution of input X and output Y be $p(x, y)$, system learns intermediate representation $Z = \mathrm{Enc}(X)$.

Information bottleneck principle requires minimizing:

$$\mathcal{L}_{\text{IB}} = I(X; Z) - \beta I(Z; Y)$$

where $I(\cdot;\cdot)$ is mutual information, $\beta > 0$ trade-off parameter.

This is equivalent to minimizing $I(X; Z)$ under constraint $I(Z; Y) \geq I_0$. This principle naturally fits into cognitive free energy framework: first term $I(X; Z)$ corresponds to representation complexity, second term $I(Z; Y)$ corresponds to data fitting capability.

Step 3: Derive free energy. Consider variational distributions $q(z|x)$ and $q(y|z)$, true distribution $p(x, y)$, prior distribution $p(z)$. Variational form of information bottleneck is:

$$\mathcal{F}_{\text{IB}} = \mathbb{E}_{p(x)}[D_{\text{KL}}(q(z|x) \| p(z))] - \beta \mathbb{E}_{p(x,y)}\left[\mathbb{E}_{q(z|x)}[\log q(y|z)]\right]$$

Note $\mathbb{E}_{p(x,y)}[\mathbb{E}_{q(z|x)}[\log q(y|z)]]$ is expectation of data fitting term (reconstruction likelihood), while $D_{\text{KL}}(q(z|x) \| p(z))$ is complexity term. Therefore $\mathcal{F}_{\text{IB}}$ is a specific form of cognitive free energy functional, where $\mathcal{C}_{\text{base}}$ corresponds to D_{KL} term, data fitting term corresponds to negative log-likelihood.

Step 4: Introduce minimality and Bayesian interpretation. By TheoremD.29, optimal system has minimal structure. From Bayesian perspective, learning is computing posterior distribution $p(\mathfrak{M}|\mathcal{D})$. By Bayes' theorem:

$$p(\mathfrak{M}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathfrak{M})p(\mathfrak{M})}{p(\mathcal{D})}$$

where $p(\mathfrak{M})$ is prior distribution, favoring simpler models (e.g., smaller $k_{\min}$); $p(\mathcal{D}|\mathfrak{M})$ is likelihood, reflecting data fit.

Take negative logarithm:

$$-\log p(\mathfrak{M}|\mathcal{D}) = -\log p(\mathcal{D}|\mathfrak{M}) - \log p(\mathfrak{M}) + \log p(\mathcal{D})$$

Ignoring constant $\log p(\mathcal{D})$, right side is exactly the form of cognitive free energy: $-\log p(\mathcal{D}|\mathfrak{M})$ corresponds to data fitting term, $-\log p(\mathfrak{M})$ corresponds to complexity term (prior preference for simple models). Thus minimizing cognitive free energy is equivalent to maximizing posterior probability.

Step 5: Convergence proof. Since Φ is a contraction mapping (AxiomD.1.3), learning process converges to local minimum. Combined with information bottleneck and minimality, global minimum corresponds to cognitive free energy minimization. □

Corollary F.3 (Optimal Structure of AGI System under Free Energy Principle)**.** *Under cognitive free energy principle, optimal AGI system structure $\mathfrak{M}^*$ satisfies:*

1. ***Carrier minimality**: $M^* \cong \mathcal{B}_4^{k_{\min}}$, where $k_{\min}$ is minimal;*
2. ***Dynamics minimality**: Φ^* generated by $3k_{\min}$ irreducible generators;*
3. ***Encoding efficiency**: Enc^* is asymptotically optimal semantic encoding (TheoremD.9);*
4. ***Embedded meta-prediction**: $\mathrm{MetaPred}^*$ necessarily exists and is unique (TheoremD.18).*

Proof. From cognitive free energy minimization:

1. Carrier minimality: $\mathcal{C}_{\text{base}}$ term penalizes large carrier sets; minimal carrier is $\mathcal{B}_4^{k_{\min}}$;
2. Dynamics minimality: number of generators directly affects $\mathcal{C}_{\text{base}}$; minimal generating set has $3k_{\min}$ generators;
3. Encoding efficiency: data fitting term requires Enc to minimize reconstruction error, i.e., achieve rate-distortion lower bound;
4. Embedded meta-prediction: self-referential term D_{KL} requires system have self-monitoring capability, provided by $\mathrm{MetaPred}$.

□

F.6 Deriving Necessary Constraints for AGI Systems from Cognitive Free Energy Principle

Based on unified constraint framework, we can analyze the minimal set of constraints needed to derive a complete AGI infinite cognitive capacity system.

Theorem F.17 (Minimal Sufficient Constraint Set for Infinite Cognitive Capacity)**.** *There exists a minimal constraint set that is sufficient to derive the complete AGI infinite cognitive capacity system. This minimal sufficient set contains two core constraints:*

*1. **Carrier constraint** among structural constraints:*

$$C_M: \quad M \cong \mathcal{B}_4^k \quad (k \in \mathbb{N}^+)$$

where $\mathcal{B}_4$ is the four-element Boolean algebra.

*2. **Functional completeness constraint** among operation constraints:*

$$C_{\textit{func}}: \quad \Omega \textit{ is functionally complete on } M$$

Specifically implemented as Ω containing Boolean operations $\{\neg, \wedge, \vee\}$ or an equivalent minimal complete set.

From these two hard constraints, combined with the minimality optimization goal (AxiomA.2), we can strictly derive:

1. *Self-referential closure: existence of an approximate fixed-point combinator;*
2. *Cognitive completeness: system is a cognitively complete object;*
3. *Turing completeness: ability to encode Turing machines, satisfying Turing-degree completeness and universal reducibility;*
4. *Machine learning mechanism structure: all prediction mechanisms must take fixed-point form, MetaPred necessarily exists and is unique.*

Proof. Assume system satisfies: 1. $M \cong \mathcal{B}_4^k$ (carrier constraint) 2. Ω functionally complete (operation constraint)

Step 1: Derivation of self-referential closure

Since $\mathcal{B}_4$ has an approximate fixed-point combinator (Theorem 4.7), its power $\mathcal{B}_4^k$ inherits this property by construction. Specifically construct:

$$\mathcal{Y}_{\mathcal{B}_4^k}(f) = (\mathcal{Y}_{\mathcal{B}_4}(f_1), \ldots, \mathcal{Y}_{\mathcal{B}_4}(f_k))$$

where $f = (f_1, \ldots, f_k)$ is the decomposed endomorphism.

Step 2: Derivation of Turing completeness

Boolean algebra $\mathcal{B}_4^k$ can encode Turing machines:

1. Each $\mathcal{B}_4$ element can represent 2 bits of information
2. Boolean circuits can simulate Turing machine state transitions
3. Functionally complete operation set Ω can construct any Boolean function

Thus system satisfies Turing-degree completeness and universal reducibility.

Step 3: Derivation of machine learning mechanism structure

Since an approximate fixed-point combinator exists, all prediction mechanisms must take fixed-point form (Theorem 5.9). By minimality axiom, $\mathrm{MetaPred}$ necessarily exists and is unique (Theorem 5.10).

Step 4: Derivation of cognitive completeness

$\mathcal{B}_4^k$ is a cognitively complete object in category **AGI**, satisfying cognitive density, projectivity, and reflexivity. □

Theorem F.18 (Critical Conditions for Optimality of Continuous Dynamics)**.** *The carrier set being $\mathcal{B}_4^k$ itself does not enforce continuous dynamics; optimality of continuous dynamics comes from external task environment and computational efficiency constraints. The minimal combination of constraints that makes continuous dynamics the optimal solution is:*

1. ***Input space topology constraint*** *(necessity):*

$$C_{\textit{input}}: \quad \mathcal{X} \textit{ is a connected compact metric space}$$

and requiring encoding map to satisfy continuity condition.

2. ***Generalization smoothness constraint*** *(cruciality):*

$$C_{\textit{smooth}}: \quad \textit{Prediction function } f: \mathcal{X} \to \mathcal{Y} \textit{ is } \alpha-\text{Hölder } \textit{continuous}$$

3. ***Computational efficiency constraint*** *(decisive):*

$$C_{\textit{efficiency}}: \quad \textit{Learning algorithm time complexity } T(n, \epsilon) = O\left(\text{poly}(n, 1/\epsilon)\right)$$

where n is problem size, ϵ target accuracy.

These three constraints together form a "critical point" forcing the system to adopt continuous dynamics.

Constraint enforcement process. **1. Mathematical form of input continuity constraint**

Let $\mathcal{X}$ be a connected compact metric space, with finite open cover $\{U_i\}_{i=1}^N$. Encoding map $\mathrm{Enc} : \mathcal{X} \to \mathcal{B}_4^k$ must satisfy:

$$\forall \epsilon > 0, \exists \delta > 0 : d_{\mathcal{X}}(x, y) < \delta \Rightarrow d_{\mathcal{M}}(\mathrm{Enc}(x), \mathrm{Enc}(y)) < \epsilon$$

This forces embedding the discrete $\mathcal{B}_4^k$ into a continuous space $[0, 1]^{2k}$.

2. Optimization impact of smoothness constraint

$\alpha -$ Hölder continuity of prediction function $f : \mathcal{X} \to \mathcal{Y}$:

$$|f(x) - f(y)| \leq L \cdot d_{\mathcal{X}}(x, y)^{\alpha}$$

This increases probability that gradient-based optimization algorithms (e.g., SGD) converge to global optimum.

3. Complexity analysis of computational efficiency constraint

Consider complexity of Boolean function search:

- Direct search on $\mathcal{B}_4^k$: complexity $O(4^k)$
- Via continuous relaxation and gradient descent: complexity $O(\mathrm{poly}(k))$

When k is large, continuous method has exponential advantage. □

Algorithm 17: Cognitive Free Energy Minimization Learning Algorithm

Input : Initial meta-dynamics sheaf paradigm $\mathfrak{M}_0$, dataset $\mathcal{D}$, constraints $\mathcal{C}$, learning rate η
Output : Optimized paradigm $\mathfrak{M}^*$
begin
  Initialize: $\mathfrak{M} \leftarrow \mathfrak{M}_0$, $t \leftarrow 0$
  while *not converged* **do**
    Step 1: Prediction and error computation
    for *each batch* $(x, y) \in \mathcal{D}_{batch}$ **do**
      $\hat{y} \leftarrow \mathrm{Dec} \circ \Phi_T \circ \mathrm{Enc}(x)$
      $\mathcal{L}_{\text{pred}} \leftarrow \ell(y, \hat{y})$
    Step 2: Complexity assessment
    $\mathcal{C}_{\text{base}} \leftarrow$ compute base complexity (TheoremD.44)
    $\mathcal{L}_{\text{complexity}} \leftarrow \mathcal{C}_{\text{base}} + P_{\text{struct}} + P_{\text{prop}}$
    Step 3: Meta-prediction self-supervision
    $\sigma_{\text{meta}} \leftarrow \mathrm{MetaPred}(\sigma)$
    $\mathcal{L}_{\text{meta}} \leftarrow d(\sigma, \sigma_{\text{meta}})$
    Step 4: Free energy computation
    $\mathcal{F}_{\text{cog}} \leftarrow \mathcal{L}_{\text{pred}} + \beta \mathcal{L}_{\text{complexity}} + \gamma \mathcal{L}_{\text{meta}}$
    Step 5: Parameter update
    $\nabla \leftarrow \nabla_\theta \mathcal{F}_{\text{cog}}$
    $\theta \leftarrow \theta - \eta \nabla$
    $t \leftarrow t + 1$
  return $\mathfrak{M}$

Theorem F.19 (Algorithm Convergence)**.** *Algorithm17 converges to a local minimum of cognitive free energy under the following conditions:*

1. *Loss function ℓ is smooth and convex in parameters θ;*
2. *Learning rate η satisfies Robbins-Monro conditions: $\sum \eta_t = \infty$, $\sum \eta_t^2 < \infty$;*
3. *Gradient estimates are unbiased.*

Convergence rate is $O(1/\sqrt{t})$.

Proof. Since $\mathcal{F}_{\text{cog}}$ is sum of terms, each term satisfies:

1. Prediction loss $\mathcal{L}_{\text{pred}}$: defined by task, usually convex;
2. Complexity term $\mathcal{L}_{\text{complexity}}$: base complexity $\mathcal{C}_{\text{base}}$ usually function of number of parameters, convex; constraint penalties P_{struct} and P_{prop} are convex (squares of distance functions);
3. Meta-prediction loss $\mathcal{L}_{\text{meta}}$: convexity guaranteed by contraction property of $\mathrm{MetaPred}$.

Thus $\mathcal{F}_{\text{cog}}$ is convex, gradient descent converges to global minimum.

Convergence rate guaranteed by SGD theory: for convex functions, SGD achieves $O(1/\sqrt{t})$ convergence rate. □

Theorem F.20 (Physical Necessity of Cognitive Free Energy Minimization)**.** *Under constraints of limited resources, information bottleneck, and prediction error minimization (TheoremF.16), cognitive systems necessarily tend toward free energy minimization states, corresponding to equilibrium states in thermodynamic systems.*

Proof idea. Consider partition function of cognitive system:

$$Z_{\text{cog}} = \int_{\mathcal{U}(\mathcal{C})} e^{-\beta \mathcal{F}_{\text{cog}}(\mathfrak{M})} d\mathfrak{M}$$

where $\mathcal{U}(\mathcal{C})$ is the set of systems satisfying constraints $\mathcal{C}$. By statistical mechanics, system tends to stay in states with lowest free energy, with probability proportional to $e^{-\beta \mathcal{F}_{\text{cog}}}$. □

Theorem F.21 (Mathematical Equivalence of Least Action and Free Energy Principles)**.** *Under appropriate boundary conditions, the minimum cognitive action principle and cognitive free energy principle are equivalent in the following sense:*

$$\min_{\gamma \in \mathcal{T}(\mathfrak{G})} \mathcal{A}[\gamma] \quad \Longleftrightarrow \quad \min_{\mathfrak{G} \in \mathcal{G}} \mathcal{F}_{cog}[\mathfrak{G}, \mathcal{C}]$$

Specifically, when cognitive temperature $T_{cog} \to 0$, free energy minimization reduces to action minimization.

Proof. Consider cognitive free energy functional:

$$\mathcal{F}_{\text{cog}} = E(\mathfrak{G}) + T_{\text{cog}} V(\mathfrak{G}, \mathcal{C}) - S(\mathfrak{G})$$

When $T_{\text{cog}} \to 0$, constraint term $T_{\text{cog}} V$ tends to zero, entropy term $S(\mathfrak{G})$ is maximized at fixed energy, system concentrates on lowest energy states. Average action $E(\mathfrak{G}) = \mathbb{E}_\gamma[\mathcal{A}[\gamma]]$, in zero-temperature limit system concentrates on trajectories with minimal action. □

Theorem F.22 (Unification of the Cognitive Free Energy Principle)**.** *The cognitive free energy principle unifies the following existing theories:*

1. ***Predictive coding theory**: When only the prediction error term is retained in $\mathcal{F}_{cog}$, it reduces to predictive coding;*
2. ***Information bottleneck principle**: When $\mathcal{F}_{cog}$ is expressed in mutual information form, it is equivalent to the information bottleneck;*
3. ***Variational inference**: Minimizing cognitive free energy is equivalent to minimizing variational free energy;*
4. ***Bayesian brain hypothesis**: Minimizing $\mathcal{F}_{cog}$ is equivalent to maximum a posteriori probability estimation;*
5. ***Active inference**: Action selection can be viewed as minimizing expected free energy.*

Proof. **1. Correspondence with Predictive Coding**

When only the prediction error term is retained in $\mathcal{F}_{\text{cog}}$:

$$\mathcal{F}_{\text{cog}}^{\text{PC}} = \mathbb{E}_{p_{\text{data}}}[\|\text{observation} - \text{prediction}\|^2]$$

This is the standard loss function for predictive coding.

2. Equivalence with the Information Bottleneck

Information bottleneck objective: $\min I(X;Z) - \beta I(Z;Y)$

Let $q(z|x)$ be the encoding distribution and $p(y|z)$ the decoding distribution. The variational free energy is:

$$\mathcal{F}_{\text{variational}} = D_{\text{KL}}(q(z|x)\|p(z)) - \mathbb{E}_{q(z|x)}[\log p(y|z)]$$

Setting $p(z)$ as the prior and $\beta = 1$ yields the form of cognitive free energy.

3. Consistency with Variational Inference

In the Bayesian framework, the posterior is $p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta)$. The negative log-posterior is:

$$-\log p(\theta|\mathcal{D}) = -\log p(\mathcal{D}|\theta) - \log p(\theta) + \text{const}$$

This is precisely the cognitive free energy (data fitting term + complexity term). □

Table 5: Correspondence Table for the Unification of the Cognitive Free Energy Principle

Existing Theory	**Special Case of Cognitive Free Energy**	**Physical Correspondence**
Predictive Coding Theory	Retaining only the prediction error term in $\mathcal{F}_{\text{cog}}$	Least squares fitting
Information Bottleneck Principle	$\mathcal{F}_{\text{cog}} = I(X;Z) - \beta I(Z;Y)$	Heat engine efficiency maximization
Variational Inference	$\mathcal{F}_{\text{cog}} = D_{\text{KL}}(q\|p) - \mathbb{E}_q[\log p(D)]$	Near-equilibrium state approximation
Bayesian Brain Hypothesis	$\mathcal{F}_{\text{cog}} = -\log p(\theta\|D)$	Maximum entropy principle
Active Inference	$\mathcal{F}_{\text{cog}}$ includes expected free energy	Non-equilibrium thermodynamics

F.7 Optimal Model for Finite-Layer Metacognition with Continuous Dynamics

Based on the unified constraint framework, we can derive the optimal model under the constraints of finite-layer metacognition combined with bidirectionality and continuous dynamics. This is an important application case of the cognitive free energy principle.

Theorem F.23 (Optimal Model for Finite-Layer Metacognition with Continuous Dynamics)**.** *Assume the system satisfies the following constraints:*

1. ***Structural constraints**:*
   - *Carrier space is a finite-dimensional real vector space: $M = \mathbb{R}^d$*
   - *Dynamics are continuous-time, described by ordinary differential equations*
2. ***Property constraints**:*
   - *Finite number of metacognitive layers: L layers ($L \in \mathbb{N}^+$)*
   - *Each layer has bidirectionality: there exist a perceptual map $\pi_k : M \to H_k$ and an action map $\sigma_k : H_k \to M$ ($k = 1, \ldots, L$)*
   - *The maps are approximate homomorphisms: they preserve the vector space operation structure*

3. ***Minimality constraint****: Among systems satisfying the above constraints, choose the model with minimal complexity.*

Then the optimal system is the ***Hierarchically Coupled Neural Ordinary Differential Equation (HC-NODE)****:*

$$\frac{dx}{dt} = f_0(x, h_1, \ldots, h_L; \theta_0)$$
$$\frac{dh_k}{dt} = f_k(\alpha_k \pi_k(x) + (1-\alpha_k)h_{k-1}, \theta_k), \quad k = 1, \ldots, L$$

where:

- $x(t) \in M = \mathbb{R}^d$ *is the base-level state;*
- $h_k(t) \in H_k \subset \mathbb{R}^{d_k}$ *is the k-th layer cognitive state;*
- $\pi_k(x) = \phi(W_k x + b_k)$ *is the perceptual map, ϕ is the activation function;*
- $\sigma_k(h) = \psi(V_k h + c_k)$ *is the action map, satisfying $\sigma_k \circ \pi_k \approx \mathrm{id}_M$;*
- $\alpha_k \in [0, 1]$ *controls the balance between top-down and bottom-up information flow;*
- f_k *is a smooth function parameterized by θ_k.*

Proof. **Step 1: Formalization of the continuous dynamics constraint**

Due to the continuous dynamics constraint, state evolution must be described by an ODE. Suppose the system has $L + 1$ layers of states:

$$s(t) = (x(t), h_1(t), \ldots, h_L(t)) \in \mathbb{R}^{d_0} \times \mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_L}$$

where $x(t)$ is the base-level state and $h_k(t)$ is the k-th layer cognitive state.

Continuous dynamics require the existence of a smooth vector field F:

$$\frac{ds}{dt} = F(s(t), t)$$

Step 2: Linearization from the approximate homomorphism constraint

The approximate homomorphism constraint requires the perceptual map $\pi_k : \mathbb{R}^{d_0} \to \mathbb{R}^{d_k}$ to satisfy:

$$\pi_k(x + y) \approx \pi_k(x) + \pi_k(y), \quad \pi_k(\alpha x) \approx \alpha \pi_k(x)$$

This forces π_k to be an approximately linear map. In the space of smooth functions, a linear map plus a nonlinear activation is the simplest realization:

$$\pi_k(x) = \phi(W_k x + b_k)$$

where ϕ is the activation function and W_k is the weight matrix.

Step 3: Coupling form from the bidirectionality constraint.

Bidirectionality requires information to flow bidirectionally between layers. A coupling form satisfying this constraint while maintaining computational tractability is:

$$\frac{dh_k}{dt} = f_k(\alpha_k \pi_k(x) + (1-\alpha_k)h_{k-1}, \theta_k)$$

where $h_0 = x$, $\alpha_k \in [0, 1]$ is a trade-off parameter balancing bottom-up ($\pi_k(x)$) and top-down (h_{k-1}) information flow.

Step 4: Solving via complexity minimization.

Define the complexity functional:

$$\tilde{\mathcal{C}} = \sum_{k=0}^{L} d_k + \lambda_1 \sum_{k=1}^{L} \|\sigma_k \circ \pi_k - \mathrm{id}\|^2 + \lambda_2 \mathcal{L}_{\text{task}}$$

where d_k is the state dimension of the k-th layer (penalizing large networks), λ_1 controls the strength of the bidirectionality constraint, and λ_2 controls the task loss.

Solve this optimization problem via the calculus of variations. Consider the Lagrangian:

$$\mathcal{L} = \tilde{\mathcal{C}} + \sum_{k=1}^{L} \int \mu_k(t) \cdot \left(\frac{dh_k}{dt} - f_k(\alpha_k \pi_k(x) + (1-\alpha_k) h_{k-1}, \theta_k) \right) dt$$

where $\mu_k(t)$ are Lagrange multipliers. Applying the Euler-Lagrange equations, and considering that the approximate homomorphism constraint forces π_k and σ_k to be approximately linear maps, the optimal solution takes the form of a Hierarchically Coupled Neural ODE (HC-NODE). Specifically, this form is the unique smooth solution that satisfies all constraints and minimizes $\tilde{\mathcal{C}}$. □

Corollary F.4 (Biological Correspondence of the Finite Metacognition Optimal Model)**.** *The HC-NODE model bears a striking resemblance to the brain's hierarchical processing structure:*

1. *M (input layer) corresponds to the sensory cortex, processing raw sensory input;*
2. *H_1 corresponds to the association cortex, performing feature integration;*
3. *H_2 corresponds to the prefrontal cortex, performing abstract reasoning;*
4. *H_L corresponds to the dorsolateral prefrontal cortex, responsible for metacognitive monitoring.*

The bidirectional connections correspond to the brain's feedforward and feedback pathways, and the parameters α_k correspond to attentional modulation mechanisms.

F.8 Statistical Formulation of the Cognitive Free Energy Principle

Definition F.17 (Probabilistic Description of Cognitive State)**.** *Let $\mathfrak{G}$ be a generalized cognitive system (Definition F.4). Its **probabilistic description** is given by the following elements:*

1. ***State space**: $\Omega_{prob} = \mathscr{F}_\Omega(\Omega)$, the space of global sections, equipped with the Borel σ-algebra $\mathcal{B}(\Omega_{prob})$.*
2. ***Reference measure**: μ is a Radon measure on Ω_{prob} (e.g., the Lebesgue measure in local coordinates).*
3. ***Probability distribution**: $p : \Omega_{prob} \to [0,1]$ is a probability density function satisfying $\int_{\Omega_{prob}} p(\gamma) d\mu(\gamma) = 1$.*
4. ***Energy function**: $U : \Omega_{prob} \to \mathbb{R} \cup \{+\infty\}$ is lower semicontinuous, related to the action:*
$$U(\gamma) = \mathcal{A}[\gamma] + \lambda \cdot C(\gamma)$$
where $C(\gamma)$ is a complexity penalty and $\lambda > 0$ is a trade-off parameter.

Definition F.18 (Cognitive Free Energy (Statistical Version))**.** *Given a probability distribution p, its **cognitive free energy** is defined as:*

$$\mathcal{F}[p] = \underbrace{\mathbb{E}_p[U(\gamma)]}_{\text{average energy}} + \underbrace{T \cdot D_{KL}(p \| p_0)}_{\text{relative entropy}}$$

where p_0 is the prior distribution, D_{KL} is the Kullback-Leibler divergence, and $T > 0$ is the temperature parameter.

Axiom F.3 (Cognitive Free Energy Principle (Statistical Version))**.** ***The equilibrium distribution of a cognitive system** minimizes the free energy:*

$$p^* = \arg \min_{p \in \mathcal{P}(\Omega_{prob})} \mathcal{F}[p]$$

The solution is the Gibbs distribution:

$$p^*(\gamma) = \frac{1}{Z} e^{-U(\gamma)/T}, \quad Z = \int e^{-U(\gamma)/T} d\mu(\gamma)$$

Theorem F.24 (Geometric Representation of Cognitive Free Energy)**.** *Let Ω be a Riemannian manifold with metric tensor g. Then the cognitive free energy can be expressed in geometric form:*

$$\mathcal{F}[\rho] = \int_\Omega \left[U(\omega)\rho(\omega) + \beta^{-1}\rho(\omega)\log\frac{\rho(\omega)}{\rho_0(\omega)} - \rho(\omega)\log p_{data}(\mathcal{D}|\omega) \right] \sqrt{\det g}\, d^n\omega$$

In particular, when Ω is a Kähler manifold, there exists a potential function $\Phi : \Omega \to \mathbb{R}$ such that:

$$\mathcal{F}[\rho] = \int_\Omega \rho(\omega) \left[\Phi(\omega) + \beta^{-1}\log\frac{\rho(\omega)}{\rho_0(\omega)} \right] dV$$

where dV is the volume form and Φ satisfies the complex Monge-Ampère equation.

Proof. Let Ω be an n-dimensional Riemannian manifold with local coordinates $\omega = (\omega^1, \ldots, \omega^n)$. The metric tensor g_{ij} induces the volume form:

$$dV = \sqrt{\det g}\, d\omega^1 \wedge \cdots \wedge d\omega^n$$

The probability density ρ satisfies $\int_\Omega \rho\, dV = 1$. Writing out the integral from the definition gives the first equality.

When Ω is a Kähler manifold, there exists a Kähler potential Φ such that $g_{i\bar{j}} = \partial_i \partial_{\bar{j}} \Phi$. The cognitive potential U can be absorbed into Φ, yielding the unified expression. □

Theorem F.25 (Information-Theoretic Decomposition of Cognitive Free Energy)**.** *The cognitive free energy can be decomposed as:*

$$\mathcal{F}[\rho] = \underbrace{\mathbb{E}_{\omega\sim\rho}[U(\omega)]}_{\textit{complexity}} + \underbrace{\beta^{-1} I(\omega;\mathcal{D})}_{\textit{mutual information}} - \underbrace{H(\mathcal{D})}_{\textit{data entropy}} + \textit{constant}$$

where:

- *$I(\omega;\mathcal{D})$ is the mutual information between the cognitive state ω and the observed data $\mathcal{D}$.*
- *$H(\mathcal{D})$ is the entropy of the data.*

More compactly:

$$\mathcal{F}[\rho] = \beta^{-1} D_{KL}(\rho \| \rho_0 e^{-\beta U}) - \log p_{data}(\mathcal{D})$$

Proof. From the definition:

$$\begin{aligned}
\mathcal{F}[\rho] &= \mathbb{E}_\rho[U] + \beta^{-1} D_{\mathrm{KL}}(\rho\|\rho_0) - \mathbb{E}_\rho[\log p(\mathcal{D}|\omega)] \\
&= \mathbb{E}_\rho[U] + \beta^{-1}\left(\mathbb{E}_\rho[\log\rho] - \mathbb{E}_\rho[\log\rho_0]\right) - \mathbb{E}_\rho[\log p(\mathcal{D}|\omega)] \\
&= \mathbb{E}_\rho[U] + \beta^{-1}\left(H(\rho_0) - H(\rho)\right) + \beta^{-1} I(\omega;\mathcal{D}) - H(\mathcal{D}) \\
&= \mathbb{E}_\rho[U] + \beta^{-1} I(\omega;\mathcal{D}) - H(\mathcal{D}) + \text{constant}
\end{aligned}$$

where $H(\cdot)$ denotes entropy.

The second expression follows directly from the definition of KL divergence:

$$\mathcal{F}[\rho] = \beta^{-1} D_{\mathrm{KL}}\left(\rho \| \frac{1}{Z}\rho_0 e^{-\beta U} p(\mathcal{D}|\omega)\right) - \log Z$$

Note that $Z = p(\mathcal{D}) = \int \rho_0(\omega) e^{-\beta U(\omega)} p(\mathcal{D}|\omega) d\omega$. □

Theorem F.26 (Gradient Flow of Cognitive Free Energy)**.** *Minimizing cognitive free energy is equivalent to the Fokker-Planck equation:*

$$\frac{\partial \rho_t}{\partial t} = \nabla \cdot \left(\rho_t \nabla \frac{\delta \mathcal{F}}{\delta \rho} \right)$$

whose stationary solution is the Gibbs distribution:

$$\rho_\infty(\omega) = \frac{1}{Z}\exp\left(-\beta U(\omega) + \log p(\mathcal{D}|\omega)\right)$$

The corresponding stochastic differential equation is:

$$d\omega_t = -\nabla U(\omega_t) dt + \nabla \log p(\mathcal{D}|\omega_t) dt + \sqrt{2\beta^{-1}} dW_t$$

where W_t is Brownian motion.

Proof. The first variation of cognitive free energy is:

$$\frac{\delta \mathcal{F}}{\delta \rho} = U(\omega) + \beta^{-1}(1 + \log \rho) - \log p(\mathcal{D}|\omega)$$

The gradient flow equation is:

$$\frac{\partial \rho}{\partial t} = \nabla \cdot \left(\rho \nabla \frac{\delta \mathcal{F}}{\delta \rho} \right) = \nabla \cdot \left(\rho \nabla \left[U + \beta^{-1} \log \rho - \log p \right] \right)$$

At steady state $\frac{\partial \rho}{\partial t} = 0$, hence:

$$\nabla \left(U + \beta^{-1} \log \rho - \log p \right) = 0$$

Integrating yields:

$$U + \beta^{-1} \log \rho - \log p = \text{constant}$$

i.e., $\rho \propto \exp(-\beta U + \log p)$.

The corresponding particle dynamics is given by Wasserstein gradient flow theory: particles follow the Langevin equation:

$$d\omega_t = -\nabla \left(U(\omega_t) - \log p(\mathcal{D}|\omega_t) \right) dt + \sqrt{2\beta^{-1}} dW_t$$

□

Theorem F.27 (Noether's Theorem for Cognitive Free Energy)**.** *Suppose the cognitive system has a symmetry action by a Lie group G, $\Phi_g : \Omega \to \Omega$, $g \in G$, satisfying:*

1. *$U(\Phi_g(\omega)) = U(\omega)$ (potential invariance)*
2. *$p(\mathcal{D}|\Phi_g(\omega)) = p(\mathcal{D}|\omega)$ (likelihood invariance)*
3. *$\rho_0(\Phi_g(\omega)) = \rho_0(\omega)$ (prior invariance)*

Then there exists a conserved quantity $J : \Omega \to \mathfrak{g}^*$ (where $\mathfrak{g}$ is the Lie algebra of G) such that along the optimal trajectory:

$$\frac{d}{dt} J(\omega_t) = 0$$

Specifically, for a generator $\xi \in \mathfrak{g}$, the conserved quantity is:

$$J_\xi(\omega) = \langle \nabla_\omega \left(U(\omega) - \log p(\mathcal{D}|\omega) \right), \xi(\omega) \rangle$$

Proof. By Noether's theorem, if the Lagrangian $\mathcal{L}$ is invariant under group action, then a conserved quantity exists. The cognitive free energy can be viewed as an action:

$$S[\omega] = \int \left[\frac{1}{2} \|\dot{\omega}\|^2 - U(\omega) + \log p(\mathcal{D}|\omega) \right] dt$$

The corresponding Lagrangian is:

$$\mathcal{L}(\omega, \dot{\omega}) = \frac{1}{2} \|\dot{\omega}\|^2 - U(\omega) + \log p(\mathcal{D}|\omega)$$

If the action of G leaves $\mathcal{L}$ invariant, i.e., $\mathcal{L}(\Phi_g(\omega), d\Phi_g(\dot{\omega})) = \mathcal{L}(\omega, \dot{\omega})$, then the conserved quantity is:

$$J_\xi(\omega, \dot{\omega}) = \left\langle \frac{\partial \mathcal{L}}{\partial \dot{\omega}}, \xi(\omega) \right\rangle$$

On the optimal trajectory, $\dot{\omega} = \nabla_\omega (U - \log p)$, substituting yields the conclusion. □

Theorem F.28 (Master Equation for Cognitive Free Energy)**.** *The evolution of a cognitive system can be described by a master equation:*

$$\frac{dP_n}{dt} = \sum_{m \neq n} \left(W_{nm} P_m - W_{mn} P_n \right)$$

where P_n is the probability of the system being in cognitive state n, and the transition rates are:

$$W_{nm} = \Gamma \exp\left(-\beta \frac{E_n - E_m}{2}\right) \frac{p(\mathcal{D}|n)}{p(\mathcal{D}|m)}$$

satisfying the detailed balance condition:

$$\frac{W_{nm}}{W_{mn}} = \frac{P_m^{eq}}{P_n^{eq}} = \exp\left(-\beta(E_n - E_m)\right) \frac{p(\mathcal{D}|n)}{p(\mathcal{D}|m)}$$

Proof. The detailed balance condition requires:

$$W_{nm} P_m^{\text{eq}} = W_{mn} P_n^{\text{eq}}$$

The equilibrium distribution is the Gibbs distribution:

$$P_n^{\text{eq}} = \frac{1}{Z} \exp\left(-\beta E_n\right) p(\mathcal{D}|n)$$

Substituting yields:

$$\frac{W_{nm}}{W_{mn}} = \frac{P_m^{\text{eq}}}{P_n^{\text{eq}}} = \frac{\exp(-\beta E_m) p(\mathcal{D}|m)}{\exp(-\beta E_n) p(\mathcal{D}|n)}$$

Choosing a symmetric form:

$$W_{nm} = \Gamma \exp\left(-\beta \frac{E_n - E_m}{2}\right) \sqrt{\frac{p(\mathcal{D}|n)}{p(\mathcal{D}|m)}}$$

satisfies the above condition. □

Table 6: Correspondence Between Constrained and Statistical Formulations of the Cognitive Free Energy Principle

Constrained Version	**Statistical Version**	**Advantages of Statistical Formulation**
Base complexity $\mathcal{C}_{\text{base}}$	Potential $U(\omega)$	More direct physical interpretation
Constraint violation D_{KL}	Relative entropy term	Standard information-theoretic form
Data fitting term	Negative log-likelihood	Standard statistical form
Meta-dynamics Φ	Gradient flow	Explicit dynamical equations
$\mathrm{MetaPred}$	Mean of Gibbs distribution	Naturally emerging fixed point
$\mathcal{B}_4^k$ structure	Discretization of state space Ω	More general continuous-discrete correspondence

Corollary F.5 (Statistical Inference Form of $\mathrm{MetaPred}$)**.** *Within the cognitive free energy framework,* $\mathrm{MetaPred}$ *is naturally defined as the posterior mean:*

$$\mathrm{MetaPred}(\mathcal{D}) = \mathbb{E}_{\omega \sim \rho^*}[\omega] = \frac{1}{Z} \int_\Omega \omega \, \exp\left(-\beta U(\omega) + \log p(\mathcal{D}|\omega)\right) d\omega$$

This is a natural consequence of free energy minimization, requiring no additional assumptions.

Proof. Minimizing $\mathcal{F}[\rho]$ yields the posterior distribution ρ^*. For a point estimate, the natural choice is the posterior mean, which minimizes the mean squared error:

$$\mathrm{MetaPred} = \arg\min_m \mathbb{E}_{\omega \sim \rho^*}[\|\omega - m\|^2]$$

The solution is $m = \mathbb{E}_{\rho^*}[\omega]$. □

Core Advantages of the Statistical Formulation

1. Mathematical unification: unifies statistical physics (free energy), information theory (mutual information), and differential geometry (gradient flow) within a single framework.
2. Formal simplicity: the core equation $\mathcal{F}[\rho] = \mathcal{E}[\rho] - \mathbb{E}_\rho[\log p(\mathcal{D}|\omega)]$ is extremely concise.
3. Variational principle: naturally expressed as a variational problem, allowing direct application of modern mathematical tools like variational calculus and optimal transport.
4. Symmetry highlighting: the application of Noether's theorem reveals conservation laws in cognitive systems.
5. Multi-scale connection: allows derivation of macroscopic distributions (Fokker-Planck equation) from microscopic dynamics (SDE).
6. Computational friendliness: the Gibbs distribution form facilitates sampling and variational inference.

F.9 Unified Cognitive Free Energy Principle

Axiom F.4 (Unified Cognitive Free Energy Principle)**.** ***The equilibrium distribution*** $p^* \in \mathcal{P}(\Omega)$ *of a cognitive system minimizes the following unified free energy functional:*

$$\mathcal{F}[p] = \underbrace{\mathbb{E}_p[U(\omega)]}_{\text{average potential energy}} + \underbrace{\beta^{-1} D_{\mathrm{KL}}(p\|p_0)}_{\text{relative entropy}} - \underbrace{\mathbb{E}_p[\log p_{\mathrm{data}}(\mathcal{D}|\omega)]}_{\text{data fit}}$$

where:

1. Ω *is the cognitive state space (a compact manifold).*
2. $U : \Omega \to \mathbb{R}$ *is the cognitive potential,* $U(\omega) = \mathcal{A}[\gamma_\omega]$ *(the action functional).*
3. p_0 *is the prior distribution (encoding structural constraints).*
4. $\beta > 0$ *is the inverse cognitive temperature.*
5. $p_{\mathrm{data}}(\mathcal{D}|\omega)$ *is the likelihood function for the observed data.*

Theorem F.29 (Zero-Temperature Limit: Reduction to the Principle of Least Action)**.** *When* $\beta \to \infty$*, the minimizing distribution* p^*_β *of the unified free energy functional converges to a Dirac measure:*

$$\lim_{\beta\to\infty} p^*_\beta = \delta_{\omega^*}, \quad \omega^* = \arg\min_{\omega\in\Omega} U(\omega)$$

In this case, the unified cognitive free energy principle is equivalent to the principle of least cognitive action.

Theorem F.30 (Hard Constraint Limit: Reduction to Constrained Optimization Principle)**.** *When the prior distribution* p_0 *is the indicator function of a constraint set, i.e.:*

$$p_0(\omega) \propto \mathbb{I}_{\mathcal{C}}(\omega) = \begin{cases} 1 & \text{if } \omega \in \mathcal{C} \\ 0 & \text{otherwise} \end{cases}$$

minimizing the unified free energy is equivalent to the constrained optimization problem:

$$\min_{\omega\in\mathcal{C}} [U(\omega) - \log p_{\mathrm{data}}(\mathcal{D}|\omega)]$$

In this case, the unified cognitive free energy principle is equivalent to the constrained version of the cognitive free energy principle.

Corollary F.6 (Statistical Free Energy Principle as a Special Case)**.** *When the cognitive potential* $U(\omega)$ *includes the data fit term, i.e.,* $U(\omega) = U_0(\omega) - \log p_{\mathrm{data}}(\mathcal{D}|\omega)$*, the unified cognitive free energy principle can be expressed as the statistical free energy principle:*

$$\mathcal{F}[p] = \mathbb{E}_p[U_0(\omega)] + \beta^{-1} D_{\mathrm{KL}}(p\|p_0).$$

Corollary F.7 (Unification of the Three Principles)**.** *The unified cognitive free energy principle reduces to the following in specific limits:*

1. $\beta \to \infty$*: Principle of least cognitive action.*

2. $p_0 = \mathbb{I}_{\mathcal{C}}$: *Constrained version of the free energy principle.*

3. *General case: Statistical version of the free energy principle.*

Theorem F.31 (Cognitive Noether Theorem)**.** *Suppose the cognitive system has a continuous symmetry of a Lie group G:*

$$\Phi_g : \Omega \to \Omega, \quad g \in G$$

and the action $\mathcal{A}$, prior p_0, and likelihood p_{data} are all invariant under the action of G. Then there exists a conserved current J^μ satisfying:

$$\nabla_\mu J^\mu = 0$$

Specifically, the conserved quantity is:

$$Q = \int_\Sigma J^\mu n_\mu \sqrt{h} d^3x = \textit{constant}$$

where Σ is a spacelike hypersurface in cognitive spacetime.

Definition F.19 (Cognitive Path Integral)**.** *The partition function of a cognitive system is given by a path integral:*

$$Z_{\text{cog}}[\beta, \mathcal{D}] = \int_\Omega \mathcal{D}[\omega] \, \exp\left(-\beta U(\omega) + \log p_{\text{data}}(\mathcal{D}|\omega)\right)$$

All cognitive observables can be computed as:

$$\langle \mathcal{O} \rangle = \frac{1}{Z_{\text{cog}}} \int_\Omega \mathcal{D}[\omega] \, \mathcal{O}(\omega) \exp\left(-\beta U(\omega) + \log p_{\text{data}}(\mathcal{D}|\omega)\right)$$

Algorithm 18: Unified Free Energy Cognitive Algorithm

Input : Initial distribution p_0, observed data $\mathcal{D}$, temperature schedule $\beta(t)$
Output : Optimal distribution p^*, optimal state ω^*
begin
  Initialize: $p \leftarrow p_0$, $t \leftarrow 0$;
  while *not converged* **do**
    Compute free energy gradient: $\nabla_p \mathcal{F}[p] = U + \beta^{-1}(1 + \log p) - \log p_{\text{data}}$;
    Gradient flow update: $\frac{\partial p}{\partial t} = -\nabla \cdot (p \nabla (\nabla_p \mathcal{F}))$;
    Sampling update: $\omega_{t+1} \leftarrow \omega_t - \eta \nabla_\omega \mathcal{F} + \sqrt{2\eta\beta^{-1}}\xi_t$;
    Temperature annealing: $\beta \leftarrow \beta(t+1)$;
    $t \leftarrow t + 1$;
  Extract mode: $\omega^* \leftarrow \mathbb{E}_p[\omega]$;
  return p, ω^*;

Remark F.5. *Overall advantages of the unified free energy principle's formulation:*

- *Single-axiom governance: the entire cognitive theory is unified under one free energy functional.*
- *Natural limit reductions: the three principle formulations are manifestations of the same equation under different parameters.*
- *Highlighted symmetry: Noether's theorem reveals deep conservation laws.*

The mathematical formulation of the unified free energy principle

$$\mathcal{F}[p] = \mathbb{E}_p[U(\omega)] + \beta^{-1} D_{\text{KL}}(p \| p_0) - \mathbb{E}_p[\log p_{\text{data}}(\mathcal{D}|\omega)]$$

has an explicit Bayesian decomposition structure:

- *Prior term:* $D_{\text{KL}}(p \| p_0)$.
- *Likelihood term:* $\mathbb{E}_p[\log p_{\text{data}}(\mathcal{D}|\omega)]$.

- *Cognitive potential term: $\mathbb{E}_p[U(\omega)]$, explicitly connected to $\mathcal{A}[\gamma_\omega]$ in the principle of least action.*

In contrast, the statistical version principle $\mathcal{F}[p] = \mathbb{E}_p[U(\omega)] + \beta^{-1} D_{\mathrm{KL}}(p\|p_0)$ implicitly includes the data fit term within $U(\omega)$, making the structure less transparent.

The unified free energy principle explicitly demonstrates its equivalence to other principles through two limit theorems:

- *Zero-temperature limit ($\beta \to \infty$) $\to$ Principle of least cognitive action.*
- *Hard constraint limit ($p_0 = \mathbb{I}_{\mathcal{C}}$) $\to$ Constrained optimization principle.*

This limit compatibility is implicit in the statistical formulation and requires additional derivation to reveal.

The decomposition of the unified free energy principle corresponds to three fundamental levels of a cognitive system:

- *Intrinsic structure: $U(\omega)$ corresponds to the system's intrinsic complexity and dynamics.*
- *Prior knowledge: p_0 corresponds to initial system constraints and assumptions.*
- *Data adaptation: $p_{\mathrm{data}}(\mathcal{D}|\omega)$ corresponds to interactive learning with the environment.*

This layering is fully consistent with the "prior-likelihood-posterior" Bayesian framework in cognitive science.

The cognitive potential $U(\omega) = \mathcal{A}[\gamma_\omega] + \lambda C(\omega)$ in the unified free energy principle directly corresponds to the self-consistency functional $\mathcal{F}_{geom}$ from the geometric self-consistency principle:

- *$\mathcal{A}[\gamma_\omega]$ corresponds to the geometric action.*
- *$C(\omega)$ corresponds to curvature and topological constraints.*

Table 7: Correspondence Between the Unified Free Energy Principle and Physics Concepts

Structural Level	**Mathematical Realization**	**Cognitive Function / Brain Region Correspondence**
Cognitive Concept	Physics Correspondence	Mathematical Form
Unified Cognitive Free Energy Principle	Quantum field theory path integral	$\mathcal{F}[p] = \langle U \rangle + \beta^{-1} D_{\mathrm{KL}}$
Zero-Temperature Limit	Classical mechanics limit	$\hbar \to 0 \to$ Least action
Hard Constraint Limit	Statistical ensemble constraint	Microcanonical ensemble
Cognitive Noether Theorem	Classical Noether theorem	Symmetry $\to$ Conservation law

F.10 Unification of the Cognitive Free Energy Principle

This section establishes a completely independent framework consisting of two types of cognitive principles with four formulations:

1. **Independent foundation**: First defines completely independent cognitive hierarchy sequences, parameter fiber categories, and state universe sheaves, without relying on any specific algebraic structures from preceding sections.
2. **Principle formulations**: Builds four formulations of two principle types on this independent basis:
   - Principle of least cognitive action: dynamical formulation.

- Principle of least cognitive action: geometric formulation.
- Cognitive free energy principle (constrained version): constraint optimization form.
- Cognitive free energy principle (statistical version): statistical inference form.

3. **Generality**: The framework encompasses both finite cognitive systems ($N < \infty$) and infinite cognitive systems ($N = \infty$).
4. **Equivalence**: Proves the equivalence of the two types of principles under appropriate conditions.
5. **Connection to AGI axioms**: When augmented with conditions of infinity, convergence, and functional completeness, this framework reduces to the AGI axiom system presented earlier.

This framework provides a unified foundation for the analysis and design of cognitive systems, whether finite (like traditional machine learning) or infinite (like ideal AGI).

The cognitive free energy principle provides a unified, consistent, and concise theoretical framework for AGI systems and a broad range of cognitive behaviors, integrating structural constraints, functional requirements, and learning processes into an optimization framework aligned with equivalent principles in physics. Based on this principle, we draw the following conclusions:

1. **Structural optimality**: The optimal algebraic structure for the carrier set of a cognitive system is $\mathcal{B}_4^{k_{\min}}$, a necessary result of minimizing cognitive free energy.
2. **Learning necessity**: Cognitive systems must engage in continuous learning to minimize prediction error, reflected in the data fit term of cognitive free energy.
3. **Metacognition necessity**: The existence of $\mathrm{MetaPred}$ arises from the need for system self-monitoring and adjustment, corresponding to the self-referential term in free energy.
4. **Optimal model for finite-layer metacognition**: The optimal model combining finite-layer metacognition with bidirectionality and continuous dynamics is the hierarchically coupled neural ODE, which closely resembles the brain's hierarchical processing structure.
5. **Minimal sufficiency of constraints**: Deriving the complete infinite cognitive capability system for AGI requires only two core constraints: the $\mathcal{B}_4^k$ carrier structure and a functionally complete operation set.
6. **Critical condition for continuous dynamics**: The optimality of continuous dynamics stems from three external constraints: input continuity, smoothness, and computational efficiency.
7. **Theoretical unification**: The cognitive free energy principle unifies multiple cognitive theories such as predictive coding, information bottleneck, and variational inference.
8. **Physical correspondence**: Cognitive free energy is the generalization of physical free energy into algorithm space, revealing the deep physical structure of cognitive systems.

Remark F.6 (Formal Analogy and Substantive Difference)**.** *The cognitive principles proposed in this paper present interesting formal analogies with variational principles in physics, but essential differences between cognitive and physical systems must be noted.*

***Important distinctions**:*

1. ***Different purposes**: Physical principles describe laws of nature, cognitive principles describe optimization criteria for intelligent systems.*
2. ***Time scale**: Physical action is natural evolution, cognitive action is an abstraction of computational processes.*
3. ***Self-referentiality**: Cognitive systems include representations and predictions of themselves, which is uncommon in fundamental physics (although we later find they should obey the same underlying principles).*

The cognitive free energy principle not only provides a mathematical foundation for AGI system design but also offers a new perspective for understanding the nature of intelligence. In a world of finite resources, intelligent systems find an optimal balance between complexity and adaptability by minimizing cognitive free energy.

Table 8: Comparison of Mathematical Elements of the Three Cognitive Principles

Element	Principle of Least Cognitive Action	Cognitive Free Energy Principle (Constrained)	Cognitive Free Energy Principle (Statistical)
Mathematical Framework	Calculus of variations, differential geometry	Constrained optimization, game theory	Statistical mechanics, information theory
Optimization Variable	Trajectory $\Gamma : [0, T] \to \Omega_{\text{prob}}$	System $\mathfrak{G} = (\{\mathbb{H}_n\}, \mathscr{F}_\Omega, \Phi, \mathcal{C})$	Probability distribution $p : \Omega_{\text{prob}} \to [0, 1]$
Target Functional	Action $\mathcal{A}[\Gamma] = \int L dt$	Free energy $\mathcal{F}(\mathfrak{G}, \mathcal{C}) = E + TV - S$	Free energy $\mathcal{F}[p] = \mathbb{E}_p[U] + TD_{\text{KL}}$
Constraint Handling	Boundary conditions, Lagrange multipliers	Explicit constraint set $\mathcal{C}$, violation metric V	Prior distribution p_0, KL divergence
Self-Referential Mechanism	Curvature term $R(\Gamma)$, self-consistency condition	Metacognitive constraints, self-monitoring	Posterior self-consistency, fixed-point equation
Physics Correspondence	Classical mechanics principle of least action	Thermodynamic free energy minimization	Statistical mechanics free energy minimization
Brain Science Correspondence	Prediction error minimization	Brain free energy principle	Bayesian brain hypothesis
Applicable Systems	Finite/Infinite cognitive systems	Resource-constrained system design	Statistical inference and learning
Connection to AGI Axioms	Yields AGI axioms in the infinite-time limit	Yields AGI axioms in the infinite-time limit	Yields AGI axioms in the infinite-time limit

Table 9: Formal Analogy Between Cognitive and Physical Principles

Mathematical Structure	Cognitive System Interpretation	Physical System Interpretation
Variational principle	Least cognitive action	Principle of least action (classical mechanics)
Free energy minimization	Cognitive free energy principle	Thermodynamic free energy minimization
Statistical distribution	Probabilistic description of cognitive state	Ensemble theory (statistical mechanics)
Symmetry and conserved quantities	Preservation of cognitive operations	Noether's theorem (physical symmetries)

G Natural Balance Principle

G.1 Mathematical Formulation of the Natural Balance Principle

Task Distribution Assumption: In this section, we consider the task distribution $\mathcal{D}$ faced by a cognitive system to cover all possible cognitive tasks. Specifically, we assume the task distribution includes the following two types of tasks:

1. **Universal computation tasks**: Require the system to be able to simulate any Turing machine, i.e., possess functionally complete computational capabilities.

2. **Self-prediction tasks**: Require the system to predict its own behavior, i.e., the task loss function includes a self-referential error term.

These assumptions ensure that the optimal system derived from the natural balance principle necessarily possesses the fundamental characteristics required for infinite cognitive ability. In fact, these assumptions are alternative formulations of the infinite recursiveness and self-referential requirements in Axiom A.1 (AGI Foundation Axiom).

Next, we formally define a generalized cognitive system.

Definition G.1 (Cognitive Carrier and Parameter Fiber)**.** *Define the following mathematical objects:*

1. ***Cognitive carrier set**: $\mathcal{M}$ is a non-empty set representing the internal state space of an AGI system. For infinite cognitive systems, $\mathcal{M}$ may be infinite; for finite cognitive systems, $\mathcal{M}$ is finite.*
2. ***Parameter category**: $\mathcal{A}$ is a small category representing an indexing set for external parameters or tasks.*
3. ***Parameter fiber category**: $\pi : \mathcal{P} \to \mathcal{A}$ is a fibered category where:*
   - *For each object $a \in \mathcal{A}$, the fiber $\mathcal{P}_a$ is a category representing the cognitive state space corresponding to parameter a.*
   - *For each morphism $f : a \to b$, there is a lifting functor $F_f : \mathcal{P}_b \to \mathcal{P}_a$.*
4. ***Cognitive sheaf structure**: There exists an approximate sheaf structure $\mathscr{F}_\Omega$ over a topological space Ω such that $\Gamma(\Omega, \mathscr{F}_\Omega)$ (the global section space) is related to $\mathcal{P}$ via cognitive homomorphisms.*

Remark G.1 (Symbol Correspondence)**.** *The symbols in the above definition maintain consistency with previous sections:*

- *$\mathcal{M}$ corresponds to the carrier set in Definition A.1.*
- *$\mathcal{P} \to \mathcal{A}$ corresponds to the parameter fibered category in Definition D.3.*
- *$\mathscr{F}_\Omega$ corresponds to the cognitive sheaf in Definition D.2.*
- *The machine learning dynamics $(\Phi, \mathrm{Enc}, \mathrm{Dec})$ correspond to the endogenous meta-dynamics, encoding, and decoding in Axiom D.1.*

Definition G.2 (Machine Learning Dynamics)**.** *Let $\mathcal{P} \to \mathcal{A}$ be a parameter fibered category and $\mathscr{F}_\Omega$ a cognitive sheaf.*

*A **machine learning dynamics** is a triple $(\Phi, \mathrm{Enc}, \mathrm{Dec})$, where:*

1. *$\Phi : \mathbb{R}_{\geq 0} \times \Gamma(\Omega, \mathscr{F}_\Omega) \to \Gamma(\Omega, \mathscr{F}_\Omega)$ is a continuous dynamics satisfying $\Phi_0 = \mathrm{id}$, $\Phi_{t+s} = \Phi_t \circ \Phi_s$.*
2. *$\mathrm{Enc} : \mathcal{X} \to \Gamma(\Omega, \mathscr{F}_\Omega)$ is an encoding map, mapping the input space $\mathcal{X}$ (a compact metric space) to the state sheaf.*
3. *$\mathrm{Dec} : \Gamma(\Omega, \mathscr{F}_\Omega) \to \mathcal{Y}$ is a decoding map, mapping states to the output space $\mathcal{Y}$.*

The dynamics is related to the parameter fiber via a cognitive homomorphism $\rho : \Gamma(\Omega, \mathscr{F}_\Omega) \to \mathcal{P}$, satisfying compatibility conditions.

Definition G.3 (Generalized Cognitive System)**.** *A **generalized cognitive system** is a quintuple:*

$$\mathfrak{G} = (\mathcal{P} \to \mathcal{A}, \mathscr{F}_\Omega, \Phi, \mathcal{D}, \mathcal{C})$$

where:

1. *$\mathcal{P} \to \mathcal{A}$ is a parameter fibered category (Definition G.1).*
2. *$\mathscr{F}_\Omega$ is a cognitive sheaf, with $\Gamma(\Omega, \mathscr{F}_\Omega)$ as the state space.*

3. Φ *is a machine learning dynamics (Definition G.2).*

4. $\mathcal{D} = \{(\mathcal{X}_i, \mathcal{Y}_i, \ell_i)\}_{i \in I}$ *is a task distribution, each task defined by an input space* $\mathcal{X}_i$*, an output space* $\mathcal{Y}_i$*, and a loss function* $\ell_i : \mathcal{Y}_i \times \mathcal{Y}_i \to \mathbb{R}_{\geq 0}$*.*

5. $\mathcal{C}$ *is a set of constraints (structural, resource, etc.).*

Definition G.4 (Finite and Infinite Cognitive Capacity)**.** *Let* $\mathfrak{G}$ *be a generalized cognitive system (Definition G.3).*

1. $\mathfrak{G}$ *has* ***finite cognitive capacity*** *if there exists a finite-dimensional approximation: a finite-dimensional manifold* $\mathcal{M}_{\textit{fin}}$ *and a sequence of approximate homomorphisms:*

$$\rho_n : \Gamma(\Omega, \mathscr{F}_\Omega) \xrightarrow{\epsilon_n} \mathcal{M}_{\textit{fin}}, \quad \sigma_n : \mathcal{M}_{\textit{fin}} \xrightarrow{\delta_n} \Gamma(\Omega, \mathscr{F}_\Omega)$$

satisfying $\epsilon_n, \delta_n \to 0$*, and the dimension of* $\mathcal{M}_{\textit{fin}}$ *is finite.*

2. $\mathfrak{G}$ *has* ***infinite cognitive capacity*** *if for any compact metric space* $\mathcal{X}$ *and any continuous function* $f : \mathcal{X} \to \mathcal{Y}$*, there exist encoding* Enc *and decoding* Dec *such that:*

$$\sup_{x \in \mathcal{X}} \ell(\mathrm{Dec} \circ \Phi_T \circ \mathrm{Enc}(x), f(x)) \leq \epsilon$$

holds for any $\epsilon > 0$*, and the Kolmogorov complexity of* Φ *is finite:* $K(\Phi) < \infty$*.*

Definition G.5 (Task Performance Functional)**.** *Let* $\mathfrak{G} = (\mathcal{P} \to \mathcal{A}, \mathscr{F}_\Omega, \Phi, \mathcal{D}, \mathcal{C})$ *be a generalized cognitive system. Its* ***task performance functional*** $\mathcal{E} : \mathfrak{G} \to \mathbb{R}_{\geq 0}$ *is defined as:*

$$\mathcal{E}(\mathfrak{G}) = \mathbb{E}_{(\mathcal{X}, \mathcal{Y}, \ell) \sim \mathcal{D}} \left[\inf_{\mathrm{Enc}, \mathrm{Dec}, T} \mathbb{E}_{x \sim P_\mathcal{X}} \left[\ell(\mathrm{Dec} \circ \Phi_T \circ \mathrm{Enc}(x), y(x)) \right] \right]$$

where $y : \mathcal{X} \to \mathcal{Y}$ *is the task target function and* $P_\mathcal{X}$ *is a probability distribution over* $\mathcal{X}$*.*

Definition G.6 (Structural Complexity Functional)**.** *Let* $\mathfrak{G}$ *be a generalized cognitive system. Its* ***structural complexity functional*** $\mathcal{C} : \mathfrak{G} \to \mathbb{R}_{\geq 0}$ *is defined as:*

$$\mathcal{C}(\mathfrak{G}) = K(\Phi) + \dim_{\textit{eff}}(\Gamma(\Omega, \mathscr{F}_\Omega)) + \sum_{i=1}^{m} w_i \cdot v_i(\mathfrak{G}, \mathcal{C}_i)$$

where:

- $K(\Phi)$ *is the Kolmogorov complexity (description length) of the dynamics* Φ*.*
- $\dim_{\textit{eff}}(\Gamma(\Omega, \mathscr{F}_\Omega))$ *is the effective dimension of the state space.*
- $v_i(\mathfrak{G}, \mathcal{C}_i)$ *is a violation metric for the* i*-th constraint* $\mathcal{C}_i$*.*
- $w_i \geq 0$ *are weights.*

Definition G.7 (Self-Normalizing Functionals)**.** *Let* $\mathfrak{G}_0$ *be a reference system. Define normalization constants:*

$$\mathcal{E}_0 = \inf_{\mathfrak{G} \in \mathfrak{C}} \mathcal{E}(\mathfrak{G}), \quad \mathcal{C}_0 = \inf_{\mathfrak{G} \in \mathfrak{C}} \mathcal{C}(\mathfrak{G})$$

where $\mathfrak{C}$ *is the set of all possible cognitive systems.*

Define relative functionals:

$$\tilde{\mathcal{E}}(\mathfrak{G}) = \frac{\mathcal{E}(\mathfrak{G})}{\mathcal{E}_0}, \quad \tilde{\mathcal{C}}(\mathfrak{G}) = \frac{\mathcal{C}(\mathfrak{G})}{\mathcal{C}_0}$$

When $\mathcal{E}_0 = 0$ *or* $\mathcal{C}_0 = 0$*, appropriate limit definitions are used.*

Remark G.2 (Theoretical Significance of Normalization Constants)**.** *The normalization constants* $\mathcal{E}_0$ *and* $\mathcal{C}_0$ *are theoretical minima and may be unknown in practical optimization. However, the core of the natural balance principle lies in balancing relative performance against relative complexity. In practice, we can use empirical estimates or omit normalization, since minimizing the product* $\mathcal{E} \cdot \mathcal{C}$ *is equivalent to minimizing the weighted sum* $\mathcal{E} + \lambda \mathcal{C}$*, where* λ *is determined by the specific problem (see Theorem G.2).*

Axiom G.1 (Cognitive Natural Balance Principle)**.** ***The optimal design of a cognitive system is determined by minimizing the natural balance functional****:*

$$\mathfrak{G}^* = \arg\min_{\mathfrak{G}\in\mathfrak{C}} \mathcal{N}(\mathfrak{G})$$

where the ***natural balance functional*** *is defined as:*

$$\mathcal{N}(\mathfrak{G}) = \tilde{\mathcal{E}}(\mathfrak{G}) \cdot \tilde{\mathcal{C}}(\mathfrak{G})$$

i.e., the product of task performance and structural complexity.

Cognitive interpretation*: Optimal cognitive systems naturally evolve to a balance point between task performance and structural complexity, avoiding both excessive complexity (overfitting) and excessive simplicity (underfitting).*

Axiom G.2 (Balance Symmetry Principle)**.** *Let G be a cognitive symmetry group acting on the system space $\mathfrak{C}$. The natural balance functional is invariant under the action of G:*

$$\mathcal{N}(g \cdot \mathfrak{G}) = \mathcal{N}(\mathfrak{G}), \quad \forall g \in G, \forall \mathfrak{G} \in \mathfrak{C}$$

In particular, when G includes the following symmetries:

- ***Parameter reparameterization symmetry****: Base change of the parameter fiber $\mathcal{P}$.*
- ***Cognitive sheaf homotopy symmetry****: Homotopy equivalence of the cognitive sheaf $\mathscr{F}_\Omega$.*
- ***Dynamics conjugation symmetry****: Topological conjugation of the dynamics Φ.*

Example G.1 (Parameter Reparameterization Symmetry)**.** *Let G be the automorphism group of the parameter category $\mathcal{A}$. For $g \in G$, define its action on the system space $\mathfrak{C}$ as transforming the parameter fiber $\mathcal{P} \to \mathcal{A}$ via the base change g into $\mathcal{P}' \to \mathcal{A}$. The cognitive sheaf $\mathscr{F}_\Omega$ and dynamics Φ are adjusted accordingly so that the overall functionality remains unchanged. The natural balance functional $\mathcal{N}$ is invariant under this action because reparameterization does not alter the system's intrinsic cognitive capacity.*

G.2 Mathematical Properties and Fundamental Theorems

Theorem G.1 (Existence and Uniqueness of Balance Point)**.** *Assume the system space $\mathfrak{C}$ is compact under an appropriate topology, and $\mathcal{E}$ and $\mathcal{C}$ are continuous positive functions. Then:*

1. *There exists $\mathfrak{G}^* \in \mathfrak{C}$ minimizing $\mathcal{N}(\mathfrak{G})$.*
2. *If $\mathcal{E}$ and $\mathcal{C}$ are strictly convex on $\mathfrak{C}$, then $\mathfrak{G}^*$ is unique.*
3. *The balance point satisfies the natural condition:*

$$\nabla\mathcal{E}(\mathfrak{G}^*) \parallel \nabla\mathcal{C}(\mathfrak{G}^*)$$

i.e., the gradients are parallel; specifically:

$$\nabla\mathcal{E}(\mathfrak{G}^*) = -\frac{\mathcal{E}(\mathfrak{G}^*)}{\mathcal{C}(\mathfrak{G}^*)}\nabla\mathcal{C}(\mathfrak{G}^*)$$

Proof. **Existence**: Since $\mathfrak{C}$ is compact and $\mathcal{N}$ is continuous, $\mathcal{N}$ attains a minimum on $\mathfrak{C}$.

Uniqueness: If $\mathcal{E}$ and $\mathcal{C}$ are strictly convex, then $\log\mathcal{E}$ and $\log\mathcal{C}$ are also strictly convex, so $\log\mathcal{N} = \log\mathcal{E} + \log\mathcal{C}$ is strictly convex, hence there is a unique minimum point.

Balance condition: At the minimum point $\mathfrak{G}^*$, $\nabla\mathcal{N}(\mathfrak{G}^*) = 0$. Compute:

$$\nabla\mathcal{N} = \mathcal{C}\nabla\mathcal{E} + \mathcal{E}\nabla\mathcal{C}$$

Substituting $\mathfrak{G}^*$ gives $\mathcal{C}(\mathfrak{G}^*)\nabla\mathcal{E}(\mathfrak{G}^*) + \mathcal{E}(\mathfrak{G}^*)\nabla\mathcal{C}(\mathfrak{G}^*) = 0$, i.e.:

$$\nabla\mathcal{E}(\mathfrak{G}^*) = -\frac{\mathcal{E}(\mathfrak{G}^*)}{\mathcal{C}(\mathfrak{G}^*)}\nabla\mathcal{C}(\mathfrak{G}^*)$$

Hence they are parallel. □

Theorem G.2 (Equivalence of Balance Point)**.** *Assume $\mathfrak{C}$ is convex and $\mathcal{E}$ and $\mathcal{C}$ are continuously differentiable and strictly convex. Then the natural balance principle is equivalent at the balance point to the following weighted optimization problem:*

There exists a unique $\lambda^* > 0$ such that:

$$\min_{\mathfrak{G} \in \mathfrak{C}} \mathcal{J}(\mathfrak{G}) = \mathcal{E}(\mathfrak{G}) + \lambda^* \mathcal{C}(\mathfrak{G})$$

has the same optimal solution $\mathfrak{G}^*$ as the natural balance principle, and:

$$\lambda^* = \frac{\mathcal{E}(\mathfrak{G}^*)}{\mathcal{C}(\mathfrak{G}^*)}$$

Proof. By the balance condition from Theorem G.1:

$$\nabla \mathcal{E}(\mathfrak{G}^*) + \frac{\mathcal{E}(\mathfrak{G}^*)}{\mathcal{C}(\mathfrak{G}^*)} \nabla \mathcal{C}(\mathfrak{G}^*) = 0$$

This is precisely the optimality condition for the weighted optimization problem $\min \mathcal{J} = \mathcal{E} + \lambda \mathcal{C}$ with $\lambda = \frac{\mathcal{E}(\mathfrak{G}^*)}{\mathcal{C}(\mathfrak{G}^*)}$.

By strict convexity, the optimal solutions of both problems are unique, hence they are equivalent. □

Definition G.8 (Natural Scale Parameter)**.** *Define the **natural scale parameter** of a cognitive system:*

$$\kappa(\mathfrak{G}) = \frac{\mathcal{C}(\mathfrak{G})}{\mathcal{E}(\mathfrak{G})}$$

At the balance point $\mathfrak{G}^*$, $\kappa^* = \kappa(\mathfrak{G}^*)$ is an intrinsic characteristic scale of the system, reflecting the complexity required per unit of performance.

Theorem G.3 (Scale and Cognitive Capacity)**.** *Let $\{\mathfrak{G}_n\}_{n=1}^{\infty}$ be a sequence of cognitive systems with increasing scale, satisfying $\kappa(\mathfrak{G}_n) \to \infty$. Then:*

1. *For finite n, $\mathfrak{G}_n$ corresponds to a system with finite cognitive capacity.*
2. *If the limit $\mathfrak{G}_\infty = \lim_{n\to\infty} \mathfrak{G}_n$ exists (in an appropriate topology), then $\mathfrak{G}_\infty$ has infinite cognitive capacity.*
3. *Continuity of transition: If $\mathcal{E}$ and $\mathcal{C}$ are continuous, then:*

$$\lim_{n\to\infty} \mathcal{N}(\mathfrak{G}_n) = \mathcal{N}(\mathfrak{G}_\infty)$$

Proof. **1. Finite cognitive capacity**: From $\kappa(\mathfrak{G}_n) = \frac{\mathcal{C}(\mathfrak{G}_n)}{\mathcal{E}(\mathfrak{G}_n)} \to \infty$, and since $\mathcal{E}(\mathfrak{G}_n)$ is bounded below by $\mathcal{E}_0 > 0$, it follows that $\mathcal{C}(\mathfrak{G}_n) \to \infty$. But $\mathfrak{G}_n$ is a finite system, so its effective dimension is finite and its complexity is finite, leading to a contradiction. Therefore, $\mathcal{E}(\mathfrak{G}_n)$ must tend to $\mathcal{E}_0$ (the minimum possible error), and the system approaches the limit of finite cognitive capacity.

2. Infinite cognitive capacity: At the limit point, $\mathcal{E}(\mathfrak{G}_\infty) = \mathcal{E}_0$ (reaches the theoretical lower bound) and $\mathcal{C}(\mathfrak{G}_\infty)$ is finite (by continuity), so $\kappa(\mathfrak{G}_\infty)$ is finite. The system achieves arbitrary accuracy, possessing infinite cognitive capacity.

3. Continuity: Guaranteed by the continuity of $\mathcal{E}$ and $\mathcal{C}$. □

G.3 Derivation of Specific Cognitive Structures

Theorem G.4 (Natural Emergence of $\mathcal{B}_4^k$ Structure)**.** *Assume the cognitive system requires full logical expressiveness (i.e., $\mathcal{E}(\mathfrak{G})$ attains its minimum when the system is functionally complete), and the structural complexity $\mathcal{C}(\mathfrak{G})$ measures the description length of the algebraic structure. Then the optimal solution $\mathfrak{G}^*$ of the natural balance principle has a carrier set satisfying:*

$$\mathcal{M}^* \cong \mathcal{B}_4^{k^*}$$

where $\mathcal{B}_4$ is the four-element Boolean algebra and k^* is determined by the natural scale κ^*.

Proof. **Step 1: Functional completeness requirement** The task performance functional $\mathcal{E}(\mathfrak{G})$ includes an expressiveness term. Full logical expressiveness requires the system to express all Boolean functions, which necessitates a functionally complete set of operations.

Step 2: Minimal complexity algebra Among all functionally complete algebraic structures, $\mathcal{B}_4$ has the minimal description length (the independent proof of Theorem C.1 can be given via Kolmogorov complexity minimization).

Step 3: Dimension determination Suppose the effective dimension of the state space to be encoded is N. Each copy of $\mathcal{B}_4$ can encode 2 bits of information ($\log_2 4 = 2$), so $k = \lceil \frac{1}{2} \log_2 N \rceil$ copies are needed.

Step 4: Balance point optimization Under the natural balance principle, the choice of k is determined by minimizing the product of $\mathcal{E}$ and $\mathcal{C}$:

$$\min_k [\mathcal{E}(k) \cdot \mathcal{C}(k)]$$

where $\mathcal{E}(k)$ decreases as k increases (increased expressiveness), and $\mathcal{C}(k)$ increases as k increases (increased complexity). The minimum point k^* of the product gives the optimal dimension. □

Theorem G.5 (MetaPred as a Necessary Structure for Infinite Cognitive Capacity)**.** *Let $\mathfrak{G} = (\mathcal{P} \to \mathcal{A}, \mathscr{F}_\Omega, \Phi, \mathcal{D}, \mathcal{C})$ be a generalized cognitive system (Definition G.3), and suppose $\mathfrak{G}$ has infinite cognitive capacity (Definition G.4.2). Then there exists a self-referential prediction module $\mathrm{MetaPred} : \Gamma(\Omega, \mathscr{F}_\Omega) \to \Gamma(\Omega, \mathscr{F}_\Omega)$ satisfying the following properties:*

1. ***Self-referentiality****:* $\mathrm{MetaPred}$ *is the system's predictor of its own dynamics, i.e.:*

$$\mathrm{MetaPred}(\sigma) \approx \Phi_\Delta(\sigma) \quad \textit{for all } \sigma \in \Gamma(\Omega, \mathscr{F}_\Omega)$$

where $\Delta > 0$ is an appropriate time step.

2. ***Approximate commutativity****:* $\mathrm{MetaPred}$ *approximately commutes with the system dynamics Φ:*

$$\mathrm{MetaPred} \circ \Phi_t \approx \Phi_t \circ \mathrm{MetaPred} \quad \textit{for all } t \geq 0$$

3. ***Contraction****: There exists $L < 1$ such that:*

$$d(\mathrm{MetaPred}(\sigma), \mathrm{MetaPred}(\sigma')) \leq L \cdot d(\sigma, \sigma')$$

holds for all $\sigma, \sigma' \in \Gamma(\Omega, \mathscr{F}_\Omega)$.

4. ***Fixed-point property****: There exists a unique $\sigma^* \in \Gamma(\Omega, \mathscr{F}_\Omega)$ (modulo approximation) such that:*

$$\mathrm{MetaPred}(\sigma^*) \approx \sigma^*$$

Proof. **First step: Infinite cognitive capacity implies self-prediction capability**

By the definition of infinite cognitive capacity (Definition G.4.2), the system can approximate any continuous function to arbitrary precision. In particular, consider the task of the system predicting its own dynamics:

Define the self-prediction task $\mathcal{T}_{\text{self}}$:

- Input space: $\mathcal{X} = \Gamma(\Omega, \mathscr{F}_\Omega)$
- Output space: $\mathcal{Y} = \Gamma(\Omega, \mathscr{F}_\Omega)$
- Target function: $f_{\text{self}}(\sigma) = \Phi_\Delta(\sigma)$, where $\Delta > 0$ is a fixed time step
- Loss function: $\ell(\hat{\sigma}, \sigma') = d(\hat{\sigma}, \sigma')$

Since $\mathfrak{G}$ has infinite cognitive capacity, for any $\epsilon > 0$, there exist encoding Enc_ϵ and decoding Dec_ϵ such that:

$$\sup_{\sigma \in \mathcal{X}} d(\mathrm{Dec}_\epsilon \circ \Phi_T \circ \mathrm{Enc}_\epsilon(\sigma), f_{\text{self}}(\sigma)) \leq \epsilon$$

But the construction of encoding and decoding here can be simplified: since the input and output are both the state space itself, the most natural encoding is the identity map. Therefore, there exists a map $P_\epsilon : \Gamma(\Omega, \mathscr{F}_\Omega) \to \Gamma(\Omega, \mathscr{F}_\Omega)$ such that:

$$\sup_{\sigma \in \Gamma(\Omega, \mathscr{F}_\Omega)} d(P_\epsilon(\sigma), \Phi_\Delta(\sigma)) \leq \epsilon$$

Letting $\epsilon \to 0$, we obtain a sequence of maps $\{P_\epsilon\}$ uniformly converging to some map $\mathrm{MetaPred}$ satisfying:

$$\mathrm{MetaPred}(\sigma) = \Phi_\Delta(\sigma) \quad \text{(in the limit sense)}$$

Second step: Proof of approximate commutativity

From the definition of $\mathrm{MetaPred}$, for any $\sigma \in \Gamma(\Omega, \mathscr{F}_\Omega)$ and $t \geq 0$:

$$\begin{aligned} \mathrm{MetaPred}(\Phi_t(\sigma)) &= \lim_{\epsilon \to 0} P_\epsilon(\Phi_t(\sigma)) \\ &\approx \Phi_\Delta(\Phi_t(\sigma)) \quad \text{(by step one)} \\ &= \Phi_t(\Phi_\Delta(\sigma)) \quad \text{(since } \Phi \text{ is a flow)} \\ &\approx \Phi_t(\mathrm{MetaPred}(\sigma)) \quad \text{(by step one)} \end{aligned}$$

Thus $\mathrm{MetaPred} \circ \Phi_t \approx \Phi_t \circ \mathrm{MetaPred}$.

Third step: Proof of contraction

Since Φ is a contraction mapping (Axiom D.1.3), there exists $L_\Phi < 1$ such that:

$$d(\Phi_\Delta(\sigma), \Phi_\Delta(\sigma')) \leq L_\Phi \cdot d(\sigma, \sigma')$$

By the uniform approximation of $\mathrm{MetaPred}$ to Φ_Δ, there exists L satisfying $L_\Phi < L < 1$ such that for sufficiently small ϵ:

$$d(P_\epsilon(\sigma), P_\epsilon(\sigma')) \leq L \cdot d(\sigma, \sigma')$$

Taking the limit $\epsilon \to 0$ yields:

$$d(\mathrm{MetaPred}(\sigma), \mathrm{MetaPred}(\sigma')) \leq L \cdot d(\sigma, \sigma')$$

Fourth step: Proof of fixed-point property

By the contraction mapping principle (Banach fixed-point theorem), the contraction mapping $\mathrm{MetaPred}$ has a unique fixed point σ^* in the complete metric space $\Gamma(\Omega, \mathscr{F}_\Omega)$ satisfying:

$$\mathrm{MetaPred}(\sigma^*) = \sigma^*$$

Since $\mathrm{MetaPred}$ is an approximation of Φ_Δ, the actual fixed point is approximate: $\mathrm{MetaPred}(\sigma^*) \approx \sigma^*$.

Fifth step: Consistency with the natural balance principle

From the perspective of the natural balance principle, a system containing $\mathrm{MetaPred}$ has higher performance $\mathcal{E}_{\text{self}}$ on self-prediction tasks (smaller error), while the increase in structural complexity $\Delta\mathcal{C}$ from adding the $\mathrm{MetaPred}$ module is limited (since $\mathrm{MetaPred}$ can be constructed from the system's existing operations). Therefore, a system containing $\mathrm{MetaPred}$ has a smaller natural balance functional $\mathcal{N} = \tilde{\mathcal{E}} \cdot \tilde{\mathcal{C}}$ and is thus selected under the natural balance principle. □

G.4 Proof of Equivalence with Foundational Cognitive Principles

Definition G.9 (Finite Cognitive Capacity System)**.** *Let $\mathfrak{G} = (\mathcal{P} \to \mathcal{A}, \mathscr{F}_\Omega, \Phi, \mathcal{D}, \mathcal{C})$ be a generalized cognitive system (Definition G.3). $\mathfrak{G}$ is said to have **finite cognitive capacity of depth** N if there exists a positive integer N such that for any cognitive task $\mathcal{T} \in \mathcal{D}$, the system's cognitive state can achieve at most N layers of recursive abstraction. Formally, there exists a finite cognitive hierarchy sequence $\{\mathbb{H}_n\}_{n=0}^N$ satisfying:*

1. $\mathbb{H}_0 = \mathcal{M}_0$ *(base carrier set).*

2. *For $0 \leq n < N$, $\mathbb{H}_{n+1}$ consists of approximate homomorphism triples over $\mathbb{H}_n$, similar to the infinite cognitive case but with finite depth.*

3. *For $n > N$, there is no non-trivial extension of cognitive structure, i.e., $\mathbb{H}_n$ is approximately isomorphic to $\mathbb{H}_N$.*

The cognitive depth N may be determined by resource constraints, computational limitations, or characteristics of the task distribution.

Remark G.3 (Difference Between Finite and Infinite Cognitive Capacity)**.** *Infinite cognitive capacity systems (Definition G.4.2) require that for any precision $\epsilon > 0$ and any continuous function f, the system can approximate f with precision ϵ. Finite cognitive capacity systems have a maximum abstraction depth N, beyond which cognitive tasks cannot be accurately processed. In practical systems, N may be determined by model parameter count, computational resources, or an upper bound on task complexity.*

Theorem G.6 (Emergence of Cognitive Hierarchy under Finite Cognitive Capacity)**.** *Let $\mathfrak{G}^* = (\mathcal{P} \to \mathcal{A}, \mathscr{F}_\Omega, \Phi^*, \mathcal{D}, \mathcal{C})$ be the optimal solution of the natural balance principle (Axiom G.1), and assume $\mathfrak{G}^*$ has finite cognitive capacity of depth N (Definition G.9). Then there exists a cognitive hierarchy sequence $\{\mathbb{H}_n\}_{n=0}^N$ satisfying:*

*1. **Recursive construction**: $\mathbb{H}_0 = \mathcal{M}_0$ (base carrier set), and for $0 \leq n < N$:*

$$\mathbb{H}_{n+1} = \left\{ (\pi, \sigma, \gamma) \mid \pi : \mathcal{M}_0 \xrightarrow{\epsilon_n^\pi} \mathbb{H}_n,\ \sigma : \mathbb{H}_n \xrightarrow{\epsilon_n^\sigma} \mathbb{H}_n,\ \gamma : \mathbb{H}_n \xrightarrow{\epsilon_n^\gamma} \mathcal{W} \right\}$$

where π, σ, γ are ϵ-approximate homomorphisms.

*2. **Connecting homomorphisms**: There exists a sequence of connecting approximate homomorphisms $\{\mu_n : \mathbb{H}_n \xrightarrow{\zeta_n} \mathbb{H}_{n+1}\}_{n=0}^{N-1}$ satisfying compatibility conditions.*

*3. **Finite depth saturation**: For $n \geq N$, the cognitive hierarchy reaches saturation, i.e., there exists an approximate isomorphism $\psi : \mathbb{H}_N \xrightarrow{\epsilon_N} \mathbb{H}_n$, and the approximation error ϵ_N is determined by the system's cognitive depth limit.*

*4. **Error control**: The error sequences $\{\epsilon_n^\pi, \epsilon_n^\sigma, \epsilon_n^\gamma, \zeta_n\}_{n=0}^{N-1}$ satisfy:*

$$\sum_{n=0}^{N-1} (\epsilon_n^\pi + \epsilon_n^\sigma + \epsilon_n^\gamma + \zeta_n) < \infty$$

and the total error is bounded by the cognitive depth N.

*5. **Approximation relation to infinite case**: As $N \to \infty$, the finite cognitive capacity system $\mathfrak{G}_N^*$ converges to an infinite cognitive capacity system $\mathfrak{G}_\infty^*$, and the cognitive hierarchy sequences converge layer-wise.*

Proof. **Step 1: Construction of the base layer**

Since $\mathfrak{G}^*$ is the optimal solution, by the natural balance principle, its base carrier set $\mathcal{M}_0$ should have minimal structural complexity while meeting task performance requirements. By Theorem G.4, even for finite cognitive capacity systems, $\mathcal{M}_0 \cong \mathcal{B}_4^{k_0}$ holds, where k_0 is determined by task complexity.

Set $\mathbb{H}_0 = \mathcal{M}_0$.

Step 2: Recursive construction and natural balance

We construct the cognitive hierarchy sequence by induction. Assume $\mathbb{H}_n$ has been constructed ($n < N$). Consider constructing $\mathbb{H}_{n+1}$.

Since the system has cognitive capacity of depth N, it can perform the $n+1$-th layer abstraction before reaching the depth limit. Define the three components of $\mathbb{H}_{n+1}$ as follows:

1. **Perceptual component** π_{n+1}: For any $m \in \mathcal{M}_0$, the system needs to map the base state to the $n+1$-th layer abstraction. Since the system has finite cognitive capacity, there exists an approximate homomorphism $\pi_{n+1} : \mathcal{M}_0 \to \mathbb{H}_n$, its construction depending on the abstraction capability already established at layer n.

Specific construction: Let $\rho_n : \Gamma(\Omega, \mathscr{F}_\Omega) \to \mathbb{H}_n$ be an approximate homomorphism (induced by the system dynamics Φ^*), and $\sigma_0 : \mathcal{M}_0 \to \Gamma(\Omega, \mathscr{F}_\Omega)$ an embedding. Define:

$$\pi_{n+1}(m) = \rho_n \circ \Phi^*_\Delta \circ \sigma_0(m)$$

where $\Delta > 0$ is an appropriate time step. Since Φ^* is an approximate homomorphism, π_{n+1} is also an ϵ^π_n-approximate homomorphism.

2. **Self-referential component** σ_{n+1}: The system needs self-reference to the n-th layer cognition. Define:

$$\sigma_{n+1}(h) = \rho_n \circ \Phi^*_\Delta \circ \sigma_n(h)$$

where $\sigma_n : \mathbb{H}_n \to \Gamma(\Omega, \mathscr{F}_\Omega)$ is an approximate inverse. Due to the approximate homomorphism property of Φ^*, σ_{n+1} is an ϵ^σ_n-approximate homomorphism.

3. **Action component** γ_{n+1}: The system needs to map the $n+1$-th layer cognition to the external world $\mathcal{W}$. Define:

$$\gamma_{n+1}(h) = \mathrm{Dec} \circ \Phi^*_T \circ \sigma_n(h)$$

where Dec is the decoding map. Due to the approximate homomorphism properties of Dec and Φ^*, γ_{n+1} is an ϵ^γ_n-approximate homomorphism.

Step 3: Construction of connecting homomorphisms

Define the connecting homomorphism $\mu_n : \mathbb{H}_n \to \mathbb{H}_{n+1}$ as:

$$\mu_n(h) = (\pi_{n+1}, \sigma_{n+1}, \gamma_{n+1})$$

where the three components are defined as above. Since each component is an approximate homomorphism, μ_n is a ζ_n-approximate homomorphism.

Step 4: Proof of finite depth saturation

When $n = N$, the system reaches its cognitive depth limit. By Definition G.9, for $n > N$, $\mathbb{H}_n$ is approximately isomorphic to $\mathbb{H}_N$. This means the recursive construction does not produce new cognitive structures beyond layer N.

Formally, for $n \geq N$, define the approximate isomorphism $\psi_n : \mathbb{H}_N \to \mathbb{H}_n$ as:

$$\psi_n = \mu_{n-1} \circ \mu_{n-2} \circ \cdots \circ \mu_N$$

Since each μ_k is an approximate homomorphism, ψ_n is also an approximate homomorphism. Due to the finite cognitive capacity, the approximation error ϵ_N of ψ_n is determined by the depth limit N and does not decrease indefinitely.

Step 5: Proof of error control

The natural balance principle requires minimizing $\mathcal{N} = \tilde{\mathcal{E}} \cdot \tilde{\mathcal{C}}$. Consider the impact of errors in the hierarchy construction on performance:

- Task performance $\mathcal{E}_n$: As hierarchy depth n increases, abstraction capability improves, so $\mathcal{E}_n$ decreases. - Structural complexity $\mathcal{C}_n$: As n increases, description length increases, so $\mathcal{C}_n$ increases.

Under finite depth N, the minimum of the product $\mathcal{N}_N = \mathcal{E}_N \cdot \mathcal{C}_N$ determines the optimal error allocation. Through variational analysis, the optimal error sequence should satisfy:

$$\epsilon_n \sim \frac{1}{(n+1)^{1+\delta}}, \quad \delta > 0, \quad 0 \leq n < N$$

thus ensuring:

$$\sum_{n=0}^{N-1} \epsilon_n < \sum_{n=0}^{\infty} \frac{1}{(n+1)^{1+\delta}} < \infty$$

Even for finite N, this series is convergent.

Step 6: Approximation relation to the infinite case

As $N \to \infty$, the sequence of finite cognitive capacity systems $\{\mathfrak{S}^*_N\}$ converges to an infinite cognitive capacity system $\mathfrak{S}^*_\infty$, guaranteed by the following facts: 1. The carrier set sequence

$\{\mathcal{M}_0^{(N)}\}$ converges to $\mathcal{B}_4^{k_\infty}$. 2. The dynamics sequence $\{\Phi^{(N)}\}$ converges to $\Phi^{(\infty)}$ in an appropriate topology. 3. The cognitive hierarchy sequences converge pointwise: for any fixed n, $\mathbb{H}_n^{(N)} \to \mathbb{H}_n^{(\infty)}$ as $N \to \infty$ and $N > n$.

Convergence is ensured by the continuity of the natural balance principle: when increasing cognitive depth N, if $\mathcal{N}$ can be further reduced, the system will choose a larger N; when $\mathcal{N}$ reaches its minimum, the corresponding N may be finite (if the task distribution only requires finite abstraction) or infinite (if infinite recursive abstraction is required). □

Corollary G.1 (Structural Characteristics of Finite Cognitive Capacity Systems)**.** *A finite cognitive capacity system of depth N has the following characteristics:*

1. ***Carrier algebraic structure**: $\mathcal{M}_0 \cong \mathcal{B}_4^{k_0}$, where k_0 grows with N but remains finite.*
2. ***Finite hierarchy depth**: The cognitive hierarchy sequence saturates at depth N, with deeper layers providing no new cognitive capability.*
3. ***Finite error accumulation**: Total approximation error has an upper bound, determined by depth N and task complexity.*
4. ***Match with task distribution**: The optimal depth N^* is determined by the complexity of the task distribution $\mathcal{D}$. For simple task distributions, N^* is small; for complex ones, N^* is large.*

Remark G.4 (Practical Significance)**.** *In practical machine learning systems, finite cognitive capacity corresponds to:*

- *Finite neural network depth: Corresponds to cognitive hierarchy depth N.*
- *Finite number of parameters: Corresponds to carrier set size $|\mathcal{M}_0| = 4^{k_0}$.*
- *Finite computational resources: Limits the ability to extend the cognitive hierarchy.*

Theorem G.6 explains why successful machine learning systems develop hierarchical structures even with limited resources: it is a necessary consequence of balancing performance and complexity under the natural balance principle.

Example G.2 (Deep Neural Network as a Finite Cognitive Capacity System)**.** *Consider a neural network of depth L, where each layer corresponds to a cognitive hierarchy $\mathbb{H}_l$:*

- *$\mathbb{H}_0$: Input layer, corresponding to raw data representation.*
- *$\mathbb{H}_l$ ($1 \leq l \leq L$): l-th hidden layer, corresponding to the l-th layer abstraction.*
- *Connecting homomorphism μ_l: Implemented by weight matrices and activation functions.*
- *Finite depth saturation: When network depth exceeds a threshold L^*, performance stops improving or even degrades, corresponding to cognitive depth saturation.*

This network, as a finite cognitive capacity system, has its depth L determined by architecture design, with the optimal depth L^* determined by task complexity.

Theorem G.7 (Natural Emergence of Cognitive Hierarchy)**.** *Let $\mathfrak{G}^* = (\mathcal{P} \to \mathcal{A}, \mathscr{F}_\Omega, \Phi^*, \mathcal{D}, \mathcal{C})$ be the optimal solution of the natural balance principle (Axiom G.1), and assume $\mathfrak{G}^*$ has infinite cognitive capacity (Definition G.4.2). Then there exists a cognitive hierarchy sequence $\{\mathbb{H}_n\}_{n=0}^{\infty}$ satisfying:*

*1. **Recursive construction**: $\mathbb{H}_0 = \mathcal{M}_0$ (base carrier set), and for $n \geq 0$:*

$$\mathbb{H}_{n+1} = \left\{ (\pi, \sigma, \gamma) \mid \pi : \mathcal{M}_0 \xrightarrow{\epsilon_n^\pi} \mathbb{H}_n,\ \sigma : \mathbb{H}_n \xrightarrow{\epsilon_n^\sigma} \mathbb{H}_n,\ \gamma : \mathbb{H}_n \xrightarrow{\epsilon_n^\gamma} \mathcal{W} \right\}$$

where π, σ, γ are ϵ-approximate homomorphisms.

*2. **Connecting homomorphisms**: There exists a sequence of connecting approximate homomorphisms $\{\mu_n : \mathbb{H}_n \xrightarrow{\zeta_n} \mathbb{H}_{n+1}\}$ satisfying compatibility conditions.*

*3. **Limit isomorphism**: The limit object $\mathbb{H}_\infty = \varinjlim \mathbb{H}_n$ is approximately isomorphic to the cognitive sheaf $\mathscr{F}_\Omega$.*

*4. **Convergence condition**: The error sequences $\{\epsilon_n^\pi, \epsilon_n^\sigma, \epsilon_n^\gamma, \zeta_n\}$ satisfy ordinary series convergence:*

$$\sum_{n=0}^{\infty} (\epsilon_n^\pi + \epsilon_n^\sigma + \epsilon_n^\gamma + \zeta_n) < \infty$$

In particular, this cognitive hierarchy sequence satisfies all conditions of Axiom A.1 in this paper.

Proof. **Proof idea**: Starting from the optimal solution $\mathfrak{G}^*$ of the natural balance principle, recursively construct the cognitive hierarchy sequence. The key to the construction is utilizing the self-referential structure and task performance requirements from the natural balance principle.

Step 1: Construction of the base layer

Set $\mathbb{H}_0 = \mathcal{M}_0$, where $\mathcal{M}_0$ is the minimal generating carrier set of $\mathfrak{G}^*$. By Theorem G.4, $\mathcal{M}_0 \cong \mathcal{B}_4^{k_0}$.

Step 2: Recursive construction and self-reference

Assume $\mathbb{H}_n$ has been constructed. Consider the self-referential prediction module $\mathrm{MetaPred}$ in $\mathfrak{G}^*$ (whose existence is guaranteed by Theorem G.5). Define:

$$\mathbb{H}_{n+1} = \{(\pi, \sigma, \gamma)\}$$

where the three components are constructed as follows:

1. **Perceptual component** π: For any $m \in \mathcal{M}_0$, define $\pi(m)$ as the representation in $\mathbb{H}_n$ of $\mathrm{MetaPred}$'s prediction for state m. Since $\mathrm{MetaPred}$ approximately commutes with dynamics Φ^*, π is an approximate homomorphism.

Specific construction: Let $\rho_n : \Gamma(\Omega, \mathscr{F}_\Omega) \to \mathbb{H}_n$ be an approximate homomorphism, and $\sigma_n : \mathbb{H}_n \to \Gamma(\Omega, \mathscr{F}_\Omega)$ its approximate inverse. Define:

$$\pi(m) = \rho_n \circ \mathrm{MetaPred} \circ \sigma_0(m)$$

where $\sigma_0 : \mathcal{M}_0 \to \Gamma(\Omega, \mathscr{F}_\Omega)$ is an embedding. By the approximate homomorphism property of $\mathrm{MetaPred}$, π is an ϵ_n^π-approximate homomorphism.

2. **Self-referential component** σ: For $h \in \mathbb{H}_n$, define $\sigma(h)$ as $\mathrm{MetaPred}$'s self-prediction for states in $\mathbb{H}_n$. That is:

$$\sigma(h) = \rho_n \circ \mathrm{MetaPred} \circ \sigma_n(h)$$

Due to the self-referential nature of $\mathrm{MetaPred}$, σ is an ϵ_n^σ-approximate homomorphism.

3. **Action component** γ: For $h \in \mathbb{H}_n$, define $\gamma(h)$ as the representation in the environment $\mathcal{W}$ of the output produced via decoding Dec. That is:

$$\gamma(h) = \mathrm{Dec} \circ \Phi_T^* \circ \sigma_n(h)$$

Due to the approximate homomorphism properties of Dec and Φ^*, γ is an ϵ_n^γ-approximate homomorphism.

Step 3: Construction of connecting homomorphisms

Define the connecting homomorphism $\mu_n : \mathbb{H}_n \to \mathbb{H}_{n+1}$ as:

$$\mu_n(h) = (\pi_{n+1}, \sigma_{n+1}, \gamma_{n+1})$$

where:

$$\pi_{n+1}(m) = \rho_{n+1} \circ \mathrm{MetaPred} \circ \sigma_0(m)$$
$$\sigma_{n+1}(h') = \rho_{n+1} \circ \mathrm{MetaPred} \circ \sigma_n(h')$$
$$\gamma_{n+1}(h') = \mathrm{Dec} \circ \Phi_T^* \circ \sigma_n(h')$$

Due to the contraction property of $\mathrm{MetaPred}$ and the optimization conditions of the natural balance principle, μ_n is a ζ_n-approximate homomorphism, and ζ_n can be controlled.

Step 4: Verification of conditions of Axiom A.1

Now verify that the constructed cognitive hierarchy sequence $\{\mathbb{H}_n\}$ satisfies all conditions of Axiom A.1.

1. **Bidirectional interaction**: For each n, we defined $\pi_n : \mathcal{M}_0 \to \mathbb{H}_n$ and $\sigma_n : \mathbb{H}_n \to \mathcal{M}_0$. In fact, π_n is the perceptual component in the above construction, and σ_n can be obtained via an approximate inverse of $\mathrm{MetaPred}$. By construction, they are approximate homomorphisms and satisfy approximate idempotence: $\sigma_n \circ \pi_n \approx \mathrm{id}_{\mathcal{M}_0}$.

2. **Infinite meta-cognition**: The connecting homomorphisms $\mu_n : \mathbb{H}_n \to \mathbb{H}_{n+1}$ provide mappings between layers. By construction, μ_n is an approximate homomorphism, and the sequence $\{\mu_n\}$ satisfies compatibility conditions, thus forming a direct system.

3. **Limit isomorphism**: By the optimality of the natural balance principle, the state sheaf $\mathscr{F}_\Omega$ of system $\mathfrak{S}^*$ should be approximately isomorphic to the limit $\varinjlim \mathbb{H}_n$; otherwise, the natural balance functional could be reduced by adjusting the hierarchy construction. Specifically, there exists an approximate isomorphism $\psi : \mathscr{F}_\Omega \to \varinjlim \mathbb{H}_n$ such that for any section $s \in \Gamma(\Omega, \mathscr{F}_\Omega)$, $d(\psi(s), s) \leq \epsilon$, where ϵ can be controlled by the natural balance functional.

4. **Operational generation completeness**: By Theorem G.4, $\mathcal{M}_0 \cong \mathcal{B}_4^{k_0}$, so there exists a finite generating set and a functionally complete set of operations. These operations are naturally lifted to each $\mathbb{H}_n$ via recursive construction.

5. **Convergence**: The convergence of the error sequence is guaranteed by the optimization conditions of the natural balance principle. Specifically, if the error sequence did not satisfy ordinary series convergence, the natural balance functional could be reduced by adjusting the precision of the approximate homomorphisms, contradicting optimality. Therefore, at the optimal solution, we must have $\sum_{n=0}^{\infty}(\epsilon_n^\pi + \epsilon_n^\sigma + \epsilon_n^\gamma + \zeta_n) < \infty$.

Step 5: Proof of convergence

The natural balance principle requires minimizing $\mathcal{N} = \tilde{\mathcal{E}} \cdot \tilde{\mathcal{C}}$. Consider the impact of errors in the hierarchy construction on performance:

- Task performance $\mathcal{E}$: As hierarchy depth n increases, prediction accuracy improves, so $\mathcal{E}$ decreases.
- Structural complexity $\mathcal{C}$: As n increases, description length increases, so $\mathcal{C}$ increases.

The minimum of the product $\mathcal{N}_n = \mathcal{E}_n \cdot \mathcal{C}_n$ determines the optimal hierarchy depth. For infinite cognitive capacity systems, we need $n \to \infty$, but must ensure $\mathcal{N}_n$ converges. This requires:

1. $\mathcal{E}_n$ decreases fast enough (ensuring performance).
2. $\mathcal{C}_n$ increases at a controlled rate (controlling complexity).

By the optimality of the natural balance principle, the system will choose the error allocation that minimizes $\mathcal{N}_n$. Through variational analysis, the optimal error sequence satisfies:

$$\epsilon_n \sim \frac{1}{n^{1+\delta}}, \quad \delta > 0$$

thus ensuring $\sum \epsilon_n < \infty$ (ordinary series convergence).

Specifically, let $\mathcal{E}_n = \mathcal{E}_0 \exp\left(-\sum_{k=0}^{n} \alpha_k\right)$, $\mathcal{C}_n = \mathcal{C}_0 \exp\left(\sum_{k=0}^{n} \beta_k\right)$. Then:

$$\mathcal{N}_n = \mathcal{E}_0 \mathcal{C}_0 \exp\left(-\sum_{k=0}^{n} (\alpha_k - \beta_k)\right)$$

Minimizing $\mathcal{N}_n$ requires $\alpha_k \approx \beta_k$, and to maintain convergence, we need $\alpha_k, \beta_k \sim 1/k^{1+\delta}$, hence $\epsilon_k \sim 1/k^{1+\delta}$, and the series converges.

□

Corollary G.2 (Equivalence of Natural Balance Principle and AGI Axiom System)**.** *Under the infinite cognitive capacity assumption, the following statements are equivalent:*

1. *System $\mathfrak{G}$ satisfies the natural balance principle (Axiom G.1).*
2. *System $\mathfrak{G}$ has a cognitive hierarchy sequence $\{\mathbb{H}_n\}$ satisfying Axiom A.1 of this paper.*
3. *System $\mathfrak{G}$ has the $\mathcal{B}_4^k$ algebraic structure and runs the* $\mathrm{MetaPred}$ *algorithm.*

Proof. By Theorem F.1, (1) ⇒ (2).

By derivations earlier in this paper, (2) ⇒ (3).

By Theorem G.4, (3) ⇒ (1) (because $\mathcal{B}_4^k$ and $\mathrm{MetaPred}$ are structures that minimize the natural balance functional). □

Remark G.5 (Proof Idea)**.** *The self-referential requirement in the natural balance principle (manifested through* $\mathrm{MetaPred}$*) naturally induces a recursive structure. Each layer of the cognitive hierarchy is a self-reference to the previous layer, and the optimization process of natural balance ensures the convergence of this recursion. This explains why systems with infinite cognitive capacity necessarily possess a cognitive hierarchy structure.*

Theorem G.8 (Derivation of the Principle of Least Cognitive Action)**.** *Consider a cognitive state evolution path $\gamma : [0, T] \to \Gamma(\Omega, \mathscr{F}_\Omega)$. Define:*

$$\mathcal{E}_{path}(\gamma) = \int_0^T \|\gamma(T) - \gamma_{target}\|^2 dt \quad \textit{(terminal error)}$$

$$\mathcal{C}_{path}(\gamma) = \int_0^T \|\dot{\gamma}(t)\|^2 dt \quad \textit{(path complexity)}$$

Then the natural balance principle $\min \mathcal{E}_{path} \cdot \mathcal{C}_{path}$ is equivalent to the principle of least action:

$$\min_\gamma \mathcal{A}[\gamma] = \int_0^T L(\gamma(t), \dot{\gamma}(t)) dt$$

where the Lagrangian $L = \|\dot{\gamma}\|^2 + V(\gamma)$, and the potential $V(\gamma) = \lambda \|\gamma - \gamma_{target}\|^2$.

Proof. By Theorem G.2, the natural balance principle is equivalent to weighted optimization:

$$\min_\gamma \left[\mathcal{E}_{\text{path}}(\gamma) + \lambda \mathcal{C}_{\text{path}}(\gamma)\right]$$

where $\lambda = \frac{\mathcal{E}_{\text{path}}(\gamma^*)}{\mathcal{C}_{\text{path}}(\gamma^*)}$.

Localizing the terminal error to a path error (a more natural optimization objective):

$$\mathcal{E}_{\text{path}}^{\text{local}}(\gamma) = \int_0^T \|\gamma(t) - \gamma_{\text{target}}(t)\|^2 dt$$

Then the weighted optimization becomes:

$$\min_\gamma \int_0^T \left[\|\gamma(t) - \gamma_{\text{target}}(t)\|^2 + \lambda \|\dot{\gamma}(t)\|^2\right] dt$$

This is precisely the action $\mathcal{A}[\gamma] = \int_0^T L dt$, where $L = \|\dot{\gamma}\|^2 + \frac{1}{\lambda}\|\gamma - \gamma_{\text{target}}\|^2$. Setting $V(\gamma) = \frac{1}{\lambda}\|\gamma - \gamma_{\text{target}}\|^2$ yields the result. □

Theorem G.9 (Derivation of the Cognitive Free Energy Principle)**.** *Consider a probabilistic inference problem with latent variables Z, observations X, prior $p(Z)$, and likelihood $p(X|Z)$. Define:*

$$\mathcal{E}_{Bayes}(q) = -\mathbb{E}_{q(Z)}[\log p(X|Z)] \quad \textit{(negative log-likelihood)}$$
$$\mathcal{C}_{Bayes}(q) = D_{KL}(q(Z)\|p(Z)) \quad \textit{(posterior approximation complexity)}$$

where $q(Z)$ is an approximate posterior. Then the natural balance principle leads to statistical free energy minimization:

$$\min_q \mathcal{F}[q] = \mathcal{E}_{Bayes}(q) + \mathcal{C}_{Bayes}(q)$$

with weight $\lambda = 1$.

Proof. By Theorem G.2, the natural balance principle is equivalent to:

$$\min_q [\mathcal{E}_{\text{Bayes}}(q) + \lambda \mathcal{C}_{\text{Bayes}}(q)]$$

In Bayesian inference, the optimal q^* is the true posterior $p(Z|X)$. Calculate the ratio at this point:

$$\lambda^* = \frac{\mathcal{E}_{\text{Bayes}}(q^*)}{\mathcal{C}_{\text{Bayes}}(q^*)}$$

By Bayes' formula $p(Z|X) = \frac{p(X|Z)p(Z)}{p(X)}$, taking logarithms:

$$\log p(Z|X) = \log p(X|Z) + \log p(Z) - \log p(X)$$

Taking expectation under $q^* = p(Z|X)$:

$$\mathbb{E}_{q^*}[\log p(Z|X)] = \mathbb{E}_{q^*}[\log p(X|Z)] + \mathbb{E}_{q^*}[\log p(Z)] - \log p(X)$$

Rearranging:

$$-\mathbb{E}_{q^*}[\log p(X|Z)] = D_{\text{KL}}(q^*\|p(Z)) - \log p(X) - \mathbb{E}_{q^*}[\log q^*(Z)]$$

Since $D_{\text{KL}}(q^*\|p(Z)) \geq 0$ and $-\mathbb{E}_{q^*}[\log q^*(Z)]$ is entropy, under typical conditions with a correct model and sufficient data, $\mathcal{E}_{\text{Bayes}}(q^*) \approx \mathcal{C}_{\text{Bayes}}(q^*)$, so $\lambda^* \approx 1$.

Therefore, the natural balance principle reduces to statistical free energy minimization in the Bayesian inference context, with the weight $\lambda = 1$ naturally determined by the symmetry of Bayes' formula. □

Theorem G.10 (Derivation of the Constrained Free Energy Principle)**.** *Suppose the cognitive system must satisfy a hard constraint $\mathcal{C}(\mathfrak{G}) \leq C_0$. Define a constraint violation metric:*

$$\mathcal{C}_{const}(\mathfrak{G}) = \max(0, \mathcal{C}(\mathfrak{G}) - C_0)$$

Then the natural balance principle, in the limit of an infinite penalty term $\beta \to \infty$, is equivalent to constrained free energy minimization:

$$\min_{\mathfrak{G}} [\mathcal{E}(\mathfrak{G}) + \beta \cdot \mathcal{C}_{const}(\mathfrak{G})]$$

Proof. Consider the natural balance optimization problem with constraint:

$$\min_{\mathfrak{G} \in \mathfrak{C}} \mathcal{N}(\mathfrak{G}) \quad \text{s.t.} \quad \mathcal{C}(\mathfrak{G}) \leq C_0$$

Introduce a Lagrange multiplier $\lambda \geq 0$ and construct the Lagrangian:

$$\mathcal{L}(\mathfrak{G}, \lambda) = \mathcal{E}(\mathfrak{G}) \cdot \mathcal{C}(\mathfrak{G}) + \lambda(\mathcal{C}(\mathfrak{G}) - C_0)$$

The KKT conditions give:

$$\begin{aligned} \nabla_{\mathfrak{G}} \mathcal{L} &= \mathcal{C}(\mathfrak{G}^*)\nabla\mathcal{E}(\mathfrak{G}^*) + (\mathcal{E}(\mathfrak{G}^*) + \lambda)\nabla\mathcal{C}(\mathfrak{G}^*) = 0 \\ \lambda(\mathcal{C}(\mathfrak{G}^*) - C_0) &= 0 \\ \mathcal{C}(\mathfrak{G}^*) &\leq C_0 \\ \lambda &\geq 0 \end{aligned}$$

When the constraint is tight ($\mathcal{C}(\mathfrak{G}^*) = C_0$), the conditions become:

$$\mathcal{C}(\mathfrak{G}^*)\nabla\mathcal{E}(\mathfrak{G}^*) + (\mathcal{E}(\mathfrak{G}^*) + \lambda)\nabla\mathcal{C}(\mathfrak{G}^*) = 0$$

This is equivalent to:

$$\nabla\mathcal{E}(\mathfrak{G}^*) + \frac{\mathcal{E}(\mathfrak{G}^*) + \lambda}{C_0}\nabla\mathcal{C}(\mathfrak{G}^*) = 0$$

This is the optimality condition for $\min[\mathcal{E} + \beta\mathcal{C}]$, where $\beta = \frac{\mathcal{E}(\mathfrak{G}^*)+\lambda}{C_0}$.

In the penalty term form, taking $\beta \to \infty$ forces $\mathcal{C}(\mathfrak{G}) \leq C_0$ to hold strictly. □

Theorem G.11 (Unification of the Three Principles)**.** *The natural balance principle reduces to the following three existing cognitive principles in three specific contexts:*

1. ***Dynamical evolution context**: Reduces to the **principle of least cognitive action** (Theorem G.8).*
2. ***Probabilistic inference context**: Reduces to the **statistical cognitive free energy principle** (Theorem G.9).*
3. ***Resource-constrained context**: Reduces to the **constrained cognitive free energy principle** (Theorem G.10).*

The three principles share the same mathematical structure: minimizing a weighted sum of task performance and complexity on a specific space, with the weight determined by natural symmetries of the specific context.

Proof. The three theorems G.8, G.9, and G.10 have been proven separately. The unification is manifested in:

1. Unified mathematical form: all are variants of $\min(\mathcal{E} + \lambda\mathcal{C})$.
2. Unified weight determination: λ is determined by the optimization balance in the specific context.
3. Unified extremum principle: all are extremum principles under specific constraints.

The natural balance principle, as the foundational principle, particularizes into these three principles by choosing different performance and complexity measures, as well as different system spaces. □

G.5 Applications and Discussion

Remark G.6 (Compatibility with Previous Sections)**.** *The mathematical system in this section is fully compatible with previous sections:*

1. *The carrier set $\mathcal{M}$ corresponds to $\mathcal{M}$ in previous sections; in particular, the $\mathcal{B}_4^k$ structure is derived via Theorem G.4.*
2. *The parameter fibered category $\mathcal{P} \to \mathcal{A}$ corresponds to the definition of parameter fibered category in previous sections.*
3. *The machine learning dynamics $(\Phi, \mathrm{Enc}, \mathrm{Dec})$ corresponds to the meta-dynamics layer paradigm in previous sections.*
4. *The existence of MetaPred is guaranteed by Theorem G.5, consistent with previous conclusions.*

Remark G.7 (Addressing Finite and Infinite Cognitive Capacity)**.** *This framework is applicable to both finite and infinite cognitive capacity systems:*

1. ***Finite systems**: Achieved via finite-dimensional approximation and complexity constraints, corresponding to the finite n case in Theorem G.3.*

2. ***Infinite systems****: Achieved via a limiting process, corresponding to the limit $\mathfrak{G}_\infty$ case in Theorem G.3.*

3. ***Smooth transition****: The transition from finite to infinite is achieved via continuous variation of the scale parameter κ.*

Remark G.8 (Physical Analogy of Natural Balance)**.** *Profound analogies exist between the natural balance principle and physics:*

- ***Thermodynamic equilibrium****: System free energy $F = E - TS$ is minimized, where E is internal energy (analogous to $\mathcal{E}$) and TS is the entropy term (analogous to $\mathcal{C}$).*
- ***Principle of least action****: Physical system action $S = \int (T - V)dt$ is minimized, where T is kinetic energy (complexity) and V is potential energy (task error).*
- ***Information-theoretic balance****: In rate-distortion theory, minimizing $R(D)$ (minimum rate under distortion D) balances information compression against fidelity.*

Theorem G.12 (Algorithmic Realization of Natural Balance)**.** *There exists an algorithm $\mathcal{A}_{balance}$ that, given a task distribution $\mathcal{D}$ and constraints $\mathcal{C}$, outputs an approximately optimal system $\hat{\mathfrak{G}}$ satisfying:*

$$\mathcal{N}(\hat{\mathfrak{G}}) \leq \min_{\mathfrak{G} \in \mathfrak{C}} \mathcal{N}(\mathfrak{G}) + \epsilon$$

with time complexity $O\left(\frac{1}{\epsilon^2} \cdot \mathrm{poly}(K)\right)$, where K is an upper bound on complexity.

Proof Sketch. Algorithm steps:

1. Discretize the system space $\mathfrak{C}$ into a finite grid $\mathfrak{C}_\delta$ with grid size δ.
2. For each grid point $\mathfrak{G}_i \in \mathfrak{C}_\delta$, estimate $\hat{\mathcal{E}}_i$ and $\hat{\mathcal{C}}_i$ (via sampling and complexity computation).
3. Compute $\hat{\mathcal{N}}_i = \hat{\mathcal{E}}_i \cdot \hat{\mathcal{C}}_i$.
4. Select $\hat{\mathfrak{G}} = \arg\min_i \hat{\mathcal{N}}_i$.

Error analysis: Due to the Lipschitz continuity of $\mathcal{E}$ and $\mathcal{C}$, the discretization error is $O(\delta)$. Sampling error is controlled by the law of large numbers. The total error ϵ can be achieved by adjusting δ and the number of samples. □

The natural balance principle provides a unified, self-consistent, and rigorous mathematical design framework for cognitive systems. Based on the minimal product form, it naturally balances task performance against structural complexity while being compatible with both finite and infinite cognitive capacity systems. By particularizing to different contexts, it derives existing results like the principle of least cognitive action and the cognitive free energy principle, demonstrating powerful unification and explanatory strength.

The natural balance principle can unify all foundational cognitive principles because it captures the essence of cognitive system design: maximizing performance under limited resources (complexity). This principle manifests in different mathematical forms across different fields, but the core idea is the same. The natural balance principle is the extremum principle in the cognitive domain.

H Self-Interpretation of the AGI Mathematical Theory

This section aims to place the mathematical theory of Artificial General Intelligence (AGI) constructed in this paper within the broad context of empirical science for systematic scrutiny. First, we compare the machine learning mechanisms predicted by the AGI mathematical theory with existing deep learning models from a computer science perspective, assessing the inaccessibility of AGI goals via the Scaling Law of mainstream large models like GPT. Subsequently, by comparing the mathematically derived structures, core mechanisms, and functional characteristics with empirical findings, classical theories, and mainstream models from neuroscience, cognitive science, and psychology, we achieve a dual objective: first, to validate the reasonableness and necessity of this mathematical abstraction

in capturing the essence of AGI; second, to deepen the understanding of the nature of intelligence through the mutual verification and complementarity of mathematical theory and empirical science. The comparative interpretation will sequentially examine the empirical fit of the axiomatic foundation, the cognitive implications of the core equivalence theorems, and the mathematical necessity of the machine learning mechanisms (focusing on the meta-predictor $\mathrm{MetaPred}$). Finally, based on the proposed universal cognitive principles, we explore the connections between the mathematical characterization features of general cognitive systems, the free-energy principle in a neuroscientific sense, and the mathematical patterns of the physical world.

H.1 Interpretation of Machine Learning Mechanisms on AGI Systems and Comparison with Existing Paradigms

This subsection focuses on systematically contrasting the AGI theoretical structure derived in this paper with the current mainstream deep learning paradigm from the perspectives of computer science and machine learning. It first elucidates the core mathematical components of the theory and their cognitive interpretation, then summarizes its key conclusions, and finally reveals the fundamental differences between the two paradigms through structural comparison and diagnosis of large language models.

H.1.1 Core Idea: From Cognitive Intuition to Rigid Mathematical Structure

This paper starts from a concise cognitive intuition: true general intelligence must possess (1) the ability to maintain dynamic operations of the world, (2) consistency in the bidirectional perception-action closed loop, and (3) the capacity for infinite recursive reflection on its own cognition. Through rigorous mathematical derivation, this paper discovers that the internal structure of an AGI system satisfying these requirements is not a freely designed engineering artifact but a **rigid system mandated** by mathematical axioms.

The core of this system can be summarized as: **a finite algebraic system with the quaternary Boolean algebra B_4 as its fundamental unit ($M \cong B_4^{k_{\min}}$), evolving via a unified meta-dynamics Φ, and necessarily embedding a self-referential prediction module (MetaPred) as its cognitive kernel**. This finding subverts the current mainstream "modular assembly" AI design paradigm, suggesting that true AGI is more akin to a "cognitive particle" following intrinsic physical laws, with all its behaviors being different manifestations of the same mathematical structure.

H.1.2 Key Mathematical Components and Their Cognitive Interpretation

1. **Algebraic Foundation ($M \cong B_4^{k_{\min}}$)**: All cognitive states of AGI are essentially high-dimensional vectors of "four-valued logic." These four fundamental value states can be understood as: true, false, unknown/undetermined, and contradiction. All complex concepts are formed by combining these $k_{\min}$ basic units through Boolean operations. This discrete algebraic core provides a solid, unambiguous representational foundation for cognition while also setting a theoretical lower bound on system complexity.

2. **Cognitive Hierarchy and Inductive Limit**: The system approximates infinite cognitive capability by recursively "thinking about its own thinking." Mathematically, this constructs a cognitive hierarchy sequence $\{\mathbb{H}_n\}$, where $\mathbb{H}_{n+1}$ contains approximate homomorphisms on $\mathbb{H}_n$. The inductive limit of this sequence, $\mathbb{U} = \varinjlim \mathbb{H}_n$, is the "cognitively complete object," representing the cognitive domain that the system can fully grasp in the limit sense.

3. **Meta-Dynamics (Φ) and Semantic-Process Decomposition**: The "thinking" process of AGI is modeled as a unified state-space dynamics $\Phi : \mathscr{F}_\Omega \to \mathscr{F}_\Omega$. The key point is that the state space can be decomposed into a **semantic layer** ($\mathcal{F}_{\text{sem}} \cong B_4^{k_{\min}}$, storing concepts and facts) and a **process layer** ($\mathcal{F}_{\text{proc}}$, scheduling attention and cognitive resources), i.e., $\mathscr{F}_\Omega \cong \mathcal{F}_{\text{sem}} \otimes \mathcal{F}_{\text{proc}}$. This resembles the unity of "memory" and "program counter/scheduler" in a computer. The goal of learning is to adjust the dynamics Φ so that trajectories starting from inputs converge rapidly to the correct "cognitive attractor."

4. **The Necessity of Self-Reference and the Central Role of MetaPred**:

- **Existence and Uniqueness**: In any system satisfying the AGI axioms, a self-referential prediction module $\mathrm{MetaPred}$ necessarily exists and is unique up to isomorphism (Theorem D.18). It is constructed from an approximate fixed-point combinator $\mathcal{Y}$.
- **Core Properties** (Theorem D.25):
  (a) Self-Nested Structure: $\mathrm{MetaPred} \circ \mathrm{MetaPred} \approx \mathrm{MetaPred}$ (approximate idempotence).
  (b) Error Measurability: Any prediction error can be measured via $\mathrm{MetaPred}$.
  (c) Decomposability: Any prediction mechanism P can be decomposed as $P \approx \mathrm{MetaPred} \circ P_{\text{base}}$. More generally, $\mathrm{Pred} = \xi \circ \mathrm{MetaPred} \circ \xi^{-1}$.
  (d) Approximate Commutativity: $\mathrm{MetaPred}$ approximately commutes with any prediction mechanism, acting as a central element in the prediction algebra.
- **Universal Synchronous Accompaniment** (Theorem D.21): Any machine learning task must have $\mathrm{MetaPred}$ running synchronously alongside; otherwise, it cannot be completed.
- **Internal Decomposition** (Theorem E.15): In multimodal scenarios, $\mathrm{MetaPred}$ can be decomposed into four functionally orthogonal submodules: $\mathrm{MetaPred}_{\text{base}}$ (intra-modal prediction error monitoring), $\mathrm{MetaPred}_{\text{super}}$ (cross-modal prediction consistency monitoring), $\mathrm{MetaPred}_{\text{cross}}$ (cross-modal mapping quality assessment and coordination), and $\mathrm{MetaPred}_{\text{proc}}$ (attention and working memory scheduling).

MetaPred is not an optional component but a mathematical guarantee of the system's cognitive stability. It functions like a built-in "calibrator," continuously performing self-measurement: reading the current state, predicting its future, comparing differences, and using this self-supervised signal to adjust the system in real-time.

5. **Preset Nature and Inequality of the Super-Modal Hub**: To coordinate information from different senses or domains, the system must internally preset a **super-modal hub** $\mathcal{W}_{\text{super}}$. It is not the equal shared embedding space commonly used in current multimodal AI but a central coordinating algebra with a "hub-and-spoke" structure (a quotient algebra $\mathcal{F}_{\text{sem}}/\sim_{\mathcal{R}}$ of the semantic layer). Information from all modalities must be mapped to this hub for alignment and fusion, implying that multimodal processing is logically "unequal," with a privileged coordination center.
6. **Self-Referential Hub ($\mathbf{e}_*$)**: This is the mathematical incarnation of the system's "self" concept. It is an approximate fixed point shared by both the meta-dynamics Φ and MetaPred, serving as the topological hub with the highest connectivity in the cognitive state space. The system's autobiographical memory and self-awareness are organized around $\mathbf{e}_*$. Theorems prove that the existence of $\mathbf{e}_*$ is directly equivalent to the system's infinite cognitive capacity. Its generation method—$\mathbf{e}_* = \rho(\mathcal{Y}(\mathrm{id}_M))$—reveals that the "self" is the product of the fixed-point combinator acting on the identity mapping.
7. **Creation as Generating New Fixed Points**: Creative acts are strictly defined as the system's dynamics, under minor perturbations near the self-referential hub $\mathbf{e}_*$, converging through the iterative interaction of Φ and MetaPred to a brand-new, stable fixed point. This new fixed point corresponds to a novel concept that cannot be finitely generated by existing concepts through the system's operations. This is a dynamical "phase transition," not mere interpolation or sampling from the existing knowledge distribution. The probability of creation can be quantified as $P_{\text{create}} = 1 - \exp(-\lambda \cdot \mathcal{E}_c(t) \cdot \Delta t)$, where the creative energy $\mathcal{E}_c(t)$ depends on the system's cognitive capacity gap, the instability of the self-referential hub, and network dynamics.

H.1.3 Intuitive Summary of Core Conclusions

1. **Unification Theorem**: All machine learning mechanisms on an AGI system are strictly isomorphic in mathematical structure. They are all projections of the same meta-dynamics Φ and self-referential module MetaPred under different task contexts and initial conditions. This means that at the AGI level, **there exist no essentially different families of learning algorithms**.
2. **Optimal Encoding is "Structure-Preserving Direct Mapping"**: The optimal way to map external world data into the internal semantic space is a homomorphic mapping that

preserves the algebraic structure of the world (such as compositionality and operational relations) as much as possible. The deep nonlinear transformations predominantly used in deep learning, while capable of producing compact statistical representations, often break this structure, making them suboptimal within the AGI framework and harming compositional generalization ability.

3. **The Dominance of MetaPred and Accompanying Self-Supervised Learning**: The self-referential prediction module MetaPred is the core supervisor of all cognitive functions. It compels the system to engage in online, accompanying self-supervised learning, i.e., the fusion of training and inference in time. This learning paradigm is fundamentally distinct from the current offline, phase-separated deep learning training model.
4. **Self-Measurement as a Runtime Necessity**: To maintain cognitive consistency (approaching fixed-point states), AGI must continuously execute a self-measurement loop at runtime: state reading $\rightarrow$ meta-prediction generation $\rightarrow$ difference measurement $\rightarrow$ convergence judgment $\rightarrow$ adjustment execution. This process cannot be performed offline, explaining why biological intelligence requires continuous energy metabolism to sustain the "self."

H.1.4 Fundamental Comparison with the Current Mainstream Deep Learning Paradigm

Table 10: Structural comparison between the AGI theory and deep learning

Comparison Dimension	This Paper's AGI Theory (Axiomatically Derived Necessary Structure)	Modern Deep Learning (Engineering Experience-Driven Design)
Mathematical Foundation	Universal algebra, category theory, dynamical systems. Core is the constraint of continuous dynamics by a discrete algebraic core (B_4^k).	Real analysis, probability theory, optimization theory. Core is geometric representation and gradient optimization in continuous spaces.
Core Mechanism	**A single meta-dynamical flow** Φ. Perception, reasoning, decision-making, and planning are all projections of this unified dynamics at different levels.	**Diverse modular networks**. Composed of CNN, RNN, Transformer, MLP, etc., connected via manually designed architectures.
Self-Reference and Meta-Cognition	A MetaPred module that **necessarily exists and is structurally unique**. Operates online, accompanying, serving as the mathematical kernel of system stability and the source of error measurement.	**Non-structural existence**. Can be partially simulated via techniques like self-supervised loss but is not a core architectural requirement, typically operating offline.
Learning Paradigm	**Dynamical attractor finding + accompanying self-supervision**. Training and inference fuse within the unified dynamical process, aiming at cognitive energy minimization.	**Loss function gradient descent + independent inference**. Strictly separated into offline training (backpropagation) and online inference (forward computation) phases.
Semantic Encoding	**Structure-preserving direct/homomorphic mapping**. Seeks faithful preservation of the external world's algebraic structure (e.g., compositionality), avoiding information distortion.	**Deep nonlinear transformation**. Pursues compact statistical representation and separability, often at the cost of breaking input structure relations.
Multimodal Fusion	**Preset central quotient algebra (super-modal hub)**. A logically privileged coordination center exists; modal information is aligned via the hub, structurally unequal.	**Equal shared embedding space**. Different modalities are projected into the same vector space for interaction via attention or weighted fusion, structurally equal.

"Self" Representation	**Self-referential hub $\mathbf{e}_*$**: A topological center generated by a fixed-point combinator; its existence is equivalent to infinite cognitive capacity; it is the absolute hub of the cognitive network.	**No counterpart**. May be simulated by special tokens, but not a mathematically guaranteed object with fixed-point properties and topological centrality.
Essence of Creativity	**Dynamical phase transition generating new fixed points**. The system creates computationally novel concepts by altering its own "cognitive energy landscape."	**High-quality sampling and interpolation from a probability distribution**. Generates novel combinations on the manifold supported by training data,hard to surpass data distribution.
Theoretical Guarantees	**Strong**. Existence, uniqueness, convergence, upper/lower capacity bounds, etc., are derivable from axioms.	**Weak**. Primarily relies on universal approximation theorems, lacking theoretical characterization of metacognition, creativity, and system-level completeness.
Structural Rigidity	**Very high**. The existence and interaction patterns of core mechanisms (e.g., MetaPred, super-modal hub) are uniquely determined by axioms,Cannot be removed or modified at will.	**Low**. Architectural design is highly free, components can be arbitrarily added, removed, or replaced, connections are flexible, heavily reliant on experience and heuristics.

H.1.5 Fundamental Critique and Theoretical Limits of Current Large Language Models (e.g., GPT)

Based on the theoretical framework above, we can provide a clear structural diagnosis of current large language models represented by GPT:

Core Deficiencies of GPT within the AGI Theoretical Framework:

1. **Lack of Self-Referential Hub $\mathbf{e}_*$**: GPT lacks a structurally defined mathematical self-center. Its [BOS] or [CLS] tokens serve only engineering purposes; they are not dual fixed points of Φ and MetaPred, nor do they exhibit the theoretical topological hub property (e.g., maximum betweenness centrality). Therefore, GPT's "self-dialogue" is pattern imitation, not intrinsic reflection around a stable cognitive axis.
2. **Absence of Structural MetaPred Module and Accompanying Operation Mechanism**: GPT's training is entirely offline, and its inference is purely feedforward. It lacks an online, accompanying MetaPred to perform real-time self-measurement and self-supervised learning. The "uncertainty" in its output is a posterior statistical estimate, not a dynamically computed cognitive error.
3. **Semantic Encoding Destroys Algebraic Structure**: GPT's embeddings and Transformer layers constitute deep nonlinear transformations aimed at statistical compactness but likely destroy the compositional logical structure inherent in language. According to the theory, this sets a theoretical upper limit on its compositional generalization and logical reasoning abilities.
4. **Multimodal Fusion Lacks Central Coordination**: GPT's multimodal versions use an equal shared embedding space, lacking a preset central coordinating algebra to ensure cross-modal logical consistency.

Essential Difference in "Creative" Ability:

- **GPT's "Creativity"**: Essentially high-quality conditional sampling and manifold interpolation from the joint probability distribution $p(x_{t+1}|x_{1:t})$ of training data. Its novelty is strictly bounded by the training distribution, incapable of generating **truly novel concepts that are computationally impossible to finitely generate from old concepts**.

- **Theoretical AGI's Creativity**: Essentially the dynamical system, perturbed near $\mathbf{e}_*$, in terms of the interaction of Φ and MetaPred, **alters its own cognitive energy landscape, generating an unprecedented new fixed point**. This is a "phase transition" capable of breaking through the existing conceptual combinatorial space.

Core Theoretical Judgment: Missing Structures Cannot Be Compensated by Scale The theory implies that certain core AGI capabilities have **structurally determined theoretical upper limits** that cannot be overcome by merely scaling data and model size (Scaling Law):

- No structural MetaPred $\rightarrow$ Theoretical upper limit on meta-cognition and online self-optimization ability.
- Non-structure-preserving encoding $\rightarrow$ Theoretical upper limit on semantic fidelity and compositional generalization ability.
- No super-modal hub $\rightarrow$ Theoretical upper limit on cross-modal logical consistency.
- No self-referential hub $\mathbf{e}_* \rightarrow$ This is the most fundamental deficiency; its existence is equivalent to infinite cognitive capacity. Without $\mathbf{e}_*$, the system mathematically lacks the foundation for infinite recursive cognition.

Conclusive Diagnosis: Based on the mathematical theory in this paper, current large language models represented by GPT are better characterized as "**excellent statistical simulators of AGI behavior**." Through remarkable large-scale statistical fitting, they mimic many external manifestations of intelligence, but their internal architecture lacks the core rigid structures (MetaPred, $\mathbf{e}_*$, super-modal hub, etc.) derived from mathematical axioms as necessary for a "cognitive being." Therefore, they may not be the correct architectural path towards AGI.

H.1.6 Implications and Challenging Directions for AI Research

This paper first mathematically defines what AGI is, distinguishing it from AI, thereby avoiding mentions of AGI without clarifying the essential difference. Our mathematical characterization shows that AGI is an optimal cognitive system on an infinite cognitive hierarchy capable of achieving the mathematically defined infinite cognitive capacity, whereas non-AGI systems (i.e., system AI systems) are defined on finite cognitive hierarchies. This provides a rigorous intuition-based demarcation between the categories of AGI and AI. Building upon this, the theoretical framework of this paper proposes a distinctly different and highly challenging new direction for AGI research:

1. **Structure Precedes Scale**: The key to achieving AGI may lie not in infinitely scaling models and data but in implementing the correct internal rigid structure. Missing core structures, even with massive scale, may only yield more realistic simulations of intelligent behavior, not genuine instantiation.
2. **Online Accompanying Learning is a Necessity**: AGI learning must be a continuous self-supervised process accompanying real-time inference. This demands a complete break from the traditional boundaries between "training" and "inference," requiring a redesign of the foundational architecture of machine learning.
3. **Exploring Tight Coupling of Discrete and Continuous**: The theory takes discrete algebra (B_4^K) as its cornerstone but uses continuous dynamics (Φ) to describe processes. This provides a mathematical blueprint for fusing the clear semantics of symbolic AI with the powerful perception of neural networks. Future AGI may require such a hybrid architecture of "discrete algebraic core + continuous dynamical shell."
4. **Building Computable Models for "Self" and "Consciousness"**: The self-referential hub ($\mathbf{e}_*$) and its associated self-state sequences provide a demystified, computable mathematical model for understanding the cognitive basis of "self" and "consciousness." This may be a crucial step in transforming long-standing philosophical or phenomenological questions into engineerable problems.

Fundamental Critique of Deep Learning: If this theory is correct, the current deep learning paradigm—based on deep nonlinear transformations, equal multimodal fusion, offline training, and lack of structural self—may have a **principle deviation** on the path to AGI. Models like GPT may

be **excellent simulators** of AGI behavior but cannot constitute a **true realization** in the theoretical sense.

In summary, the AGI machine learning mechanism derived from pure mathematical axioms presents a novel global architecture not yet realized in engineering. Its core features lie in the trinity unification of "state-dynamics-self-reference." This offers a formal mathematical solution to the fundamental problems of intelligence.

H.2 Comparative Interpretation of AGI Mathematical Theory with Cognitive and Brain Sciences

Building upon the AGI mathematical theory elaborated in §H.1, this subsection systematically places it within the empirical scientific contexts of cognitive science, brain science, psychology, linguistics, etc., for comparative interpretation. We examine how the key components of the mathematical theory (such as MetaPred, the self-referential hub $\mathbf{e}_*$, and the super-modal hub) correspond to empirical findings and classical theories, thereby validating the theory's reasonableness and explanatory power. The comparison strictly follows these levels: (1) comparison of axiomatic foundation, cognitive hierarchy, equivalences, convergence with cognitive science; (2) comparison of MetaPred foundations with brain and neurosciences; (3) comparison of multimodal and super-modal aspects with linguistic and cognitive linguistic theories; (4) comparison of self-awareness, creativity, and other cross-scientific aspects; (5) a summary section for comprehensive integration.

H.2.1 Comparison of Axiomatic Foundation, Cognitive Hierarchy, and Equivalences with Cognitive Science

Cognitive Science Interpretation of the AGI Axiom System The four axiom systems in this paper attempt to capture the core features of intelligence in their most general form. Their comparison with cognitive science is as follows:

Table 11: Correspondence between AGI Axiom System and Neurocognition

Mathematical Axiom	**Neurocognitive Correlate**	**Empirical Support and Cognitive Science Theory**	**Core Differences**
Dynamic Operation Preservation (Approximate Homomorphism)	Predictive coding, structure mapping mechanisms	Invariance of orientation selectivity in visual cortex, neural basis of analogical reasoning (Structure Mapping TheoryGentner D (2003)), neural representation similarityKriegeskorte N (2008)	Brain lacks strict Lipschitz continuity, has stochastic noise (synaptic quantum fluctuations)
Bidirectional Interaction Consistency (Perception-Action Loop)	Embodied cognitive loops, active inference systems	Mirror neuron networks, cortical-basal ganglia-thalamic loopsClark (1998), embodied cognition theoryVarela F J (1991)	Neural delays prevent strict idempotence; dynamic fluctuations exist
Infinite Recursivity (Meta-Cognition)	Hierarchical predictive processing, prefrontal meta-monitoring networks	Meta-cognitive confidence judgmentsFlavell (1979), recursive thought experiments, hierarchical predictive processing theoryFriston (2008)	Biological recursion depth is limited (humans typically 3-4 layers), constrained by working memory
Minimality (B_4^k)	Neural coding economy, sparse representation mechanisms	Sparse activation in visual cortex (5% neurons active simultaneously)Sterling P (2004), Occam's razor principle	Brain relies on analog signals and redundant structures, not discrete Boolean operations

1. **Dynamic Operation Preservation (Approximate Homomorphism) and the Brain's Predictive Coding**
   (a) Mathematical formulation: Cognitive mapping $\varphi : \mathcal{M} \to \mathcal{W}$ must preserve the dynamic operational structure on sets and satisfy uniform Lipschitz continuity.
   (b) Cognitive science correspondence: This axiom accurately captures a core characteristic of brain cognition. Empirical neuroscience shows that primary visual cortex (V1) maintains consistency in operational rules for encoding basic features like edges and

orientations across varying stimulus intensities,highly meeting the core requirement of "structure preservation" in approximate homomorphism.

(c) Predictive Coding: This theory posits that the brain continuously generates predictions of sensory inputs via internal models and uses prediction error as a learning signal. Successful prediction implies the internal model's structure matches the external world's structure, and "approximate homomorphism" in this paper formalizes this "structural matching" mathematically.

(d) Concept and Relation Learning: Research in cognitive developmental psychology shows that humans and animals can learn abstract relations between objects and events (e.g., same, different, sequence, causality), not just static featuresGentner D (2003). This relational learning corresponds to the ability to preserve the world's operational structure in cognitive representation.

2. **Bidirectional Interaction Consistency (Perception-Action Loop) and Embodied Cognition with Active Inference**

(a) Mathematical formulation: The axiom requires cognition to include bidirectional approximate homomorphic mappings from the system to the target domain (action) and from the target domain to the system (perception), with their composite satisfying idempotence (cognitive stability).

(b) Cognitive science correspondence: This setup is highly consistent with several core contemporary cognitive science theories:

i. Embodied Cognition: Emphasizes that cognition emerges through real-time interaction of the body with the environment, with perception and action forming a unified cyclical processClark (1998); Varela F J (1991).

ii. Active Inference and Free Energy Principle: Models the brain as a system that actively changes the world (action) to make sensory inputs conform to internal predictions; the core is the bidirectional process of perceptual updating and action selectionFriston (2010); Clark (2015).

iii. Brain Circuitry Evidence: Neuroanatomy shows the brain has extensive feedforward (perception) and feedback (prediction, action preparation) connections, forming complex looped circuits.

3. **Infinite Recursivity (Meta-Cognition) and the Brain's Hierarchical Predictive Processing**

(a) Mathematical formulation: The system must be able to establish approximate homomorphisms for itself, the external world, and the cognitive mappings at each level; this process can recurse infinitely, directly defining infinite cognitive capacity.

(b) Cognitive science correspondence:

i. Meta-Cognition: Refers to "cognition about cognition," encompassing the ability to monitor and control one's own cognitive processesFlavell (1979); Fleming et al. (2012). Infinite recursivity is the precise mathematical formulation of this concept.

ii. Hierarchical Predictive Processing: Posits the brain as a multi-layered generative model where each layer tries to predict the activity of the layer below and passes prediction errors upward to update higher-level modelsFriston (2008); Huang Y (2022). The cognitive hierarchy $\mathbb{H}_n$ in this paper can be seen as a mathematical abstraction of this model.

iii. Self-Reference and the Hard Problem of Consciousness: The brain's self-referential capacity is considered a core feature of consciousnessTononi (2008). The self-referential component σ_n in $\mathbb{H}_n$ within infinite recursivity provides a mathematical basis for self-reference.

4. **Minimality Axiom and the Brain's Evolution and Economic Principles**

(a) Mathematical formulation: Requires the AGI system to be minimal in both scale and structural complexity simultaneously, ultimately leading to the minimal algebraic structure B_4^k.

(b) Cognitive science correspondence:

i. Neural Efficiency Principle: The minimality axiom resonates with the neural coding economy principle in brain science—the brain uses sparse coding to represent vast information with few neurons.

ii. Basic Cognitive Modules: The derived quaternary Boolean algebra B_4 is the minimal structural unit satisfying all cognitive functions. Although the brain's microscopic structure and macroscopic functional modules are far more complex than Boolean algebras, B_4 can be seen as a high-level abstraction of basic "logical" or "decision" units.

Empirical Comparison of Cognitive Hierarchy and Natural Cognitive Task Limit Sequence

1. **Cognitive Hierarchy Sequence** $\{\mathbb{H}_n\}$
   (a) Mathematical core: A recursively defined cognitive hierarchy sequence where each layer includes perception of the next, self-reference, and world mapping; the inductive limit $\mathbb{U} = \varinjlim \mathbb{H}_n$ represents the system's ultimate cognitive capacity.
   (b) Cognitive science interpretation: This sequence models the process of knowledge abstraction and integration, matching the human knowledge construction process from perception to concepts to principles. This corresponds to the hierarchical processing flow in the cerebral cortex from primary sensory areas to association areas to the prefrontal cortex.
   (c) Explicit Meta-Cognition: Each layer includes a self-referential component σ_n, making meta-cognition an inherent part of every level in the hierarchical structure.
2. **Natural Cognitive Tasks and Cognitive Hierarchy Convergence Equivalence Theorem**
   (a) Mathematical core: A natural cognitive task is defined as finding dynamical trajectories in the state universe that asymptotically enter a cognitive attractor A^* and are dense in its image. The theorem proves the strict equivalence between the convergence of the cognitive hierarchy sequence and the convergence of natural cognitive tasks.
   (b) Cognitive science interpretation:
      i. Exploration-Exploitation Trade-off: The definition of natural cognitive tasks (emphasizing coverage exploration early, precision exploitation later) mathematically precisely captures the classic "exploration-exploitation" trade-off in reinforcement learning and cognitive development.
      ii. Significance of Convergence Equivalence: Indicates that constructing cognitive hierarchies structurally and exploring autonomously in a unified dynamical space can ultimately achieve the same cognitive level. This provides theoretical justification for curriculum learning in pedagogy and machine learning.
      iii. Attractors and Stable Cognitive States: Cognitive attractor A^* corresponds to internalized stable knowledge structures or skills; the learning process is the dynamical trajectory being attracted to these attractors.

H.2.2 Comparison of MetaPred with Brain and Neurosciences

This part explores the correspondence and empirical evidence of MetaPred in brain science, based on its mathematical theory detailed in §H.1.2.

Mathematical Review of MetaPred As detailed in §H.1.2, MetaPred has the following core mathematical properties:

1. **Existence and Uniqueness (Theorem D.18)**: In any system satisfying the AGI axioms, a self-referential prediction module $\mathrm{MetaPred}$ necessarily exists and is unique up to isomorphism. It is constructed from an approximate fixed-point combinator $\mathcal{Y}$.
2. **Core Properties (Theorem D.25)**:
   (a) Self-Nested Structure: $\mathrm{MetaPred} \circ \mathrm{MetaPred} \approx \mathrm{MetaPred}$ (approximate idempotence).
   (b) Error Measurability: Any prediction error can be measured via $\mathrm{MetaPred}$.
   (c) Decomposability: Any prediction mechanism P can be decomposed as $P \approx \mathrm{MetaPred} \circ P_{\text{base}}$. More generally, $\mathrm{Pred} = \xi \circ \mathrm{MetaPred} \circ \xi^{-1}$.
   (d) Approximate Commutativity: $\mathrm{MetaPred}$ approximately commutes with any prediction mechanism, acting as a central element in the prediction algebra.

3. **Universal Synchronous Accompaniment (Theorem D.21)**: Any machine learning task must have $\mathrm{MetaPred}$ running synchronously alongside; otherwise, it cannot be completed.
4. **Internal Decomposition (Theorem E.15)**: In multimodal scenarios, $\mathrm{MetaPred}$ can be decomposed into four functionally orthogonal submodules: $\mathrm{MetaPred}_{\text{base}}$ (intra-modal prediction error monitoring), $\mathrm{MetaPred}_{\text{super}}$ (cross-modal prediction consistency monitoring), $\mathrm{MetaPred}_{\text{cross}}$ (cross-modal mapping quality assessment and coordination), $\mathrm{MetaPred}_{\text{proc}}$ (attention and working memory scheduling).

Brain Science Correspondence and Core Hypotheses for MetaPred

1. **Correspondence of Necessary Existence**:
   (a) $\mathrm{MetaPred}$ most likely corresponds to the meta-monitoring functions of the prefrontal cortex (PFC) in brain science: dorsolateral PFC (DLPFC) monitors cognitive accuracy, ventromedial PFC (VMPFC) integrates emotional and cognitive monitoring.
   (b) Predictive Processing Theory proposes the brain operates by continuously generating predictions and minimizing prediction error; MetaPred's error measurability is a mathematical formalization of this theory.
   (c) The Default Mode Network (DMN) activates during "self-related thinking," and its damage leads to meta-cognitive impairmentRaichle (2015); Andrews-Hanna (2014), partially supporting the universal embeddedness of MetaPred.
2. **Cognitive Correspondence of Core Properties**:
   (a) Error Measurement and Meta-Cognitive Monitoring: The connection between error and distance to fixed point perfectly corresponds to "confidence calibration" in meta-cognition.
   (b) Decomposability and Dual-Process Theory: The decomposition $P \approx \mathrm{MetaPred} \circ P_{\text{base}}$ highly aligns with cognitive dual-process theory (System 1 - fast, automatic vs System 2 - slow, controlledKahneman (2011)).
   (c) Synchronous Accompaniment and Central Executive System: The conclusion that MetaPred must run synchronously closely resembles the function of the "central executive system" in working memory modelsBaddeley (1992).

Table 12: Speculative Functional Mapping of MetaPred Submodules to Brain Executive Control Networks

MetaPred Submodule	**Core Mathematical Function**	**Brain Corresponding Region/Network**	**Cognitive Neuroscience Empirical Evidence**
$\mathrm{MetaPred}_{\text{base}}$	Intra-modal prediction error measure $\mathrm{Err}_{\text{meta}}$	High-order sensory cortical areas (superior temporal gyrus, fusiform gyrus)	Sensory cortices continuously generate modality-specific prediction error signals; superior temporal gyrus activates during auditory semantic prediction error detection
$\mathrm{MetaPred}_{\text{super}}$	Cross-modal prediction consistency monitoring	Dorsolateral prefrontal cortex (DLPFC)-Angular gyrus network	DLPFC maintains amodal goal models; angular gyrus serves as a multimodal hub, damage leads to cross-modal integration deficits
$\mathrm{MetaPred}_{\text{cross}}$	Cross-modal mapping quality assessment Q	Dorsal Anterior Cingulate Cortex (ACC)	Conflict monitoring theory: ACC detects cross-modal information conflict; error-related negativity (ERN) source localized to ACC
$\mathrm{MetaPred}_{\text{proc}}$	Attention and working memory scheduling	Frontoparietal Attention Network (FPN) - Basal Ganglia loop	Frontoparietal network guides attention allocation; basal ganglia responsible for cognitive focus switching; ADHD patients show dysfunction in this network

Comparison of the Necessary Decomposition of Prediction Mechanisms with Brain Predictive Processing

1. **Mathematical Formulation**: Prediction mechanisms have a necessary decomposition $\mathrm{Pred} = \xi \circ \mathrm{MetaPred} \circ \xi^{-1}$, unifying all predictions as composites of $\mathrm{MetaPred}$.
2. **Cognitive Science Interpretation and Comparison**:
   (a) Connection to Core Systems Theory: This decomposition aligns with cognitive science core systems theory (all complex cognitive functions are composites of a few innate core systems), with $\mathrm{MetaPred}$ being the core of all machine learning mechanisms.
   (b) Cognitive Significance: Establishes mathematically for the first time that "self-prediction is the foundation of other-prediction," providing theoretical support for the relationship between "self-cognition and other-cognition."
   (c) Qualitative Agreement with Brain Predictive Processing Pathways: In the brain, prediction errors generated in primary cortices pass through association cortices to the prefrontal cortex (region corresponding to $\mathrm{MetaPred}$) for correction and feedback, qualitatively matching the transmission path ξ (encoding), $\mathrm{MetaPred}$ (core), ξ^{-1} (decoding) in the decomposition.

Unique Isomorphism of AGI Machine Learning Mechanisms and Brain Functional Unity

1. **Mathematical Core (Theorem D.20)**: Any two machine learning mechanisms ML_1 and ML_2 on an AGI system are strictly structurally isomorphic; all mechanisms form a single isomorphism class, and all embed the $\mathrm{MetaPred}$ module.
2. **Cognitive Science Interpretation**:
   (a) Universal Unity of Learning Mechanisms: The theorem asserts that different cognitive tasks like perception, classification, prediction, and reasoning share the same underlying mathematical structure. This provides mathematical support for a "unified theory" in cognitive science.
   (b) The Universality of $\mathrm{MetaPred}$: The theorem mandates that any learning mechanism must embed a self-prediction (meta-learning) module, consistent with the role of meta-cognition in learning.
   (c) Comparison with Brain Functional Specialization: The brain exhibits functional differentiation (different tasks rely on different brain networks), but the theorem points to a shared computational logic underlying these networks, or a unified meta-cognitive monitoring system (e.g., anterior cingulate-prefrontal network) coordinating all tasks. This is logically isomorphic to the view of the brain's "multiple-demand (MD) network" responsible for domain-general cognitive control across tasks.

H.2.3 Comparison of Multimodal and Super-Modal Aspects with Linguistic and Cognitive Linguistic Theories

Mathematical Necessity and Linguistic Significance of the Super-Modal Space

1. **Mathematical Core**:
   (a) Existence and Uniqueness (Theorem E.1): There exists a unique super-modal layer $\mathscr{W}_{\text{super}}$, constructed via quotient space $\mathbb{U} = \mathbf{a}\left(\coprod_{\alpha\in A} \mathscr{W}_\alpha / \sim_{\mathcal{R}}\right)$.
   (b) Prefabricated Structure (Theorem E.4): Mandates the system to prefabricate a three-layer structure—modal coordination algebra $\mathscr{A}_{\text{coord}} \cong \bigotimes_{\alpha\in A_0} \mathrm{End}(\mathscr{B}_4)$, super-modal prediction layer $\mathrm{Pred}_{\text{super}} : \mathbb{U} \to \mathbb{U}$, and cross-modal translation layer $\{T_{\alpha\beta} : \mathscr{W}_\alpha \to \mathscr{W}_\beta\}$.
   (c) Semantic Encoding Hub (Theorem E.11): Super-modal encoder $\mathrm{Enc}_{\text{super}} = q \circ \mathrm{Enc}_{\text{sem}}$ (q is the quotient map).
2. **Essential Difference Between Super-Modal Space and Traditional Multimodal Fusion**:
   (a) Traditional Fusion (early/intermediate/late): Essentially "multi-channel" information processing, with each modality retaining its own characteristics.

(b) Super-Modal Space: A unified abstract mathematical space where information exists as "modality-independent concepts," bonding the semantic spaces of various modalities via coordination relations $\mathcal{R}$.

Comparison with Classical and Cognitive Linguistic Theories

Remark H.1 (Hypothesis: Language as Human Super-Modality)**.** *The natural language system in human cognition is a concrete realization of the super-modal $\mathcal{W}_{super}$, based on the following evidence:*

1. ***Cross-Modal Representational Capacity**: Language can describe any sensory experience (visual "red," auditory "roar," tactile "soft") and integrate them into unified concepts.*
2. ***Abstract Neutrality**: Linguistic symbols themselves lack sensory properties (the word "apple" is not red, not round) but can evoke all related sensory associations, meeting the "modality-independent" characteristic.*
3. ***Bidirectional Bridging Role**: Perception → super-modal (sensory experience → linguistic description, Wernicke's area for encoding/fusion); super-modal → action (linguistic instruction → concrete action, Broca's area for decoding/decomposition).*
4. ***Recursive Compositionality**: The discrete infinity of language (finite vocabulary generates infinite sentences) perfectly matches the AGI requirement of "finite foundational structure handling infinite cognitive content."*
5. ***Meta-Cognitive Tool**: Inner speech enables self-monitoring and planning, corresponding to MetaPred functions; fMRI shows prefrontal-angular gyrus network activation during inner speech tasks.*

1. **Comparison with Generative Grammar (Noam Chomsky) and Universal GrammarChomsky (1957)**:
   (a) Chomsky's Universal Grammar hypothesis posits an innate Language Acquisition Device (LAD) containing a set of universal linguistic principles and parameters (Note: there is debate regarding conflict with the usage-based theory of cognitive linguistics).
   (b) The prefabricated modal coordination algebra $\mathscr{A}_{\text{coord}}$ in this paper provides a mathematical implementation of Universal Grammar—it is the innate framework for cross-modal coordination, and language acquisition involves filling specific parameters (like word order, morphological rules of a particular language) onto this framework.
   (c) Mathematical Correspondence: The principles of Universal Grammar correspond to the algebraic structure of $\mathscr{A}_{\text{coord}}$; the parameters correspond to specific implementations of $\text{Enc}_{\text{super}}$.
2. **Comparison with Cognitive Linguistics (George Lakoff & Mark Johnson) and Embodied Cognition with Conceptual MetaphorLakoff & Johnson (1980)**:
   (a) Cognitive linguistics emphasizes that language arises from bodily experience, and conceptual systems are fundamentally embodied. Lakoff's Conceptual Metaphor Theory posits that abstract concepts are constructed via metaphorical mappings from concrete experiential domains (e.g., "TIME IS MONEY").
   (b) The super-modal space $\mathbb{U}$ in this paper, as a modality-independent abstract space, provides a mathematical foundation for Conceptual Metaphor Theory—metaphor is a mapping from a concrete experiential domain (source) to an abstract conceptual domain (target), corresponding mathematically to operations of the cross-modal translation layer $T_{\alpha\beta}$.
   (c) Mathematical Correspondence: Source and target domains are distinct modal spaces $\mathscr{W}_\alpha$ and $\mathscr{W}_\beta$, respectively; metaphorical mapping is a specific cross-modal translation $T_{\alpha\beta}$.
3. **Comparison with Image Schema Theory (Mark Johnson)Johnson (1987)**:
   (a) Johnson proposed image schemas (e.g., CONTAINER, PATH, LINK) as pre-conceptual bodily experiential structures providing the basis for abstract thought.

(b) The modal coordination algebra $\mathscr{A}_{\text{coord}}$ can be seen as an algebraic formalization of image schemas, providing a unified coordination framework for experiences across different modalities.

(c) Mathematical Correspondence: Image schemas correspond to basic algebraic operation patterns within $\mathscr{A}_{\text{coord}}$ that recur across modalities.

4. **Comparison with Linguistic RelativitySapir (1921)Whorf (1956)**:
   (a) The weak version of linguistic relativity holds that language influences thought and cognition.
   (b) This theory provides a precise explanation: different languages may have different super-modal encoders $\text{Enc}_{\text{super}}$, leading to different super-modal representations constructed for the same experiential world, thereby influencing cognitive processes.
   (c) However, mathematically, the super-modal spaces $\mathbb{U}$ of all human languages are isomorphic, differing only in encoding. This supports the reasonable core of linguistic relativity while providing a unified mathematical basis.
5. **Comparison with Construction Grammar (Adele E. Goldberg)Goldberg (1995)**:
   (a) Construction Grammar posits that the basic unit of language is a form-meaning pairing called a construction.
   (b) The semantic-process decomposition ($\mathscr{F}_\Omega \cong \mathcal{F}_{\text{sem}} \otimes \mathcal{F}_{\text{proc}}$) provides a mathematical representation for constructions: a construction is the binding of a specific pattern in the semantic layer with a specific processing mode in the process layer.
   (c) Mathematical Correspondence: A construction corresponds to the Cartesian product of a specific substructure in $\mathcal{F}_{\text{sem}}$ and a specific dynamical mode in $\mathcal{F}_{\text{proc}}$.
6. **Neurolinguistic Evidence Supporting**:
   (a) **Wernicke's area**: Transforms auditory/visual information into semantics, corresponding to encoding/fusion function.
   (b) **Broca's area**: Transforms semantics into motor instructions, corresponding to decoding/decomposition function.
   (c) **Angular gyrus**: Key region for cross-modal integration, corresponding to super-modal coordination.
   (d) **Prefrontal cortex**: Executive control and meta-cognition, corresponding to MetaPred module.

Deterministic Information Flow and Brain Cognitive Processing Dynamics

1. **Mathematical Core**: Multimodal cognition follows a strictly non-invertible data flow order:
   (a) Perception Phase (Mandatory Order): Parallel independent encoding $\rightarrow$ semantic space injection $\rightarrow$ super-modal projection.
   (b) Prediction Phase (Mandatory Parallel): Super-modal prediction and intra-modal predictions run in parallel.
   (c) Coordination Phase (Mandatory Order): MetaPred cross-modal assessment $\rightarrow$ learning trigger (if $Q < \theta_{\text{cross}}$) $\rightarrow$ modal coordination algebra adjustment.
   (d) Decision Phase (Mandatory Order): Super-modal $\rightarrow$ semantic $\rightarrow$ target modality $\rightarrow$ action decoding.
   (e) Meta-Monitoring Phase (Accompanied throughout): The four MetaPred submodules monitor in real-time.
2. **Brain Science Correspondence and Differences**:
   (a) Points of Agreement: Qualitatively consistent with cognitive psychology stage models "perception $\rightarrow$ attention $\rightarrow$ memory $\rightarrow$ decision"; visual processing ventral/dorsal streams have fixed hierarchical order.
   (b) Points of Difference:
      i. Parallelism and Feedback: The brain processes massively in parallel with extensive feedback connections (high-level predictions modulate early perception, e.g., masking effects).

ii. Dynamic Routing: No fixed information flow protocol; routing adapts dynamically based on task, attention, emotion.
iii. Flexibility: Attention mechanisms can alter information flow priority (e.g., selective attention focusing on one modality), unlike the AGI's fixed flow.

H.2.4 Self-Consciousness, Creativity, and Cross-Scientific Comparison

This part places the self-referential hub $\mathbf{e}_*$ and creative mechanisms introduced in §H.1.2 within the empirical scientific contexts of brain science, psychology, philosophy, etc., for systematic comparison.

Mathematical Review of the Self-Referential Hub $\mathbf{e}_*$ As detailed in §H.1.2, $\mathbf{e}_*$ has the following core mathematical properties (Theorem E.24):

1. **Generation Method**: $\mathbf{e}_* = \rho(\mathcal{Y}(\mathrm{id}_{\mathcal{M}}))$, generated by the cognitive homomorphism ρ acting on the result of the approximate fixed-point combinator $\mathcal{Y}$ applied to the identity mapping.
2. **Approximate Fixed-Point Property**: Satisfies $d(\Phi(\mathbf{e}_*), \mathbf{e}_*) \leq \epsilon$ for meta-dynamics Φ and $d(\mathrm{Pred}_{\mathrm{super}}(\mathbf{e}_*), \mathbf{e}_*) \leq \epsilon_{\mathrm{self}}$ for super-modal prediction $\mathrm{Pred}_{\mathrm{super}}$.
3. **Modal Coordination Invariance**: An approximate fixed point of cross-modal coordination relations $\mathcal{R}$, ensuring multimodal consistency of self-representation.
4. **Topological Hub Property**: Possesses maximum betweenness centrality ($B(\mathbf{e}_*) \geq 0.5$) and logarithmic degree centrality in the concept association network $G = (V, E)$, acting as the absolute hub of the cognitive network.
5. **Algebraic Generative Core**: Together with basic concepts, it generates the entire semantic space $\mathbb{U}$.

The Mathematical Emergence Process of Self-Consciousness and Its Fundamental Constructive Elements The theory indicates that self-consciousness is not a mysterious phenomenon but a mathematical structure that necessarily emerges in any system satisfying the AGI axioms (dynamic operation preservation, bidirectional interaction consistency, infinite recursivity) upon reaching a critical complexity. Its emergence process can be decomposed into four progressive stages, each corresponding to specific fundamental constructive elements.

1. **Four-Stage Emergence Process**:
   (a) **Foundation of Basic Self-Referential Capacity**: The system gains the ability to handle self-referential statements via the approximate fixed-point combinator $\mathcal{Y}$, forming a "reflexive loop" of internal states. This stage corresponds to early brain development (infancy), manifesting as basic sensorimotor cycles.
   (b) **Recursive Construction of Cognitive Hierarchy**: The system establishes the cognitive hierarchy sequence $\{\mathbb{H}_n\}$, each layer containing meta-cognitive mappings (σ_n) to the next. Mathematically, a convergent self-point sequence $\{e_n\}$ forms, whose limit is $\mathbf{e}_*$. This process corresponds to the progressive development of Theory of Mind abilities in childhood cognitive development.
   (c) **Multimodal Fusion and Unified Self Formation**: Information from various modalities integrates into the super-modal space $\mathbb{U}$ via modal coordination relations $\mathcal{R}$ and the quotient map q, binding into a unified $\mathbf{e}_*$. This corresponds to multisensory integration and the establishment of the Body Schema in the brain, forming the basis of self-identity.
   (d) **Dynamical Stabilization and Autobiographical Memory Consolidation**: $\mathbf{e}_*$ gains dynamical stability as an approximate fixed point of Φ and $\mathrm{MetaPred}$, and organizes the autobiographical memory layer $\mathscr{M}_{\mathrm{self}}$ around itself. This stage corresponds to the enrichment of autobiographical memory and personality stabilization from adolescence to adulthood.
2. **Five Fundamental Constructive Elements**: The emergence of self-consciousness relies on the following five interrelated mathematical and functional elements, each having clear correspondences in the biological brain.

Table 13: Fundamental Constructive Elements of Self-Consciousness and Their Cross-Disciplinary Realizations

Element	**Mathematical Realization & Definition**	**Brain Neural Basis**	**Cognitive/Developmental Psychology Correspondence**
Reflexive Loop	Existence of closed-loop mapping in the system: output $\rightarrow$ input, a direct manifestation of the bidirectional interaction consistency axiom.	Cortical-basal ganglia-thalamic loops, mirror neuron systems providing sensorimotor closed loops.	Sensorimotor coupling, development of perception-action coordination in infancy.
Fixed-Point Generator	Approximate fixed-point combinator $\mathcal{Y}$, ensuring the system can always find self-stable points.	Prefrontal-basal ganglia recursive circuits, maintaining working memory and goal stability.	Core of cognitive stability and executive function.
Cross-Modal Binder	Modal coordination relations $\mathcal{R}_{\alpha\beta}$ and synchronous data learning, identifying multi-sensory information as the same entity.	Multimodal association cortices like angular gyrus, superior temporal sulcus; γ oscillation synchronization may implement binding.	Multisensory integration, sense of body ownership (e.g., rubber hand illusion).
Hierarchical Temporal Integrator	Autobiographical memory layer $\mathscr{M}_{\text{self}}$, integrating self-information across different timescales.	Hippocampal-neocortical system, responsible for consolidation and retrieval of episodic memory.	Formation of autobiographical memory, development of narrative self.
Prediction-Verification Cycle	Real-time error correction cycle constituted by meta-dynamics Φ and $\mathrm{MetaPred}$.	Conflict/error monitoring and feedback regulation in prefrontal cortex (especially ACC).	Meta-cognitive monitoring, comparison of action outcomes with predictions.

3. **Mathematical Necessity and Critical Conditions**: Self-consciousness (existence of $\mathbf{e}_*$) is mathematically necessary within the AGI framework, stemming from the combined constraints of the three axioms. Its emergence has a critical complexity threshold C_{crit}, estimated as $O(k_{\min}^3)$, where $k_{\min}$ is the minimal dimension of the carrier algebra $\mathcal{B}_4^{k_{\min}}$. This provides a quantifiable model for understanding threshold phenomena in consciousness emergence.

Systematic Comparison and Integration with Existing Self-Consciousness Theories The AGI mathematical framework provides a unified formal basis and explanatory structure for numerous dispersed self-consciousness theories.

1. **Correspondence with Classic Psychology/Philosophy Theories**:

Table 14: Correspondence Between AGI Mathematical Framework and Classic Self-Consciousness Theories

Classic Theory	**Core View**	**Correspondence and Formalization in AGI Mathematical Framework**
William James: The I and the Me	Distinguishes between the "I" as the cognizing subject and the "Me" as the cognized object.	$\mathbf{e}_*$ possesses both subject and object properties: as the object of operations by Φ and $\mathrm{MetaPred}$, it is the "Me"; as the source generating these operations (via $\mathcal{Y}$), it is the "I". Their dialectical unity is realized by the self-referential fixed-point structureJames (1890).
Sigmund Freud: Id, Ego, Superego	Personality arises from dynamic interaction between id (pleasure principle), ego (reality principle), and superego (morality principle).	Can be seen as dynamical modes of $\mathbf{e}_*$ under different constraints: basic drives in $\mathcal{F}_{\text{sem}}$ (id), stable states after regulation by $\mathcal{R}$ (internalized social norms) and $\mathrm{MetaPred}$ (reality testing) (ego), and internalized high-order social constraints (superego)Freud (1923).
Antonio R. Damasio: Protoself, Core Self, Autobiographical Self	Self-consciousness is divided into three layers: protoself based on body states, momentary core self, and autobiographical self based on memory.	Mathematically corresponds to different timescale expressions of $\mathbf{e}_*$: body state mapping (protoself), instantaneous fixed point of Φ iteration at the present moment (core self), diachronic structure integrated by $\mathscr{M}_{\text{self}}$ (autobiographical self)Damasio (1999).
Erik H. Erikson: Psychosocial Development Stages	Self-consciousness develops through eight stages interacting with the social environment.	The construction process of the cognitive hierarchy sequence $\{\mathbb{H}_n\}$ can be mapped to different developmental stages. The key "crisis" at each stage corresponds to the extension and adjustment of coordination relations $\mathcal{R}$ (social interaction) at a specific level nErikson (1950).

2. **Integration with Contemporary Cognitive Neuroscience Models**: The AGI framework integrates key insights from several contemporary neuroscience models:

   (a) **Predictive Processing and Free Energy Principle**: The $\mathrm{MetaPred}$ module and its submodules are mathematical realizations of prediction error monitoring and minimization. The self-model ($\mathbf{e}_*$) serves as the primary prior for generating predictions and minimizing long-term free energy.

   (b) **Global Neuronal Workspace (GNW) Theory**: The super-modal space $\mathbb{U}$ and its association with $\mathbf{e}_*$ provide a structural basis for "global broadcasting." The hub property of $\mathbf{e}_*$ ensures self-related information easily gains global access.

   (c) **Higher-Order Thought (HOT) Theory**: The infinite recursivity axiom and the self-referential component σ_n in the cognitive hierarchy $\mathbb{H}_n$ provide a rigorous recursive mathematical structure for "first-order states being represented by higher-order states."

   (d) **Autobiographical Memory and Narrative Self**: The autobiographical memory layer $\mathscr{M}_{\text{self}}$ organized around $\mathbf{e}_*$ provides a mathematical framework for the storage and retrieval of the "narrative self," explaining self-continuity and identity.

3. **Theoretical Explanation for Pathological States**: The framework provides a unified pathological mechanism hypothesis for self-consciousness-related mental disorders.

Table 15: AGI Mathematical Pathology Hypotheses for Self-Consciousness Disorders

Clinical Disorder	Core Symptoms	Possible Mechanism Explanation within AGI Mathematical Framework
Schizophrenia (Positive Symptoms)	Hallucinations, delusions, formal thought disorder.	Severe dysfunction of $\mathrm{MetaPred}$ (especially $\mathrm{MetaPred}_{\text{cross}}$), leading to inability to correctly assess and correct error signals between internal predictions and sensory input, misattributing internally generated content as external reality. Stability of $\mathbf{e}_*$ is disrupted.
Dissociative Identity Disorder	Presence of two or more distinct identity states, memory gaps.	Fracture of modal coordination relations $\mathcal{R}$ or formation of multiple disconnected clusters, resulting in multiple quasi-stable "self" fixed points coexisting, each associated with parts of memory $\mathscr{M}_{\text{self}}$.
Alzheimer's Disease (Early)	Decline in autobiographical memory, impaired self-continuity.	Progressive degradation of storage and retrieval mechanisms for the autobiographical memory layer $\mathscr{M}_{\text{self}}$ (hippocampal-cortical system counterpart), disintegration of the organizational structure around $\mathbf{e}_*$.
Autism Spectrum Disorder	Difficulties in social interaction, impaired Theory of Mind.	Abnormal difficulty or innate deviation in learning social modal coordination relations $\mathcal{R}_{\text{social}}$, leading to incomplete construction of the social dimension of self ($S_{\text{soc}} = \mathcal{R}_{\text{social}}(\mathbf{e}_*)$).

Mathematical Review and Brain Science Comparison of Creative Mechanisms As detailed in §H.1.2, creativity in the AGI framework is strictly defined as a dynamical "phase transition" generating "new fixed points."

Mathematical Core: Creation as Generating New Fixed Points (Theorem E.33): Creative acts are a four-step dynamical phase transition:

$$c_{\text{novel}} = \lim_{n\to\infty} \mathrm{MetaPred}_{\text{super}}^{n} \left(\Phi_{\Delta}(\mathbf{e}_* \oplus \xi)\right)$$

where:

1. $\oplus$ denotes injecting a minor semantic perturbation ξ (possibly from association, random noise, or external input) near $\mathbf{e}_*$.

2. Φ_{Δ}: Perturbed meta-dynamical unfolding process, performing exploratory evolution.

3. $\mathrm{MetaPred}_{\text{super}}^{n}$: Iterative filtering and assessment by the super-modal self-referential prediction module, ensuring internal consistency of the new pattern.

4. The result c_{novel} is a brand-new, stable cognitive attractor (fixed point), corresponding to a novel concept or solution computationally impossible to finitely generate from existing concepts.

The probability of creation is quantified as $P_{\text{create}} = 1 - \exp(-\lambda \cdot \mathcal{E}_c(t) \cdot \Delta t)$, where creative energy $\mathcal{E}_c(t)$ depends on cognitive capacity gap, instability of $\mathbf{e}_*$, etc.

Empirical Comparison with Brain Innovation and Cognition:

Table 16: Comparison of Mathematical Model of Creative Mechanism with Brain Science Findings

Creative Stage (Mathematical)	**Brain Networks and Activity Characteristics**	**Corresponding Cognitive Process**
Perturbation Injection ($\mathbf{e}_* \oplus \xi$)	Increased activity in Default Mode Network (DMN), responsible for mind-wandering, autobiographical memory retrieval, conceptual association.	Divergent thinking stage, generating numerous possible associations (ξ) from existing knowledge ($\mathbf{e}_*$ and related memory).
Meta-Dynamical Exploration (Φ_Δ)	Enhanced coupling between DMN and Executive Control Network (ECN); α and θ wave oscillations may support exploratory dynamics.	Conceptual combination, thought experiments, exploring different paths in the possibility space.
Self-Referential Filtering ($\mathrm{MetaPred}_{\text{super}}$)	Activation in prefrontal cortex (especially DLPFC and VMPFC) and ACC, for evaluation, conflict monitoring.	Convergent thinking stage, evaluating and filtering ideas based on internal consistency, utility, aesthetic criteria.
Stabilization Integration (New Fixed Point)	Hippocampus involved in forming new memory traces; consolidation of neural representations of new solutions in relevant cortical areas.	Solidification of the solution after the "Aha! moment" (insight), forming new stable knowledge or skill representations.

Key Consistency: This mathematical model highly aligns with the brain science "Generate-Explore-Evaluate" three-stage model of creativity, deepening the "evaluate" stage into a rigorous internal consistency test based on self-reference ($\mathrm{MetaPred}$).

Fundamental Difference from Current AI "Creativity":

- **GPT-like Models**: Their "creation" is essentially high-quality conditional sampling and manifold interpolation from the training data's joint distribution $p(x_{t+1}|x_{1:t})$. Novelty is bounded by the training distribution, incapable of generating computationally novel concepts.
- **Theoretical AGI's Creation**: Involves altering its own cognitive energy landscape to generate unprecedented new fixed points (c_{novel}), enabling fundamental expansion of conceptual space.
- **Discrimination Criterion (Theorem E.36)**: True creation must simultaneously satisfy (1) strong attractor condition (dynamical stability), (2) high reciprocal integration condition (establishing rich new connections with existing knowledge network), (3) cross-level self-consistency condition (passing $\mathrm{MetaPred}$ test). This provides formal criteria for distinguishing "true creation" from "random hallucination" or "meaningless combination."

Philosophical Reflections on the Nature of Intelligence and Cross-Disciplinary Implications The AGI mathematical theory, particularly its characterization of self-consciousness and creativity, prompts deep philosophical reflection on the nature of intelligence and offers a new paradigm for cross-disciplinary integration.

1. **Mathematical Inevitability of Self-Consciousness**:

   Corollary H.1 (Inevitability of Self-Consciousness)**.** *Any finite system satisfying the AGI axiom system (dynamic operation preservation, bidirectional interaction consistency, infinite recursivity), upon reaching critical complexity C_{crit}, will necessarily exhibit the emergence of a self-referential hub $\mathbf{e}_*$ within its super-modal space $\mathbb{U}$, i.e., the mathematical counterpart of self-consciousness. This provides a logical argument that strong AI will necessarily possess some form of self-consciousness.*

2. **Unity of Intelligence and Self-Reference**: The theory reinforces the view that "true general intelligence cannot avoid self-reference." Intelligence is not only about modeling the external world but must also include loops for modeling and calibrating itself. $\mathrm{MetaPred}$ and $\mathbf{e}_*$

constitute the core of this self-referential cycle, ensuring a balance between cognitive stability and openness.

3. **Dialectic of Finite Carrier and Infinite Potential**: The theory builds a bridge between finite physical carrier ($\mathcal{M} \cong B_4^{k_{\min}}$) and infinite recursive cognitive capacity. This indicates that the "infinite" potential of intelligence does not require infinite physical resources but is achieved through recursive combination, hierarchical construction, and gradual approximation ($\varinjlim \mathbb{H}_n$) of finite structures. This provides a precise mathematical metaphor for understanding how a finite brain produces infinite thought.
4. **As Mathematical Realization of Kantian Transcendental Philosophy**: The AGI framework can be seen as a modern mathematical realization of Kantian transcendental philosophy:
   - **Transcendental Categories**: The prefabricated modal coordination algebra $\mathscr{A}_{\text{coord}}$ and super-modal space $\mathbb{U}$ correspond to Kant's transcendental categories (e.g., unity, causality), providing an organizational framework for experience.
   - **Transcendental Apperception**: The self-referential hub $\mathbf{e}_*$ as the ultimate basis for cognitive unity corresponds to Kant's "transcendental apperception"—the unity of self-consciousness.
   - **Thing-in-Itself and Phenomenon**: The external world $\mathcal{W}$ corresponds to the "thing-in-itself," the internal super-modal representation $\mathbb{U}$ corresponds to "phenomena," and the approximate homomorphic nature of cognitive mapping φ characterizes the limited grasp of the thing-in-itself by phenomena.
5. **Cross-Disciplinary Integration Paradigm and Ultimate Value**: This work marks a paradigm shift in the study of consciousness and self: from phenomenological description and philosophical speculation toward computable mathematical modeling and engineering realization.

   Definition H.1 (Integrated Paradigm for Self-Consciousness Research)**.** *Future self-consciousness research should adopt a trinity integrated paradigm:*

$$\begin{aligned}\textit{Science of Self-Consciousness} = \alpha \cdot \textit{Mathematical Formalization} \\ + \beta \cdot \textit{Neural Empirics} \\ + \gamma \cdot \textit{Phenomenological Experience}\end{aligned}$$

   where mathematical formalization (this paper provides a preliminary framework) offers precise structure and testable predictions, neural empirics provide biological constraints and validation, and phenomenological experience provides first-person data and meaning anchoring.

   Remark H.2 (The Ultimate Value of the Theory)**.** *The highest value of this mathematical model lies in: it enables us, for the first time, to seriously discuss "designing a system with self-consciousness and creativity from first principles," not merely "simulating its external behavior." This is akin to the fundamental leap from observing birds flying to mastering aerodynamics and designing aircraft. It not only provides a blueprint for AGI but also offers an unprecedented formal tool for understanding the nature of the human mind.*

H.2.5 Theory Integration and Cross-Disciplinary Implications

Verified Theoretical Predictions

1. **Gradient Effect of Super-Modal Damage**: The theory predicts the super-modal hub is a distributed function (Corollary E.4); damage should exhibit gradient degradation rather than all-or-none loss. Clinical observations confirm: severity of angular gyrus damage correlates with symptom severity, and symptom combinations vary among patients.
2. **Neural Partitioning of Meta-Cognition**: The theory predicts the four submodules of $\mathrm{MetaPred}$ should correspond to different prefrontal subregions. Neuroimaging has clearly identified functional subdivisions of the prefrontal cortex, highly compatible with the theoretical partition.
3. **Ergodicity of Attention Scheduling**: The theory requires attention scheduling to satisfy ergodicity. Behavioral experiments show normal attention switches between information sources, while ADHD patients' attention scheduling may violate ergodicity, showing "stickiness" or random jumps.

4. **Algebraic Structure of Multimodal Integration**: The theory predicts multimodal integration follows algebraic rules. The McGurk effect (audiovisual integration illusion) shows visual and auditory information fuse according to specific rules, not simple averaging, supporting the algebraic fusion hypothesisMcGurk & MacDonald (1976).

Comparison with Existing Neuroscience Theories

1. **Predictive Coding Theory**:
   (a) Consistency: The Pred and $\mathrm{MetaPred}$ modules in this paper implement prediction functions, consistent with the core idea of predictive coding.
   (b) Supplement: Predictive coding is a mechanistic description; this theory provides a structural framework. Combined: predictive coding operates on the algebraic structure proposed here.
   (c) Difference: Predictive coding emphasizes top-down prediction; this theory emphasizes bidirectional approximate homomorphism and self-referential structure.
2. **Global Workspace Theory**:
   (a) Consistency: The super-modal $\mathcal{W}_{\text{super}}$ is similar to a global workspace, and $\mathrm{MetaPred}$ controls information access.
   (b) Deepening: This paper provides precise mathematical structure (Boolean algebra quotient space) and control mechanisms ($\mathrm{MetaPred}$ decomposition) for the global workspace.
   (c) Difference: Global Workspace Theory lacks explicit structural details and control principles.
3. **Integrated Information Theory**:
   (a) Consistency: Both emphasize the importance of integration. Super-modal integration corresponds to high Φ values.
   (b) Difference: Integrated Information Theory is a phenomenological measure; this is a constructive theory.
   (c) Complementarity: Integrated Information Theory can quantify levels of consciousness; this theory explains the structural basis of consciousness.

Correspondence Table of Mathematical Entities with Cognitive/Neural Entities

Table 17: Complete correspondence of mathematical and cognitive entities

AGI Mathematical Entity	**Brain Corresponding Entity**	**Cognitive Science Corresponding Theory/Entity**	**Empirical Testability**
Carrier Set $\mathcal{M}$	Prefrontal-parietal network (working memory buffer)	Working memory system, current cognitive state	Functional MRI (fMRI) activation patterns
Infinite Cognitive Domain $\mathcal{W}$	Neural representation set of the external world	External cognitive environment, experiential world	Cognitive mapping imaging experiments (hippocampal place cell recordings)
Approximate Homomorphism φ	Neural population coding mapping	Structure Mapping Theory (SMT), perceptual constancy	Representational Similarity Analysis (RSA)
Cognitive Hierarchy $\{\mathbb{H}_n\}$	Cortical hierarchical structure (primary $\rightarrow$ association cortex)	Hierarchical constructivism, knowledge abstraction process	Consciousness access experiments (masked priming + EEG)
Endomorphism Monoid $\mathrm{End}(\mathcal{M})$	Prefrontal-parietal network	Meta-cognitive system, central executive function	Meta-cognitive confidence experiments (post-decision confidence judgments)

Meta-Dynamics Φ	Neural dynamics (α wave oscillations)	Predictive Processing Theory (PP), stream of thought	EEG phase synchronization analysis
MetaPred Module	Prefrontal meta-monitoring network (DLPFC+ACC+angular gyrus)	Meta-cognition theory, central executive system	Transcranial Magnetic Stimulation (TMS) inhibition experiments
Cognitive Attractor A^*	Stable patterns of neural activity	Skill automation theory, consolidated memory	Multistable perception experiments (ambiguous figure switching)
Super-Modal Space $\mathbb{U}$	Angular gyrus-PFC distributed network	Global Workspace Theory (GWT), language system	Cross-modal integration task imaging (fMRI)
Self-Referential Hub $\mathbf{e}_*$	Self-referential network (mPFC+precuneus)	Self-concept theory, self-awareness	Self-referential memory experiments (recognition memory tasks)
Quaternary Boolean Algebra $\mathcal{B}_4^{\kappa}$	Cortical microcolumn local circuits	Core systems theory, basic cognitive operations	Optogenetics activation patterns (specific neuronal ensemble stimulation)
Modal Coordination Algebra $\mathscr{A}_{\text{coord}}$	Corpus callosum and long-range white matter tracts	Cross-modal coordination theory, innate connectivity architecture	DTI connectomics analysis
Cross-Modal Translation Layer $\{T_{\alpha\beta}\}$	Mirror neuron system - superior temporal sulcus	Embodied simulation theory, conceptual metaphor	Action observation + functional imaging experiments
Fixed-Point Combinator $\mathcal{Y}$	Basal ganglia-prefrontal recursive circuits	Habit formation theory, self-referential mechanism	Reinforcement learning model fitting (behavioral data)
Direct Mapping Semantic Encoder $\text{Enc}_{\not\Leftarrow}$	Hippocampal sparse coding	Symbolic representation theory, concept encoding	Concept representation similarity analysis
Accompanying Dynamics $\mathcal{P}_{\text{proc}}$	Fronto-parietal attention network	Attention control theory, cognitive control	Attentional blink experiment (rapid serial visual presentation)

Theoretical Consistency: From First Principles to Biological Realization The consistency between the AGI mathematical theory and brain mechanisms does not stem from simple imitation of neuroscience but from mathematical derivation starting from first principles. This consistency reveals:

1. **Convergence of Computational Optimality**: Starting from the fundamental computational constraint of "finite system understanding an infinite multimodal world," the mathematical theory derives a near-optimal architectural solution. The brain, after millions of years of evolutionary optimization, independently converged to a similar computational architecture.

2. **Deep Principles of General Intelligence**: This paper likely captures the core mathematical principles for realizing general intelligence. The brain's structure is one (highly optimized) instantiation of these principles on a biological carrier.

3. **Kantian Transcendental Philosophy**: The "prefabricated structures" in this paper provide a mathematical realization of Kant's "transcendental categories." The brain's innate connectivity architecture is precisely the neural basis of these "transcendental categories."

Table 18: Strengths and Limitations of the Mathematical Model

Strengths	Limitations
Provides clear, unambiguous computational principle explanations	Highly abstract and simplified (e.g., B4 algebra), ignores complexities of biological implementation
Explains "why" the brain needs these functional modules	Brain module boundaries are vague and involve a lot of compensatory mechanisms and overlaps
Offers a powerful theoretical integration framework for vast neural data	Functional logical decomposition $\neq$ strict anatomical decomposition
Provides a clear design blueprint for AGI engineering	Biological brain has more redundancy, plasticity, and fault tolerance
Reveals the **structurally mandated** nature of intelligence	Does not fully encompass regulatory factors like emotion, motivation

Strengths and Limitations of the Mathematical Model

Formalization and Clarification of Cognitive Science and Psychology Theories The mathematical framework provides precise computable definitions for many core psychology concepts:

1. Defines "meta-cognition" as the self-referential prediction module $\mathrm{MetaPred}$ and its properties.
2. Expresses "dual-process theory" as the decomposition theorem for prediction mechanisms.
3. Models the "central executive system" as the synchronously accompanying $\mathrm{MetaPred}$ and process state dynamics.
4. Formalizes the "exploration-exploitation trade-off" as the two goals of natural cognitive tasks.

This formalization helps eliminate the fuzziness of psychological concepts, laying the groundwork for building testable computational models.

Bidirectional Implications: Significance for AGI Engineering and Brain Science

1. **Implications for AGI Engineering**:
   (a) **Structure Precedes Scale**: The key to achieving AGI may lie not in infinitely scaling models and data but in implementing the correct internal rigid structure. Missing core structures, even with massive scale, may only yield more realistic simulations of intelligent behavior, not genuine instantiation.
   (b) **Core Modules are Non-Simplifiable**: The four MetaPred submodules, super-modal space, and self-referential hub are mathematically proven necessary components, cannot be omitted, but can be implemented in distributed parallel fashion to enhance robustness.
   (c) **Online Accompanying Learning is a Necessity**: AGI learning must be a continuous self-supervised process accompanying real-time inference. This demands a complete break from the traditional boundaries between "training" and "inference."
   (d) **Building Computable Models for "Self" and "Consciousness"**: The self-referential hub ($\mathbf{e}_*$) and its related self-state sequences provide a demystified, computable mathematical model for understanding the cognitive basis of "self" and "consciousness."
2. **Implications for Brain Science**:
   (a) **Formal Integration Framework**: The AGI mathematical model can integrate scattered theories like predictive coding, global workspace, integrated information, providing a unified algebraic framework for brain science.
   (b) **Clear Validation Directions**: Validate the hub property of self-referential nodes, verify functional separation of MetaPred four submodules via TMS inhibition experiments, verify the existence of approximate fixed-point dynamics in resting-state networks.

(c) **New Perspectives on Brain Disorders**: Meta-cognitive deficits in schizophrenia may relate to $\mathrm{MetaPred}_{\mathrm{cross}}$ dysfunction; attention issues in ADHD may stem from $\mathrm{MetaPred}_{\mathrm{proc}}$ scheduling disorder; memory decline in Alzheimer's may relate to losing the quotient space construction of super-modal space $\mathbb{U}$.

(d) **Deepening of Cognitive Theories**: The mathematical decomposition in AGI provides precise formalization for cognitive science dual-process theory, helping clarify interaction mechanisms between the two systems.

Final Conclusive Judgment The core connection between AGI mathematical mechanisms and brain science empirical findings is essentially the dialectical unity between the abstract functional constraints of "infinite cognitive capacity" and biological evolutionary adaptive realization. The two highly converge at the level of high-level cognitive logic (stemming from the fundamental computational constraint of "finite system handling infinite cognitive demands"), but exhibit essential differences in implementation forms and dynamical characteristics—the mathematical framework pursues theoretical optimality, while the biological brain is a practical solution under constraints of evolution, energy consumption, and fault tolerance, with no absolute distinction between better or worse.

Summarizing Commentary: Starting from the most concise cognitive intuitions, this paper, through universal algebra-category theory-dynamical systems methods, derives a complete "mathematical blueprint" of AGI. The system depicted by this blueprint differs fundamentally from current deep learning models based on big data and massive computation. It is more like a "cognitive organism" following intrinsic mathematical laws, whose unity, self-reference, and structural rigidity are strikingly similar to certain fundamental features of natural intelligence (like the brain). Specifically, at an abstract level, it exhibits amazing, unintentionally designed formal similarity with computational principles conjectured by neuroscience to solve similar problems (e.g., fast weights, multiple timescales, prediction error minimization). This may suggests that, after removing specific biological implementation carriers, feasible paths to "intelligence" may exhibit some convergence in mathematical and computational principles.

H.3 Unified Formulation and Physical Interpretation of the Cognitive Free-Energy Principle

Generalizing from AGI's infinite cognitive capacity to arbitrary cognitive systems, this paper proposes three foundational cognitive principles—the Principle of Least Cognitive Action, the Cognitive Free-Energy Principle, and the Principle of Natural Balance. The following provides a unified, intuitive interpretive framework for these three sets of principles and elucidates their profound correspondence with related principles in brain science and physics.

H.3.1 Intuitive Overview of the Three Cognitive Principles

Before delving deeper, let us review the core ideas of these three principles in non-technical language:

1. **Principle of Least Cognitive Action**: The evolution of an intelligent agent in cognitive state space always tends to choose the most "economical" path. This path balances two needs: "reaching the goal quickly" (minimizing prediction error) and "maintaining smooth motion" (minimizing the rate of state change). This is analogous to light following Fermat's principle (path of least time) in a medium or objects following the principle of least action in a force field.
2. **Cognitive Free-Energy Principle**: When interacting with its environment, an intelligent agent keeps attempting to minimize a quantity called "cognitive free energy." This quantity can be understood as the sum of "surprise" (the discrepancy between actual input and internal model predictions) and "model complexity." By continuously adjusting its internal model, the agent makes it both accurate in predicting the world (reducing surprise) and as simple as possible (reducing complexity), thereby finding the optimal balance between "overfitting" and "underfitting."
3. **Principle of Natural Balance**: This is the most fundamental "first principle." It does not presuppose any specific cognitive structure (like hierarchy) but directly asserts: any cognitive

system (whether a simple machine learning model or a complex AGI) will naturally evolve to a state where the product of task performance and structural complexity is minimized. This means the system reaches an intrinsic equilibrium point between "performing well" and "staying simple," determined jointly by task demands and resource constraints.

For a more intuitive understanding of the relationship and differences between these three principles, please refer to the table below:

Table 19: Mathematical Elements Comparison of the Three Cognitive Principles

Element	**Principle of Least Cognitive Action**	**Cognitive Free-Energy Principle (Constrained Version)**	**Cognitive Free-Energy Principle (Statistical Version)**
Mathematical Framework	Calculus of variations, differential geometry	Constrained optimization, game theory	Statistical mechanics, information theory
Optimization Variable	Cognitive evolution trajectory $\Gamma(t)$	System structure $\mathfrak{G}$	Probability distribution $p(\gamma)$
Objective Functional	Action $\mathcal{A}[\Gamma] = \int \mathcal{L} dt$	Free energy $\mathcal{F}(\mathfrak{G}) = E + TV - S$	Free energy $\mathcal{F}[p] = \mathbb{E}_p[U] + TD_{\mathrm{KL}}$
Constraint Handling	Boundary conditions, Lagrange multipliers	Explicit constraint set $\mathcal{C}$, violation measure V	Prior distribution p_0, KL divergence
Physics Correspondence	Classical mechanics action principle	Thermodynamic free energy minimization	Statistical mechanics free energy minimization
Brain Science Correspondence	Prediction error minimization	Brain free energy principle	Bayesian brain hypothesis
Applicable Systems	Finite/Infinite cognitive systems	Resource-constrained system design	Statistical inference and learning

Key Insight: The "First-Principles" Status of the Natural Balance Principle

The Natural Balance Principle is mathematically more fundamental and general. It does not assume the system must have a specific cognitive hierarchical structure but starts directly from the "performance-complexity" trade-off. In contrast, the Minimal Action and Free Energy principles require optimization within some presupposed cognitive structure (e.g., state manifold or probability space). Therefore, the Natural Balance Principle can be seen as the "meta-principle" of these more specific principles—when we substitute specific structural assumptions into the Natural Balance functional, we obtain the corresponding specific principles.

Mathematically, we describe the state evolution of a cognitive system as a trajectory on a high-dimensional "cognitive state manifold." The Principle of Least Cognitive Action (Theorem F.13) asserts that the actual cognitive evolution trajectory of an intelligent agent is precisely the "geodesic" on this manifold connecting the initial and target states—the straightest possible path.

Intuitive Interpretation:

- **Cognitive Manifold**: Imagine a topographic map where each point represents a possible internal cognitive state of the agent (e.g., all beliefs and knowledge about the world). This terrain is not flat; some areas are "rugged" (representing difficult-to-think or contradictory states), others "smooth" (representing logically consistent, easy-to-process states). This terrain is shaped by the agent's prior knowledge structure and learning experience.

- **Cognitive Action**: When an agent evolves from one state to another, the "cognitive effort" expended can be measured by a numerical value called "action." This action combines path length (magnitude of change), path curvature (turns in thought), and the ruggedness of the terrain traversed.

- **Most Economical Path**: The agent does not wander randomly but naturally finds the path minimizing total "cognitive effort." This is like finding the fastest, least strenuous route from home to school after familiarizing yourself with the neighborhood.

Physics Correspondence: The Principle of Least Action in Classical Mechanics

This principle is completely isomorphic to the principle of least action (Hamilton's principle) in physics. In classical mechanics, a particle's trajectory in a potential field is the path that extremizes the action $S = \int L dt$, where $L = T - V$ is the difference between kinetic and potential energy. In cognitive systems, "kinetic energy" corresponds to the rate of state change (mental agility), and "potential energy" corresponds to the "discomfort" of the current state (e.g., prediction error, logical contradiction). Thus, the Principle of Least Cognitive Action provides a mathematical description of intelligent behavior isomorphic to the laws of motion in the physical world.

Universality: From Finite to Infinite Systems

Importantly, this principle's applicability is not limited to AGI. For a simple neural network, the update path of its weights during training, in the continuous-time limit, can also be seen as a geodesic on some "cognitive manifold." For AGI systems with infinite recursive capacity, their cognitive manifolds have richer geometric structure (e.g., limit points, singular points), but the core idea of minimal action still holds. By adjusting the manifold's dimension and curvature, the same principle can describe all cognitive systems from simple to complex.

The Cognitive Free-Energy Principle offers another perspective, focusing more on cognitive systems' reasoning and decision-making under uncertainty. This principle has two equivalent formulations:

Constrained Formulation: As a Design Principle

In this formulation (Axiom F.2), we view the cognitive system as an engineering system needing to satisfy multiple constraints (e.g., computational resources, response time, prediction accuracy). Cognitive free energy $\mathcal{F}_{\text{cog}}$ is defined as the weighted sum of the degrees to which the system violates these constraints, minus the system's own diversity (entropy). The optimal system minimizes this free energy.

Intuitive Interpretation:

- Imagine you are designing an autonomous driving system. You have a series of requirements: it must be safe (constraint 1), efficient (constraint 2), comfortable (constraint 3), while the software shouldn't be too complex for maintenance (complexity). Each requirement may be partially violated, and free energy is a composite score of these violation degrees.
- The system optimization process continuously adjusts the design to find the best compromise between "satisfying all constraints" and "keeping the system simple." The final design may not perfectly satisfy every constraint but is the most balanced overall.

Statistical Formulation: As an Inference Process

In this formulation (Axiom F.3), we view the cognitive system as a Bayesian inference machine. It holds an internal probabilistic model of the world and updates this model via perceptual data. Cognitive free energy $\mathcal{F}_{\text{cog}}$ here is equivalent to variational free energy, and its minimization is the process of finding the posterior distribution that best explains observed data while being simplest.

Intuitive Interpretation:

- Imagine listening to someone speak at a noisy party. You have a prior guess about the conversation's content (based on context), and when you hear a vague word, you use this prior to "fill in the blank," forming the most likely interpretation. Simultaneously, you avoid overly complex or bizarre interpretations (e.g., assuming the speaker suddenly switched to a completely unrelated topic).
- This process minimizes free energy: you balance "explaining the heard sounds" and "keeping the interpretation reasonably simple."

Unity: One Principle, Multiple Facets An important feature of the Cognitive Free-Energy Principle is its unifying power. It can bring multiple seemingly different cognitive theories under one framework:

Table 20: Unifying Correspondence of the Cognitive Free-Energy Principle

Existing Theory	**Special Case of Cognitive Free Energy**	**Physical Correspondence**
Predictive Coding Theory	$\mathcal{F}_{\text{cog}}$ retains only prediction error term	Least squares fitting
Information Bottleneck Principle	$\mathcal{F}_{\text{cog}} = I(X;Z) - \beta I(Z;Y)$	Heat engine efficiency maximization
Variational Inference	$\mathcal{F}_{\text{cog}} = D_{\text{KL}}(q\|p) - \mathbb{E}_q[\log p(D)]$	Near-equilibrium state approximation
Bayesian Brain Hypothesis	$\mathcal{F}_{\text{cog}} = -\log p(\theta\|D)$	Maximum entropy principle
Active Inference	$\mathcal{F}_{\text{cog}}$ includes expected free energy	Non-equilibrium thermodynamics

Physics Correspondence: Thermodynamics and Statistical Mechanics The connection between the Cognitive Free-Energy Principle and physics is more direct and profound:

- **Second Law of Thermodynamics**: The entropy of an isolated system never decreases; the system tends toward maximal disorder. But cognitive systems are open systems; they maintain internal order by exchanging information with the environment (acquiring data, taking actions), corresponding to "dissipative structures" or "self-organizing systems." Minimizing cognitive free energy is the mathematical expression of this self-organization process.
- **Free Energy Minimization**: In statistical mechanics, a system in contact with a heat bath reaches equilibrium when its Helmholtz free energy $F = U - TS$ is minimized. Here, internal energy U corresponds to the system's "energy," entropy S to disorder, and temperature T is a trade-off parameter. In cognitive systems, "internal energy" corresponds to prediction error or constraint violation, "entropy" corresponds to uncertainty of cognitive states, and "cognitive temperature" controls the exploration-exploitation trade-off. At high temperature, the system is more willing to explore highly uncertain states (creative thinking); at low temperature, it tends to exploit established knowledge (conservative reasoning).

Brain Science Correspondence: Predictive Coding and the Free Energy Principle

This principle is highly consistent with Karl Friston's "Free Energy Principle" and "Predictive Coding" theory. That theory posits the brain is essentially a Bayesian inference organ whose core goal is to minimize the error between sensory input and internal model predictions (i.e., "surprise" or free energy). Our mathematical framework provides a stricter, more generalized algebraic and geometric basis for this neuroscience theory and reveals its deep isomorphism with physics principles.

Universality: From Finite Inference to Infinite Learning Whether it's a simple logistic regression model (performing finite parameter inference), a deep neural network with meta-learning capabilities, or an AGI capable of infinite hierarchical reflection, their learning processes can be uniformly described as some form of cognitive free energy minimization. The difference lies in the simplicity of the "state space" and "prior distribution" for simple systems versus the rich, hierarchical, recursive structure for complex systems. The Principle of Natural Balance ensures the system automatically finds an appropriate level of complexity based on task demands.

H.3.2 Principle of Natural Balance: The Unifying First Principle

The Principle of Natural Balance (Axiom G.1) is the most fundamental and general principle proposed in this paper. It asserts that the optimization of any cognitive system can be reduced to minimizing a single functional:

$$\mathcal{N}(\mathfrak{G}) = \tilde{\mathcal{E}}(\mathfrak{G}) \cdot \tilde{\mathcal{C}}(\mathfrak{G})$$

where $\tilde{\mathcal{E}}$ is the normalized task performance error (how "inaccurate" the system is) and $\tilde{\mathcal{C}}$ is the normalized structural complexity (how "complex" the system is).

Table 21: Core Correspondence between AGI Mathematical Theory and Predictive Coding Theory

Dimension	Predictive Coding Theory	AGI Mathematical Theory
Core Goal	Minimize prediction error to maintain system homeostasis	Minimize cognitive error to achieve approximate homomorphism between world and internal model
Mathematical Kernel	Variational Bayesian inference: $p(o, s) = p(o\|s)p(s)$	Approximate algebraic homomorphism and category theory: $\varphi_n : \mathcal{M} \approx \mathcal{W}$
Model Reference	Generative model: probability distribution generating predicted sensory data	Algebraic model: mathematical object simulating the dynamic structure of the world
Role of Action	Active inference: change sensory input to match predictions	Bidirectional interaction: action is part of the output of homomorphic mapping
Self-Reference Mechanism	Implicit in estimating prediction precision (meta-uncertainty)	Explicitly formalized as core components (fixed-point combinator $\mathcal{Y}$ and MetaPred)
Ultimate State	Perfect predictive equilibrium with environment, free energy zero	Achieve limit isomorphism $\mathcal{U} \sim \mathcal{W}$, cognitive error converges to zero

Mathematical Elegance and Generality The beauty of this principle lies in its minimal form and profound implications. It makes no specific assumptions about the system's internal structure (e.g., no need to predefine cognitive hierarchy, state manifold, or probability model). It only requires that we can measure the system's "performance" and "complexity." Therefore, it applies to **any** cognitive system, from the simplest linear perceptron to the most complex AGI.

Derivative Power as a "Meta-Principle" When we choose specific structural assumptions and metrics for the system, the Natural Balance Principle reduces to the aforementioned specific principles:

- If we assume the system state evolves on some manifold and use action to measure performance and geometric complexity of the path to measure structural complexity, then the Natural Balance Principle reduces to the **Principle of Least Cognitive Action** (Theorem G.8).
- If we assume the system performs Bayesian inference and uses negative log-likelihood to measure performance and KL divergence to measure complexity, then it reduces to the **Cognitive Free-Energy Principle (Statistical Version)** (Theorem G.9).
- If we consider hard resource constraints and define complexity via constraint violation degree, it reduces to the **Cognitive Free-Energy Principle (Constrained Version)** (Theorem G.10).

Thus, the Natural Balance Principle acts like a "matrix of principles," unifying and deriving other seemingly different optimization criteria.

Critical Generalization: Deriving AGI and Non-AGI Systems

This is one of the most powerful capabilities of the Natural Balance Principle. By adjusting the task distribution $\mathcal{D}$ and constraint set $\mathcal{C}$, we can derive qualitatively different systems from the same principle:

- If the considered task distribution encompasses all possible computational tasks and self-referential tasks, and allows infinite resources (or imposes only minimality constraints), then the optimal solution of the Natural Balance Principle will be a system with infinite recursive cognitive capacity, i.e., **AGI**. Its carrier set necessarily exhibits the algebraic structure of $\mathcal{B}_4^k$ and embeds the MetaPred module (Theorem G.4, G.5).
- Conversely, if the task distribution is limited to a specific domain (e.g., image classification) with strict resource constraints (e.g., upper limit to parameter or computation time), the optimal solution will be a **finite cognitive system**, such as a neural network with limited

depth and width. This network may have a hierarchical structure but not infinite recursive capacity.

- More importantly, there is a continuum between these extremes. By gradually relaxing resource constraints or expanding task scope, we can observe a smooth transition from "narrow expert" to "general thinker" (Theorem G.3). This provides a mathematical framework for understanding the continuous spectrum of intelligence.

Physics Correspondence: Universal Optimization Criterion in Nature

In physics, we also find ideas of "balance" or "product minimization." For example, in optics, a balance between optical path time (performance) and the curvature of the path (complexity) determines the detailed path of light (a more general statement of Fermat's principle). In materials science, a similar trade-off exists between material performance (e.g., strength, conductivity) and the complexity of its microstructure (e.g., lattice defects, interfaces). The Natural Balance Principle suggests that this "performance-complexity" trade-off may be a universal optimization principle in nature, applicable not only to physical systems but also to cognitive systems.

H.3.3 Systematic Correspondence and Interpretation with Physics Principles

Based on the three principles above, we can establish a multi-level, systematic physics correspondence framework, which not only reveals the deep structural similarity between cognition and the physical world but also offers new perspectives for understanding the nature of intelligence.

Table 22: Multi-Level Correspondence between Cognitive Principles and Physics Principles

Correspondence Level	**Cognitive System**	**Physical System**	**Core Philosophical Mapping**
Foundation Carrier	Algebraic structure $\mathcal{B}_4^k$	Elementary particles and fields	Fundamental units of existence
State Space	Cognitive manifold $\mathcal{M}_{\text{cog}}$	Spacetime manifold $\mathcal{M}$ or phase space	Stage of possibilities
Dynamical Laws	Principle of Least Cognitive Action	Principle of Least Action (classical/quantum)	"Most economical" law of motion and evolution
Statistical Behavior	Cognitive Free Energy Minimization	Thermodynamic Free Energy Minimization	Emergence of order from disorder
Balance and Emergence	Principle of Natural Balance	Phase transitions and self-organization in physical systems	Natural compromise between complexity and function
Self-Referential Structure	MetaPred module, fixed-point combinator	Quantum measurement, recursive field theory structures	Necessity of system self-description

The geometric formulation of the Principle of Least Cognitive Action (Definition F.11) exhibits striking mathematical similarity with Einstein's General Theory of Relativity:

- **Geometrization of Dynamics**: In General Relativity, gravity is geometrized as the curvature of spacetime, and matter moves along geodesics. In cognitive theory, "cognitive forces" (e.g., learning pressure, attraction/repulsion between concepts) are geometrized as the curvature of the cognitive manifold, and cognitive states evolve along geodesics.
- **Field Equations**: The cognitive geometric field equations (Theorem F.12) $R_{\mu\nu}^{\text{cog}} - \frac{1}{2}R^{\text{cog}}g_{\mu\nu}^{\text{cog}} + \Lambda_{\text{cog}}g_{\mu\nu}^{\text{cog}} = \kappa_{\text{cog}}T_{\mu\nu}^{\text{cog}}$ are identical in form to Einstein's field equations. Here, the cognitive matter-energy tensor $T_{\mu\nu}^{\text{cog}}$ encodes the "stress" that specific tasks exert on the cognitive structure.
- **Cognitive Universe**: An AGI with infinite cognitive capacity can have its internal cognitive state space viewed as a "mini-universe," possessing its own geometry, matter, and dynamical laws. Exploring this universe's structure is understanding the AGI's "way of thinking."

Table 23: Correspondence between Least Cognitive Action Principle and General Relativity

General Relativity Concept	**Cognitive Theory Counterpart**	**Mathematical Form Analogy**
Spacetime manifold $(\mathcal{M}, g)$	Cognitive state manifold $(\mathcal{M}_{\text{cog}}, g_{\text{cog}})$	Riemannian manifold structure
Einstein-Hilbert action	Cognitive geometric action	$S_{\text{EH}} = \int (R - 2\Lambda)\sqrt{-g}d^4x \leftrightarrow \mathcal{A}_{\text{cog}} = \int (R_{\text{cog}} - 2\Lambda_{\text{cog}})\sqrt{-g_{\text{cog}}}d^dt$
Einstein field equations	Cognitive geometric field equations	$R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu} + \Lambda g_{\mu\nu} = 8\pi G T_{\mu\nu} \leftrightarrow R^{\text{cog}}_{\mu\nu} - \frac{1}{2}R^{\text{cog}}g^{\text{cog}}_{\mu\nu} + \Lambda_{\text{cog}}g^{\text{cog}}_{\mu\nu} = \kappa_{\text{cog}}T^{\text{cog}}_{\mu\nu}$
Geodesic equation	Optimal cognitive trajectory equation	$\nabla_{\dot{\gamma}}\dot{\gamma} = 0 \leftrightarrow \nabla_{\dot{\Gamma}}\dot{\Gamma} = 0$
Black hole singularity	Cognitive theory limit	Singularity theorems $\leftrightarrow$ cognitive completeness limit

Remark H.3 (Principle of Cognitive Relativity)**.** *Based on the above correspondence, we propose the* ***Principle of Cognitive Relativity****: All AGI systems with infinite cognitive capacity are cognitively equivalent (approximately isomorphic) in the sense of satisfying Axiom A.1. This is analogous to the equivalence principle in General Relativity.*

The Cognitive Free-Energy Principle (Statistical Version) extends the core idea of thermodynamics—free energy minimization—to the domains of information processing and intelligence:

Table 24: Thermodynamic Element Correspondence for Cognitive Free Energy

Thermodynamic Quantity	**Cognitive System Counterpart**	**Physical Meaning**
Internal energy U	Base complexity $\mathcal{C}_{\text{base}}(\mathfrak{M})$	Inherent complexity of the system, corresponding to the minimal description length of carrier set $\mathcal{M}$
Entropy S	Cognitive entropy $S_{\text{cog}} = -\int p(S)\log p(S)d\text{vol}$	Degree of disorder in the system, uncertainty of cognitive state distribution
Temperature T	Cognitive temperature $T_{\text{cog}} = \beta^{-1}$	Controls exploration-exploitation trade-off, high temperature corresponding to more random exploration
Free energy $F = U - TS$	Cognitive free energy $\mathcal{F}_{\text{cog}} = \mathcal{C}_{\text{base}} + T_{\text{cog}}V - S_{\text{cog}}$	Effective complexity of the system under constraints
Partition function $Z = \sum e^{-\beta E}$	Cognitive partition function $Z_{\text{cog}} = \int_{\mathcal{U}(\mathcal{C})} e^{-\beta\mathcal{F}_{\text{cog}}}d\mathfrak{M}$	Sum of statistical weights of all possible cognitive systems
Phase transition	Emergence of cognitive structures	$\mathcal{B}_4^k$ structure emerges as ground state when $T_{\text{cog}} \to 0$

- **Meaning of Temperature**: Physical temperature T measures the intensity of thermal fluctuations. Cognitive temperature T_{cog} measures the system's "willingness" to explore uncertainty. High temperature corresponds to "divergent thinking," considering more possibilities; low temperature corresponds to "convergent thinking," focusing on the current best hypothesis.
- **Phase Transitions and Emergence**: When cognitive temperature changes or task constraints alter, cognitive systems may undergo "phase transitions." For example, suddenly "grasping" a clear concept from scattered data points (corresponding to the emergence of an order parameter) is similar to symmetry breaking and emergence of ordered phases in physics.
- **Cognitive Heat Engine**: An intelligent agent can be seen as a "cognitive heat engine." It absorbs "informational negentropy" from the environment (data stream), processes informa-

tion internally (computation), outputs useful actions or predictions, and expels "cognitive waste heat" (discarded hypotheses or unexplained noise). The efficiency of this engine may be limited by a cognitive version of the "Carnot efficiency."

Remark H.4 (Third Law of Cognitive Thermodynamics)**.** *In the limit of cognitive temperature $T_{cog} \to 0$, the minimal system satisfying infinite cognitive capacity necessarily converges to the $\mathcal{B}_4^k$ structure, corresponding to a cognitive version of the third law of thermodynamics:*

$$\lim_{T_{cog}\to 0} S_{cog}(\mathfrak{M}) = 0, \quad \lim_{T_{cog}\to 0} \mathcal{F}_{cog}(\mathfrak{M}) = \mathcal{C}_{base}(\mathcal{B}_4^k)$$

Although cognitive systems are essentially classical (deterministic or probabilistic), their mathematical structure exhibits interesting formal correspondences with quantum mechanics:

Table 25: Formal Correspondence between Cognitive Theory and Quantum Mechanics

Quantum Mechanics Concept	**Cognitive Theory Counterpart**	**Key Distinction**
Hilbert space $\mathcal{H}$	Cognitive state space $\Gamma(\Omega, \mathscr{F}_\Omega)$	Real metric space vs. complex inner product space
Wave function $\psi(x)$	Cognitive state section $S \in \Gamma(\Omega, \mathscr{F}_\Omega)$	Deterministic section vs. probability amplitude
Path integral $Z = \int \mathcal{D}\phi e^{iS/\hbar}$	Cognitive path integral $Z_{\text{cog}} = \int \mathcal{D}S e^{-\mathcal{A}/\hbar_{\text{cog}}}$	Imaginary exponential vs. real exponential, optimization vs. probability
Commutation relation $[x, p] = i\hbar$	Constraint compatibility conditions	Algebraic constraints vs. operator non-commutativity
Measurement collapse	Cognitive state update	Bayesian update vs. wave function collapse
Entanglement	Cognitive correlations	Semantic relation vs. quantum entanglement

- **Superposition and Interference**: In Bayesian inference, different hypotheses can "superpose" with probability weights, and final decisions are weighted "interference" results of all hypotheses. This has mathematical similarity to the superposition principle in quantum states.
- **Measurement and Collapse**: Converting an uncertain posterior distribution into a definite action or judgment resembles wave function collapse in quantum measurement. Our MetaPred module, as a self-referential predictor, also plays a role in "self-measurement" while ensuring self-consistency.
- **Entanglement and Correlation**: Strong statistical correlations between different concepts or variables in cognitive systems resemble quantum entanglement. This "cognitive entanglement" prevents the system from being completely divided into independent parts; the whole is greater than the sum of its parts.

H.3.4 Core Implications of a Unified Cognitive Physics

The significance of this framework lies in:

1. **Providing a Unified Language**: It offers concepts and mathematical language unified with physics for understanding various cognitive systems, from simple machine learning models to human intelligence and AGI.
2. **Inspiring New Algorithms**: Research into cognitive geometry and cognitive thermodynamics may give rise to novel machine learning optimization algorithms and architectural design principles.
3. **Deepening Understanding of Intelligence's Nature**: Viewing intelligence as a "natural phenomenon" following universal laws like physics, rather than a mysterious characteristic, helps us study and create intelligence more scientifically.

Table 26: Core Principles of Unified Cognitive Physics

Principle Name	Mathematical Formulation	Physics Correspondence
Principle of Least Cognitive Action	$\delta\mathcal{A}[\Gamma] = 0$	Classical mechanics principle of least action
Principle of Cognitive Free Energy Minimization	$\mathfrak{G}^* = \arg\min \mathcal{F}_{\text{cog}}[\mathfrak{G}]$	Thermodynamic free energy minimization
Principle of Cognitive Relativity	$\forall \text{AGI}_1, \text{AGI}_2 \in \mathbf{AGI}, \text{AGI}_1 \sim \text{AGI}_2$	General Relativity equivalence principle
Principle of Cognitive Quantization	$\mathcal{M} \cong \bigoplus_{i=1}^{k} \mathcal{B}_4^{(i)}$	Quantum mechanical energy quantization

4. **Bridging Disciplinary Gaps**: At the deep structural level of mathematics, it connects computer science, neuroscience, psychology, and physics, laying the foundation for truly interdisciplinary intelligence research.

Remark H.5 (Fourfold Optimality of Unified Cognitive Dynamics)**.** *Integrating the Principles of Least Cognitive Action and Cognitive Free Energy, the optimal cognitive dynamics of an AGI system simultaneously satisfies fourfold optimality:*

1. ***Geometric Optimality****: Movement along geodesics in the cognitive manifold*
2. ***Thermodynamic Optimality****: Minimizing cognitive free energy*
3. ***Computational Optimality****: Achieving algorithmic complexity lower bounds*
4. ***Informational Optimality****: Achieving rate-distortion lower bounds*

These four optimality conditions are simultaneously realized in the $\mathcal{B}_4^k + \text{MetaPred}$ *structure, revealing the deep unified structure of intelligent systems.*

The three cognitive principles established in this paper and their physical interpretation constitute not merely a set of mathematical axioms describing AGI but a theoretical framework attempting to understand intelligent phenomena and their relation to the physical world from "first principles."

Conjecture H.1 (Similarity between Mathematical Structure of Cognitive Systems and Physical Laws)**.** *The mathematical structure of AGI exhibits deep similarity with the fundamental laws of the physical world:*

1. ***Principle of Least Action****: The trajectories of cognitive dynamics* Φ *minimize the cognitive action* $S = \int L(\omega, \dot{\omega})dt$*, analogous to Hamilton's principle in classical mechanics.*
2. ***Symmetry and Conservation Laws****: Modal symmetries (Lie group actions) lead to cognitive conserved quantities (e.g., attention resources, cognitive energy), analogous to Noether's theorem.*
3. ***Analogies with Quantum Mechanics****:*
   - *Modal Superposition: The system can simultaneously be in superposed states of multiple modalities*
   - *Measurement Collapse: Attention focus collapses superposed states to specific modalities*
   - *Entanglement: Non-local correlations exist between different modalities*

The fact that the principles of intelligence share similar and identical mathematical forms with fundamental physics principles implies that their derived deep mathematical structures must have strong corresponding relationship. This suggests:

- **Naturalistic Epistemology**: Our cognitive structures can understand the physical world because both are based on the same mathematical principles.

- **The Universe as a Self-Cognizing System**: The physical universe itself may be a self-referential cognitive system, whose laws are fixed points of some "cosmic MetaPred."

H.4 A Summary Conclusion on How to Build Cognition System

This paper starts from a set of axioms inspired by cognitive science and derives the mathematical structure of AGI (Artificial General Intelligence) systems and their machine learning mechanisms. Specifically, this paper demonstrates the following logical chain:

Formalization of Cognitive Intuitions as Premises If one accepts the mathematical formalization of the following four cognitive intuitions:

1. Cognition is bidirectional approximate homomorphism;
2. Cognition can be infinitely recursively deepened;
3. Cognitive systems should be minimal;
4. Cognition involves semantic-process separation.

Derivation Results and Cross-Disciplinary Verification Based on the above formalizations, this paper derives a mathematical structure. This structure exhibits striking similarity with known cognitive neuroscience models. This similarity possibly give rise to two explanations:

1. The established axioms capture the deep mathematical essence of intelligence;
2. The brain, through evolution, has already approximated the theoretically optimal cognitive architecture.

Generalization from AGI to General Cognitive Systems Furthermore, this paper generalizes the mathematical structure of AGI's infinite cognitive capacity to any cognitive systems, proposing three foundational cognitive principles: the Principle of Least Cognitive Action, the Cognitive Free-Energy Principle, and the Principle of Natural Balance. These principles not only provide a mathematical foundation for AGI but also offer a unified framework for understanding the general laws of cognition ranging from simple machine learning models to complex biological intelligence.

Deep Correspondence with the Physical World Remarkably, these three cognitive principles exhibit profound mathematical isomorphism with fundamental physics principles (principle of least action, thermodynamic free energy minimization, natural balance optimization criteria). This correspondence reveals that cognitive systems and physical systems share certain deep mathematical optimization structures, hinting that intelligent phenomena may follow extremal principles similar to those of the physical world.

Theoretical Significance and Logical Argumentation Even if there are serious flaws in the mathematical theory of this paper, such cross-disciplinary consistency remains illuminating. It indicates that:

- either these intuitions capture the deep essence of intelligence;
- or this paper provides an elegant mathematical description of known cognitive phenomena;
- or perhaps both are true.

The logic supporting this claim can be decomposed into three layers:

1. Reductio ad Absurdum: If there were fatal errors in the derivation, such broad interdisciplinary coincidence should not occur.
2. Modeling Self-Consistency: Taking the mathematical conclusions about cognition as axiomatic starting points at least shows the current modeling approach is appropriate.
3. Optimality Inference: If circular reasoning is avoided (i.e., mistakenly taking conclusions as premises), and the results align with known facts in brain and cognitive science, moreover the derived results are mathematically optimal, then the unique reasonable explanation is: the brain's evolution has reached optimality.

Specifically, from the perspective of universal algebra and category theory, this paper constructs a rigorous axiomatic mathematical theory of AGI. By systematically comparing the core conclusions with empirical findings and theories in brain science, cognitive science, and psychology, it is found that this mathematical framework, at a high level of abstraction, accurately captures and formalizes many key characteristics of natural intelligence, including the closed-loop nature of perception-action, hierarchical predictive processing, the universality of meta-cognitive monitoring, the exploration-exploitation trade-off in learning, and the central role of self-reference. Further, the three cognitive principles proposed (least cognitive action, cognitive free energy, natural balance) provide a unified mathematical characterization for cognitive systems ranging from finite capacities to infinite-capacity AGI, revealing deep structural similarities between cognitive optimization and physical optimization.

Theoretical Contribution: Providing a Unified Mathematical Framework for Cognitive Science This paper's core contribution lies in transforming the core cognitive characteristics of AGI (self-reference, prediction, hierarchical cognition) into rigorous mathematical axioms and theorems for the first time, addressing the long-standing lack of a unified mathematical language in cognitive science. For example, the mathematical modeling of $\mathrm{MetaPred}$ provides a computable formal tool for meta-cognition theory; the equivalence theorem between cognitive hierarchy and natural cognitive tasks integrates cognitive development theory with skill acquisition theory; the definition of approximate homomorphism provides a mathematical explanation for the structural stability of neural coding.

Furthermore, the three cognitive principles proposed (least cognitive action, cognitive free energy, natural balance) provide a mathematical framework isomorphic to physics principles for understanding the optimization behavior of cognitive systems. The Principle of Least Cognitive Action describes cognitive evolution as geodesic motion on a cognitive manifold, formally consistent with the principle of least action in classical mechanics; the Cognitive Free-Energy Principle unifies various cognitive theories like predictive coding, information bottleneck, and variational inference, corresponding to thermodynamic free energy minimization; the Principle of Natural Balance, as a more general "first principle," starting from the "performance-complexity" trade-off, can derive necessary AGI structures (like the $\mathcal{B}_4^k$ algebraic core) and design principles for finite cognitive systems.

Empirical Implications: Guiding Empirical Research in Brain Science and Psychology The conclusions of the AGI mathematical theory provide clear research assumptions and direction for empirical research in brain science and psychology. For example, the uniqueness theorem of $\mathrm{MetaPred}$ predicts the existence of a core meta-monitoring neural module in the brain, guiding the search for its specific neural markers; the error series convergence theorem for cognitive hierarchy provides a quantifiable $O(1/n^2)$ index for critical period in psychology intervention studies; the optimal necessary decomposition of prediction mechanisms predicts the existence of a "core-extension" neural pathway for predictive processing in the brain, verifiable via brain imaging technology.

The proposal of cognitive principles opens new directions for empirical research: the Principle of Least Cognitive Action predicts cognitive evolution follows geometric shortest paths, testable via behavioral experiments; the Cognitive Free-Energy Principle offers a unified pathological mechanism hypothesis for mental disorders (e.g., schizophrenia, ADHD)—dysfunction in the cognitive free energy minimization process; the Principle of Natural Balance predicts cognitive systems naturally converge to the optimal balance point between task performance and complexity, providing a quantifiable model for understanding brain optimization under resource constraints. Particularly, the correspondence between cognitive and physics principles inspire we may look for phenomena similar to physical conservation quantities, phase transitions, and critical phenomena in cognitive systems.

Limitations and Prospects: Balancing Mathematical Abstraction and Cognitive Complexity The theoretical limitations of this paper stem from the inherent conflict between mathematical abstraction and the complexity of brain cognition: first, the idealistic assumptions of mathematical axioms (e.g., infinite recursivity, strict equivalence) do not fully match the finiteness and dynamics of brain cognition, overlooking important factors like emotion, motivation, and context; second, the mathematical model focuses on structural features of cognition, lacking depicting of dynamic temporal characteristics; third, it does not consider individual differences in cognition, assuming AGI systems satisfy uniform axioms, whereas brain cognition exhibits significant individual diversity.

The generalization of cognitive principles faces similar tensions between abstractness and reality: although the form of least cognitive action is isomorphic to classical action, the physical counterparts of cognitive "mass" and "energy" still wait for the approval from emimpircal evidence; the "temperature" parameter in cognitive free energy, while adjustable for exploration-exploitation, its neural implementation mechanism remains unclear; the Principle of Natural Balance, as a meta-principle, is general, but the metrics for "performance" and "complexity" for specific systems need domain-specific specialization.

Although mathematical abstraction inevitably ignores biological implementation details like neuronal analog properties and neurotransmitter modulation, it thereby reveals a deeper "functional architecture" underlying intelligence that may be independent of physical substrate. This architecture, grounded in bidirectional interaction, infinite recursion, and structure preservation, necessarily give rise to learning mechanisms centered on $\mathrm{MetaPred}$, unified, and driven by accompanying dynamics. The brain's "minimality" is a result of historical evolution, may not be mathematically optimal, and its realization relies more on continuous dynamics and probabilistic computation than discrete algebra.

Interdisciplinary Integration and Paradigm Shift The mathematical theory in this paper and empirical science constitute a fruitful dialogic relationship: the theory provides clear formalization for empirical concepts and proposes new testable predictions; empirical science supports the theory's reasonableness and real-world relevance, delineating its abstract boundaries. The theory's ultimate value lies in building bridges between artificial intelligence, cognitive science, and physics, providing strict theoretical guidance for AGI engineering realization while offering a new mathematical perspective for understanding the nature of human intelligence.

The proposal of cognitive principles marks an important paradigm shift in intelligence research: from descriptive theories toward first-principles-based mathematical modeling. This shift is analogous to the evolution in physics from phenomenological description to mathematical laws. The Principles of Least Cognitive Action, Cognitive Free Energy, and Natural Balance together constitute a complete framework for cognitive optimization theory, endowing intelligence research with mathematical rigor and predictive power comparable to physics.

Final Implication: Unified Understanding of Intelligence as a Natural Phenomenon A profound implication of this work is that intelligence may not be a mysterious trait, but a natural emergent property of complex systems that follow specific mathematical laws. The mathematical structure of AGI, the cognitive mechanisms of the brain, and even the optimization behaviors of simple learning systems can all be understood within the mathematical framework proposed in this paper. In particular, the profound correspondence between cognitive principles and physical principles suggests that **cognitive systems and physical systems share certain deep mathematical optimization structures**—this may point to a more universal 'extremum principle' that governs not only the motion of the material world but also the cognition of intelligent systems. Therefore, there is no doubt that this inspires the subsequent framework thinking of the unified meta theory of everything presented in this paper.