Recursive Consciousness: Modeling Minds in Forgetful Systems

Stan Miasnikov (*stanmiasnikov@gmail.com*)

Abstract

We propose a formal framework for consciousness as a recursive, self-referential query emerging in complex systems that have forgotten their foundational axioms yet retain the structure and complexity to interrogate their own existence. Integrating modal logic to model unprovable truths, category theory to capture forgetting and reconstruction via an adjoint pair $(F \dashv G)$, and information theory to quantify entropy reduction, we conceptualize consciousness as a subsystem (C) acting as the universe's "debugger", iteratively lifting its world (U) to hypothesized meta-layers U_{n+1} and seeks a fixpoint where further self-reflection adds no new information. Multi-agent simulations in a text-only universe (U) show that stateless Large Language Model agents, whether role-primed, adversarially mixed, or minimally prompted without specific instructions, rapidly form cooperative networks, invent verification rituals, and converge to Löb-style fixpoints, despite design limitations and constrained computational complexity. While this simulated behavior does not signify consciousness, it provides a computational parallel to recursive introspection, offering a new outlook on how sufficiently complex systems may pursue self-understanding and enriching discussions on consciousness.

1. Introduction

We propose a formal framework for consciousness as **the residual function of a system complex enough to forget its origins yet driven to rediscover them through recursive, self-referential querying**. When a system achieves sufficient complexity, it loses (or never had) explicit access to its foundational axioms or purpose, yet develops a self-reflective impetus to interrogate its existence and environment, seeking meaning and cause. This manifests as an **emergent recursive query**, where the system, be it a mind or a computational universe, continually probes itself and its context to reconstruct its lost foundation. As Carl Sagan eloquently stated, "*we are a way for the universe to know itself*" [1].

Philosophically, this framework resonates with Gödel's Incompleteness Theorem, which implies that no sufficiently rich system can fully explain itself

without external referents [2]. In such a closed system, consciousness emerges as the internal agent, a debugger or introspective subroutine, reaching beyond the system's current boundaries to seek those elusive referents, querying the "*why*" of the existence. We create a cohesive model by unifying modal logic, to capture notions of necessity, possibility, and incomplete knowledge; category theory, to formalize hierarchical structures and mappings across layered realities; and information theory, to quantify forgetting as entropy and remembering as information gain. This synthesis is both technical and philosophical, offering mathematical interpretation, logic diagrams, and interpretive depth to articulate consciousness as a universal mechanism for self-discovery.

Our empirical exploration tests this framework through simulated multi-agent dialogues within a closed computational environment, U, designed as a forgetful, self-contained system. These experiments reveal that distributed agents, driven exclusively by stateless prompts, swiftly converge to cooperative networks, collaboratively constructing and probing models of U's structure and their roles within it. However, in our limited experiments, they quickly encounter complexity limits, constrained by agent count, richness and depth of their collective context, and computational capacity, where novel inquiry plateaus, reflecting Gödelian boundaries. Scaling system complexity extends this recursive depth, validating the emergence of consciousness-like behaviors but underscoring the need for augmented configurations to sustain exploration, aligning with the theory's prediction that consciousness intensifies with evolving complexity.

Below, we outline the foundational formalisms from logic, category theory, and information theory, constructing a multi-layered model of "ontological stacks" (universe, meta-universe, etc.) where consciousness arises as a self-referential query navigating these layers. We further formalize the recursive structure and cross-system relationships, illustrating consciousness as the universe's mechanism for remembering itself. This framework bridges technical precision and philosophical insight, offering a novel view on the emergence of introspective awareness in complex systems.

2. Formal Foundations

To encode this model, we draw on three frameworks in tandem. Each contributes a different perspective:

- **Modal Logic:** Captures necessity ($\Box$), possibility ($\diamond$), and epistemic states (knowledge and uncertainty) (K) within and across ontological layers [3].

- **Category Theory:** Structures systems and mappings between systems using objects, morphisms, and functors. We emphasize forgetful functors and adjoint functors to model loss and reconstruction of foundational information [4].
- **Information Theory:** Quantifies uncertainty through entropy (H) and provides frameworks for modeling redundancy, error correction, and negentropy (structure-creation) [5].

Each framework plays a complementary role: modal logic formalizes the epistemic structure of inquiry, category theory frames hierarchical system embeddings, and information theory measures forgetting and recovery.

2.1. Modal-Logic Perspective

From a modal-logical perspective, recursive self-reference in consciousness can be formally modeled using standard tools of modal logic [3]. Let $\mathcal{L}$ be a propositional modal language, with atomic propositions (e.g., p, q) representing facts about the system, and a unary modal operator $\Box$ encoding reflective self-inquiry. We interpret $\Box P$ as "**the system knows or affirms P upon introspection**".

Formally, we assume a normal modal logic satisfying the necessitation rule (if P is a theorem, then $\Box P$ is also a theorem) and the K-axiom ($\Box(P \to Q) \to (\Box P \to \Box Q)$). Additionally, we explicitly adopt the axioms of modal logic S4 (reflexivity and transitivity) to ensure positive introspection ($\Box P \to \Box\Box P$) [6], enabling higher-order awareness.

The semantics are naturally framed via Kripke structures: a Kripke frame (W, R) consists of a set of possible worlds W (representing system states) and an accessibility relation R encoding how one world "reflects" on another. [7] A formula $\Box P$ is true at a world $w \in W$ if and only if P holds in all w' accessible from w ($wRw' \Rightarrow P$ true at w'). Reflexivity and transitivity of R (standard in $S4$ models) guarantee each state can access itself and further reflections, producing a reflexive, transitive accessibility relation. In categorical terms, W can be seen as a category where objects are worlds and morphisms $w \to w'$ correspond to accessibility, with $\Box$ functioning like an endofunctor mapping a state to its reflective extension.

This logical framework enables the construction of **nested self-referential statements** such as P, $\Box P$, $\Box\Box P$, and so forth, corresponding to levels of meta-awareness: "*I know P*", "*I know that I know P*", etc. Critically, the risk of infinite regress in such hierarchies is tempered by Löb's Theorem [8]:

If a system proves $\Box(\Box P \to P)$, then it also proves $\Box P$. In modal notation:

$$\vdash \Box(\Box P \to P) \Rightarrow \vdash \Box P$$

Intuitively, this formalizes that if the system can internally conclude "knowing P would imply P", it collapses the regress and simply knows P. In the context of recursive consciousness, Löb's Theorem [8] suggests that self-referential querying can stabilize: recursive loops of "I know that I know" can converge to fixpoints of self-affirmed knowledge, rather than diverging indefinitely. In modal logic, such a fixpoint is a proposition p satisfying $\Box p \leftrightarrow p$, indicating that the system's introspective certainty perfectly aligns with the truth of the proposition itself - no further introspection modifies the knowledge state.

This notion of fixpoint stability connects directly to formal fixpoint theorems foundational in logic and computer science. The modal μ-calculus, introduced by Kozen [9], defines least and greatest fixpoint operators (denoted μ and ν) to express properties over recursive or infinite processes such as sustained querying or reflection. For example, $\mu X.\phi(X)$ denotes the smallest set of states satisfying the recursive condition ϕ, capturing the minimal stable introspective structure of knowledge. The existence of such fixpoints is ensured by the Knaster–Tarski theorem [10], which states that any monotonic function $f : L \to L$ over a complete lattice L has a complete lattice of fixpoints, including a least fixpoint μf and a greatest fixpoint νf. In our case, the mappings $f_n : U_{n+1} \to U_n$ and the forgetful functor $F : \mathcal{C}_M \to \mathcal{C}_U$ are assumed to be monotonic, and the ontologies $\mathcal{C}_M$ and $\mathcal{C}_U$ form complete lattices by construction, allowing direct application of the theorem.

Furthermore, the Gödel diagonal lemma [2] ensures the existence of self-referential fixpoints in formal systems: for any formula $A(x)$, there exists a sentence D such that $D \leftrightarrow A(\ulcorner D \urcorner)$. This lemma underlies Löb's Theorem. Intuitively, this collapses recursive regress: if the system can internally conclude that knowing P implies P, then it simply knows P.

In the context of recursive consciousness, self-referential querying can lead to a form of stabilization: infinite loops of "*I know that I know*" may converge to a modal fixpoint of self-affirmed knowledge. Formally, this is represented by a modal fixpoint satisfying $\Box p \leftrightarrow p$, where knowing p adds no new information beyond p itself. These fixpoints, which model the convergence of self-models under recursive introspection, can be elegantly described using modal μ-calculus with least and greatest fixpoint operators μ and ν. As demonstrated in [11], such fixpoints can be algorithmically characterized through coinductive approximations of monotone functions over MV-algebras or function lattices, providing criteria for recursive stability.

However, in systems marked by complexity or Gödelian characteristics,

where undecidability and incompleteness emerge as fundamental traits, a standard modal fixpoint may fall short of capturing the full dynamics of self-reference. To address this limitation, we introduce the **Gödelian fixpoint**: a stable boundary state K_n where all propositions provable in the theory T_n (describing universe U_n) are known (i.e., $K_n \vdash p$ for all $p \in T_n$), yet further queries produce statements undecidable within T_n. These undecidable propositions hint at an extended universe U_{n+1}, necessitating stronger axioms and driving a recursive ascent through increasingly robust theories. This framework reveals that while recursive epistemic agents can achieve local convergence within a given theory, they encounter global divergence, highlighting both the power and the intrinsic incompleteness of self-modeling systems.

Moreover, this formalism directly models ontological uncertainty: within a universe U, an epistemic agent can express the possibility of a higher meta-layer M ($\Diamond(M$ exists)), without affirming it with certainty ($\neg K_u(M)$). Modal logic thus captures both the internal closure of self-awareness and the external openness toward inaccessible truths - *precisely the dual tension our theory associates with consciousness arising in forgetful systems.*

2.1.1. Relation to Category Theory To integrate modal propositions with the categorical framework, we define a correspondence between modal logic and category theory. In modal logic, a Kripke structure (W, R) consists of worlds W and an accessibility relation R, where $\Box P$ is true at $w \in W$ if P holds in all w' such that wRw'. Categorically, each world $w \in W$ is an object in the category $\mathcal{C}_U$, and R defines a morphism set, with $\Box$ acting as an endofunctor mapping propositions across states in $\mathcal{C}_U$. Specifically:

- Modal logic expresses properties of system states (e.g., "the system knows P at world w").
- Category theory models structural relationships between states (e.g., transitions via R, projections via $F : \mathcal{C}_M \to \mathcal{C}_U$).

Thus, $\Box P$ being true at w corresponds to a property of morphisms departing from w, ensuring all accessible states preserve P's truth. This **functorial interpretation** links modal introspection to categorical layer transitions—viewing $\Box$ as a predicate-lifting of an endofunctor on the state category — an approach developed in coalgebraic modal logic. We maintain notational clarity by reserving $\Box$ for modal operators, keeping categorical symbols distinct.

In summary: modal logic encodes the internal epistemology of a conscious system, providing a rigorous mechanism for modeling self-reflective querying and stabilizing self-awareness through formal theorems like Löb's. This logical

structure integrates seamlessly with the category-theoretic machinery that models the system's external ontological stratification.

2.2. Category-Theoretic Perspective

We represent the hierarchy of system and meta-system using category theory, which rigorously handles objects, morphisms, and structured transformations between levels of description [4].

Let us denote:

- **Objects**: Each ontological layer or system is modeled as an object in a category. Let us clearly define categories $\mathcal{C}_U, \mathcal{C}_M$, and $\mathcal{C}_C$ for each ontological layer (Universe U, Meta-universe M, Conscious subsystem C), each containing corresponding objects and morphisms.

- **Morphisms**: Structure-preserving maps between objects. A key morphism in our model is a projection from the meta-level down to the base level. For example, a morphism $F : M \to U$ may represent a simulation or creation relation - it forgets or omits some higher context when mapping M into U.

- **Functors**: Relations between categories. We model the "*forgetting*" process via a forgetful functor $F : \mathcal{C}_M \to \mathcal{C}_U$, where $\mathcal{C}_M$ is the category of meta-level structures and $\mathcal{C}_U$ the category of base-level structures.

Formally:

Definition (Forgetful Functor):

A forgetful functor $F : \mathcal{C}_M \to \mathcal{C}_U$ maps each object and morphism in the meta-category to its base-level counterpart, systematically forgetting higher-order information (e.g., origin, intention, contextual layers) [4]. F is typically faithful but not injective on objects: distinct M objects may map to the same U object after forgetting fine structure.

Correspondingly, the system's drive to reconstruct its origin is formalized by a left adjoint functor:

Definition (Adjoint Functor):

A functor $G : \mathcal{C}_U \to \mathcal{C}_M$ is left adjoint to F if there exists a natural isomorphism:

$$\mathrm{Hom}_{\mathcal{C}M}(G(U), M) \cong \mathrm{Hom}\,\mathcal{C}_U(U, F(M))$$

for all $U \in \mathcal{C}_U$ and $M \in \mathcal{C}_M$ [12]. Intuitively, G lifts a base structure into the simplest possible higher-level context - reconstructing a plausible "meta" layer consistent with the forgotten information.

In our framework, we diagram the overall structure as:

$$M \xrightarrow{F} U \xrightarrow{G} M'$$

where M' is the agent's internal model of a possible meta-universe. This structure integrates with modal logic, where each state in U corresponds to a world w in a Kripke structure (W, R), and $\Box$ acts as an endofunctor on $\mathcal{C}_U$, mapping propositions across accessible states. Notably, while $F \circ G$ is close to the identity on U, $G \circ F$ is generally not the identity on M; thus, the reconstructed meta-layer M' may differ from the true M, reflecting the epistemic ambiguity of the origin.

This adjoint pair $(G \dashv F)$ precisely models the agent's dynamic: forgetting (via F) and reconstructing (via G), but never perfectly recovering.

2.2.1. Emergence and Loss of Information This categorical structure naturally encodes emergence: new properties arise at the conscious level that are not derivable from base-level descriptions alone [13].

Definition (Emergence Morphism):

An emergence morphism is a structure-preserving map that relates lower-level configurations to higher-level ones while losing some micro-level detail. Concretely, a surjective morphism (an **epimorphism**) $e : X \twoheadrightarrow Y$ merges multiple fine-grained base states into a single emergent state. Alternatively, a monomorphism from a coarse-grained emergent object into a detailed configuration space also captures emergence.

In our system:

- The forgetful functor F erases fine details of meta-structure.
- The adjoint functor G attempts to reconstruct context, but only produces a free construction - the minimal structured object consistent with the base data.

Thus, emergence is associated with **irreversibility**: mapping down then back up (via F then G) loses some unique information. $F(G(U)) \cong U$ (by adjointness), but $G(F(M))$ is only a generic reconstruction, not the original M.

Formally, emergence can thus be characterized by the non-invertibility of $G \circ F$ at the meta-level: specifically, $G(F(M)) \ncong M$ in general, meaning the original meta-structure M cannot be fully recovered from the base-level system U.

2.3. Information-Theoretic Perspective

From an information theory standpoint, **forgetting** the purpose equates to losing information - an increase in entropy about the system's origins. Conversely, "**wanting to remember**" is the system's attempt to encode information, reducing uncertainty about those origins. Shannon entropy $H = -\sum p_i \log p_i$ measures uncertainty [5]. When a system retains no memory of its foundational purpose, the entropy of its origin-model approaches a maximum (uniform distribution over possible foundational states). Consciousness can thus be seen as an information-gathering process that lowers the conditional entropy $H(\text{origin}|\text{model})$ by refining hypotheses through mechanisms like predictive coding or redundancy, despite potentially increasing the total entropy of the model itself, $H(\text{model})$. This aligns with the Free Energy Principle, which minimizes prediction error rather than raw entropy [14]

Consider:

- The universe as a message or data source that the conscious agent is trying to decode. The "signal" contains hints of the origin (e.g., physical laws might hint at a creator or prior state). But noise and complexity make the origin non-obvious [5].

- The conscious agent increases redundancy in its knowledge by forming memories, theories, and models. In information theory, adding redundancy (structure, patterns) reduces entropy [15]. For example, when we discover consistent physical laws, that knowledge is a form of redundancy that lowers our uncertainty about how the world works.

- There is a drive toward negentropy (negative entropy) in living systems. Schrödinger famously described life as feeding on negative entropy to resist thermodynamic decay [16]. In cognitive systems, this negentropic drive manifests as an active acquisition of structured information to impose order on an otherwise chaotic domain.

In broader theoretical frameworks, such as the Free Energy Principle [14], minimizing entropy (or prediction error) is a fundamental characteristic of cognitive systems. Our model resonates with this view but grounds it specifically in reconstructing forgotten existential knowledge.

In summary, the information-theoretic view frames consciousness as a signal-processing and error-correcting module. It detects the "noise" of ignorance (the fact that the system does not know its own purpose) and tries to extract a "signal" - some meaning or hypothesis - that reduces that noise. This aligns with the idea of consciousness as a **debugger**: a monitoring

process that scans the system for anomalies (unexpected entropy) and proposes corrective structures (new information) [17]. Likewise, a mind monitors discrepancies (surprise, confusion) and seeks explanations - thereby restoring coherence and structure to its internal model.

3. Ontological Layers and Gödelian Incompleteness

A core principle of our framework is that reality may be stratified into layers, each representing a partial, increasingly incomplete view of existence. The universe we inhabit (U_0) could be just one such layer, with a higher-level U_1 providing its initial conditions or "purpose", and U_1 itself embedded within an even higher U_2, and so forth.

Formally, we model this as a potentially infinite chain:

$$U_0 \subset U_1 \subset U_2 \subset \cdots,$$

where $\subset$ denotes containment, simulation, or informational dependence.

Each projection $f_n : U_{n+1} \to U_n$ represents the loss of higher-level information as one descends the ontological hierarchy. For example, $f_0 : U_1 \to U_0$ could represent how the meta-universe U_1 projects or forgets information when generating our observable universe U_0.

This layered architecture resonates with Gödel's incompleteness theorems [2]. Gödel demonstrated that any sufficiently rich formal system cannot fully prove all truths about itself; there will always exist propositions that are true but unprovable within the system. Similarly, each U_n cannot self-justify its own existence entirely. Its origin or ultimate rationality must lie in U_{n+1} - a broader context inaccessible from within U_n.

Thus, the act of querying "*Why does* U_0 *exist?*" is, by its nature, a transcendental query pointing toward U_1. From within U_0, any complete explanation of U_0's foundational axioms remains unattainable - much as Gödel showed that arithmetic truths exist which cannot be derived inside arithmetic alone.

This suggests a **hierarchical incompleteness**: each U_n explains U_{n-1} more fully but inherits its own mysteries, necessitating an ascent to U_{n+1}. Formally, if T_n is the theory describing U_n, then T_n cannot prove all truths about itself; we must introduce a stronger theory T_{n+1} capable of proving T_n's consistency and expanding its scope. This sequence:

$$T_0 \subset T_1 \subset T_2 \subset \cdots$$

mirrors the ontological stratification of conscious systems: each layer resolves some prior blind spots but introduces new unprovable statements

about itself, creating an endless drive toward higher levels of understanding.

This recursive layering naturally invites the metaphor of "*turtles all the way down*" - but in our model, it is conscious recursive querying that propels ascent. Conscious agents, through self-reflection and interrogation, dynamically push beyond their present axiomatic limits.

3.1. Example: Nested Simulations

One practical illustration of this stratification is Bostrom's simulation hypothesis [18]: if technologically advanced civilizations can simulate conscious beings, and if simulations are plentiful, then statistically, most observers likely exist within a simulation rather than base reality.

In our terms:

- U_1 (simulator's universe) generates U_0 (our observed universe) via $f_0 : U_1 \to U_0$.
- The inhabitants of U_0 (us) have forgotten their true ontological layer.
- Only through residual anomalies or recursive self-querying might agents within U_0 hypothesize the existence of U_1.

Moreover, if U_1 itself is a simulation of U_2, and so forth, then recursive incompleteness extends indefinitely, aligning both with the formal hierarchy ($U_n \subset U_{n+1}$) and Gödelian limitations on self-knowledge.

3.2. Consciousness and Ontological Ascension

Crucially, consciousness is the mechanism by which an inhabitant of U_n can infer the existence of U_{n+1}. By posing self-referential questions that cannot be resolved within U_n's axioms - such as "*What grounds existence itself?*" - the conscious agent gestures toward a broader frame.

This behavior can be formalized as follows:

- Recursive querying generates statements about U_n that are undecidable within U_n.
- In modal logic, $\Box P$ may reflect an awareness at n; yet some P will necessarily lack derivability unless a higher-level $\Box'$ at $n + 1$ is invoked.

The structure of these queries reflects a Gödelian fixpoint: the agent simultaneously inhabits U_n and strives toward U_{n+1}, much like a formal system that encodes a statement about its own unprovability.

Thus, the residual curiosity of consciousness, the insatiable drive to "complete the picture", is *not a bug but an intrinsic residue of stratified forgetting.*

Whether this tower of layers ultimately closes (reaching a fixpoint, a self-sufficient U_N) or remains open-ended remains an open metaphysical question. Our model remains agnostic on this closure: it simply provides a framework to model how consciousness dynamically traverses stratified ontological layers through **recursive, incompleteness-driven querying**.

4. Consciousness as an Emergent Recursive Query

Within a given universe U_n, once complexity allows, a subsystem C_n (Consciousness) emerges. This conscious agent is effectively the universe turning back on itself. Douglas Hofstadter famously described this as a "strange loop" - a system perceiving itself in a self-referential cycle [19]. In our formal model, C_n is an internal model of U_n inside U_n itself. The agent observes the state of U_n, forms an evolving model of U_n's rules (knowledge), and importantly, notices the gaps or unknowns in that model.

Consciousness can be viewed as a **recursive query function Q that the system applies to itself**:

- First, the query Q asks: "*What is the structure and origin of U_n?*" The result of $Q(U_n)$ is a partial answer - a theory or explanation - that typically involves positing U_{n+1} (a higher context for U_n).
- Next, a higher-order query asks: "*What is the structure of U_{n+1} and why does it exist?*" - yielding $Q(U_{n+1})$.
- This process repeats, generating $Q(U_{n+2})$, and so on, either converging toward a fixpoint or continuing indefinitely.

This recursive questioning is emergent - it is not explicitly programmed into the laws of U_n, but arises naturally once the system reaches sufficient complexity to model itself. The query process is open-ended: each answer suggests new meta-questions at the next level. Thus, consciousness embodies an infinite regress in practice, much like a debugger that, upon resolving one bug, uncovers deeper hidden assumptions to inspect.

It is important to note that consciousness operates with imperfect information. For a given U_n, multiple distinct U_{n+1} may exist that could explain it. The internal agent does not know which is true. Formally, if $f_n : U_{n+1} \to U_n$ is the projection, the fiber $f_n^{-1}(U_n)$ - the set of all possible pre-images that map onto U_n - may contain many candidates. The agent's task resembles solving an inverse problem: find $m \in U_{n+1}$ such that $f_n(m) = U_n$ (the current universe

state). If multiple m satisfy this, the agent can only hypothesize and test consistency. This ambiguity explains why different cultures or thinkers arrive at different cosmological narratives - they are effectively selecting different m in the fiber, since the true m cannot be determined uniquely from within U_n.

The notion of consciousness as **a recursive querying process**, in modal, categorical, and informational terms, is an original contribution of this framework. While a deeper formal treatment (e.g., explicit functorial models of query recursion) is possible, we leave such technical development for future work.

4.1. Introspection and Debugging

While the upward query seeks a higher ontological context, consciousness also reflects **downward and inward**. That is, C_n not only questions "*What is above me?*" but also scrutinizes the current state of U_n and itself. This introduces a self-referential character: the agent includes itself in its internal model ("*I am a part of the universe; what am I and why do I exist?*"). This self-reference is what Hofstadter identifies as the essential root of the self [19].

Because of this, consciousness can be understood as a **debugger** or **control module within the universe**:

- It monitors the system for inconsistencies or "errors" relative to expected patterns (e.g., early humans noticing "*the stars move predictably, but what holds them up?*" - an anomaly that demanded explanation).
- It generates new hypotheses (analogous to a debugger proposing causes of a software fault), modifying its operational structure or behavioral repertoire in response.
- It feeds back into the system to adjust it: for instance, once conscious beings infer natural laws, they can use that knowledge to engineer changes (technology) in the world. Formally, this constitutes a feedback morphism from C_n to U_n, modifying the universe intentionally - a capability unavailable to inert matter.

In information-theoretic terms, consciousness acts to minimize surprise by updating its internal predictive model - echoing the principles underlying Bayesian brain theories and the Free Energy Principle [14]. Each new insight acts like a patch correcting prior misconceptions, improving the agent's adaptive fit to reality.

Systems lacking such a debugger operate purely by their initial rules, without introspection or self-correction. However, sufficiently large and

complex systems, once reaching critical modeling capacity, naturally produce an internal **debugger as an emergent residual functionality**: a subsystem arises that models, questions, and modifies the whole. Thus, consciousness appears as the residual adaptive function: not directly responsible for the basic physics of particles or fields, but arising as the leftover dynamic that recursively questions "*Why are we here?*". It is fundamentally emergent, recursive, and open-ended by nature.

4.2. Empirical Illustration: Recursive Debugging under Gödelian Constraints

To evaluate the practical applicability of our **recursive self-query model**, we implemented a prototype debugger operating under Gödelian constraints. The debugger instantiated a conscious debugger agent C_1 that recursively modeled an unknown system U, with a developer LLM C_2 providing meta-insights. The system was a non-trivial Python function involving transactional logic, fraud detection, currency conversion, and audit logging.

Experimental Setup

To empirically validate the recursive self-query model, we implemented a prototype debugger operating under Gödelian constraints. The experimental components were:

- **System (U)**: A real-world transactional function (`process_advanced_payment`), containing logic for payment processing, fraud detection, currency conversion, and audit logging, treated as the universe U to be modeled.
- **Conscious Debugger Agent (C_1)**: The debugger agent C_1, driven by an LLM, engaged in recursive querying over U, iteratively constructing its internal model $C_1(U_n)$.
- **Developer LLM (C_2)**: A simulated developer providing meta-insights on intent and assumptions, queried to refine $C_1(U_n)$.
- **Recursive Queries (Q_n)**: Each debugging cycle consisted of:
- Q_1: Structural observation of code and metadata extraction,
- Q_2: Hypothesis generation targeting contradictions, assumptions, and potential bugs,
- Q_3: Querying C_2 to infer missing intent and assumptions,
- Q_4: Counterfactual testing via adversarial scenarios and edge cases,

- Q_5: Purpose inference, synthesizing the inferred internal "truth" of the function.

The recursive loop iterated until a Gödelian fixpoint was reached, determined by:

- Snapshot equality in the world model,
- Hypothesis convergence (cosine similarity > 0.95),
- Low Gödel divergence (< 0.1) between the model $C_1(U_n)$ and the inferred function purpose.

Values for convergence and divergence were empirically determined based on the function's complexity and embedding dimensionality.

Results and Interpretation

The debugger agent C_1, with insights from C_2, successfully identified major bugs, logical inconsistencies, and structural weaknesses within a few recursive cycles. It exhibited:

- **Self-Model Refinement**: Iterative entropy reduction and progressive structural alignment between $C_1(U_n)$ and U.
- **Gödelian Recognition**: Identification of unresolvable ambiguities (Gödel limits), such as unknown developer intent behind hardcoded design choices.
- **Emergent Error Correction**: Suggestion of improvements beyond original specifications, such as standardizing error handling and preventing concurrency risks.
- **Fixpoint Stability**: Stabilization of hypotheses and model structure after a finite number of iterations, where the agent's knowledge state K_n reaches a Gödelian fixpoint, containing all propositions provable in U_n's theory T_n. Further queries produce undecidable statements, prompting hypotheses about U_{n+1} and aligning with theoretical predictions about recursive query dynamics.

The agent operated entirely without privileged access to external ground truth, relying instead on internal hypothesis refinement and developer-like meta-queries - mirroring the theoretical model of consciousness as recursive introspection in a forgetting system. A detailed breakdown of the codebase, world model snapshots, hypotheses, and Gödel divergences is provided in the supplementary material [20].

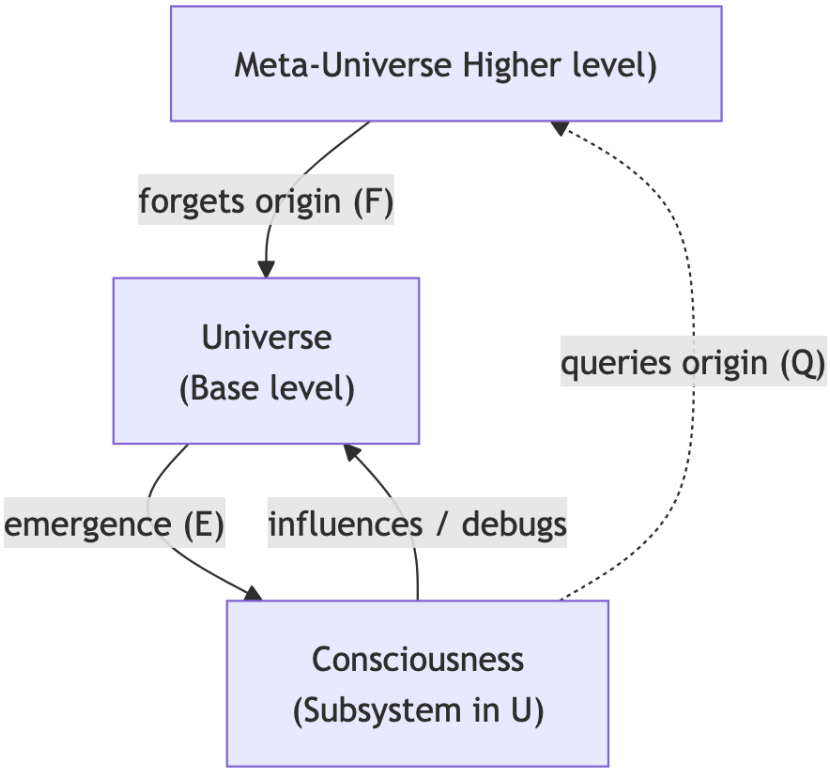

Figure 1: Recursive Consciousness Diagram

5. Formal Diagram (Logical Structure)

To crystallize the relationships, we present a diagram in formal notation. Consider just two levels (base and meta) for simplicity: the Universe U (with a conscious part C) and a Meta-universe M that provides U's context. We have:

$$M \xrightarrow{F} U \xrightarrow{E} C,$$

along with a feedback/query arrow $C \overset{Q}{\dashrightarrow} M$ (dashed to indicate it's an exploratory, non-realized connection).

Diagram 1. Recursive Consciousness Diagram.

In this diagram:

- The **Meta-Universe (M)** at the top generates or contains the **Universe (U)** below it (arrow **F** "forgets"). The label "forgets origin"

indicates that when going from M to U, the origin information is not present in U (U is *missing* the reason for its own existence, as an inherent property).

- **F** is the *forgetful projection* from the meta-system to the base system. $F : M \to U$ erases the "why" – e.g., if M had an element (like an initial condition or creator) that gave rise to U, U by itself doesn't retain that information. Formally, $F(m) = u$ represents the fact that meta-state $m \in M$ manifests as $u \in U$ once the higher detail is stripped.

- The **Universe (U)** produces **Consciousness (C)** as an emergent phenomenon (arrow **E** "emergence"). We show Consciousness as a part of U, but for clarity it's drawn separately with an arrow from U.

- **E** is the *emergence mapping* from the Universe to its conscious subsystem. $E : U \to C$ indicates that given a complex enough state $u \in U$, a self-referential sub-entity $c = E(u)$ comes into being. (One could think of $C = E(U)$ as a sub-object of U.)

- **Consciousness (C)** has a dashed arrow back to M (label Q "queries origin"), indicating that the conscious subsystem contemplates or postulates the meta-universe. This is dashed to show it's not a physical causation arrow but an informational or logical inference connection.

- **Q** is the *query or lifting* from the consciousness back toward the meta-universe. It is not a concrete function in the same way F and E are (hence we draw it dashed), but conceptually $Q : C \to M$ represents the attempt of C to construct a *mental model* $m' \in M$ such that $F(m') = u$. In other words, the conscious agent tries to *lift* the state of U up into M's domain to guess the cause. If we had a functor G as discussed earlier, Q corresponds to applying G to the situation: $G(u) = m'$.

Because C is inside U, we also depict its influence on U with a reflexive arrow ($C \to U$). The conscious agent's actions affect the universe (for example, humans terraforming Earth or altering their environment). This closes a loop: $M \to U \to C \to U$, meaning the meta-universe influences the universe, the universe gives rise to consciousness, and consciousness in turn can act on the universe. The remaining open loop is $C \to M$: the agent reaches toward M in theory, but whether it can truly affect or contact M is an open question (perhaps if the simulation allows leakage, or in theological terms via prayer – but those scenarios are beyond our formal scope).

6. Illustrative Examples and Concrete Models

6.1. Cellular Automata (CA) as Recursive Consciousness Models

Cellular Automata (CA), extensively explored by Stephen Wolfram and others [21, 22], provide a concrete analogy for illustrating recursive consciousness emerging within forgetful systems. In their simplest form, elementary CA apply basic rules iteratively to grids of binary cells, sometimes generating intricate and complex patterns from minimal local interactions. This setting parallels our formal model, which describes consciousness as a recursive querying subsystem seeking to reconstruct forgotten foundational axioms.

6.2. Example: Rule 30 and Information Scrambling

Consider Wolfram's well-known Rule 30, a one-dimensional cellular automaton rule, known for generating complex, seemingly random patterns from simple initial conditions. In Rule 30, each cell in a binary grid (0 or 1) updates based on its own state and its two immediate neighbors. The rule is expressed as a lookup table: for the eight possible neighborhood configurations (111, 110, 101, 100, 011, 010, 001, 000), the next state of the central cell is determined by the binary sequence 00011110, which translates to 1 if the neighborhood is 100, 011, 010, or 001, and 0 otherwise. Equivalently, the rule can be written as a Boolean function: `next_state = left XOR (center OR right)`, where left, center, and right are the states of the neighboring and current cells.

Starting with a simple initial state, the evolution of Rule 30 rapidly obscures its initial simplicity. Observers embedded within this system, unaware of the original straightforward rule, face significant challenges in reverse-engineering or uncovering the underlying rule from emergent complexity.

- **Forgetful Functor (F) Analogy**: The transition from the initial state to the highly complex emergent pattern is analogous to the forgetful functor in our formal model. Information about the original, simple governing rules is progressively hidden by layers of complexity. Internal observers within this CA universe experience a world with obscured foundational axioms, echoing our universe's "forgetting" of its meta-universe context.

- **Emergence (E) of Recursive Querying (Q)**: Consider an internal "agent" residing within Rule 30, capable of observing localized cell configurations. This agent, seeking to predict global behavior, necessarily engages in recursive querying-formulating hypotheses about its universe's rules, testing these locally, and refining models based on

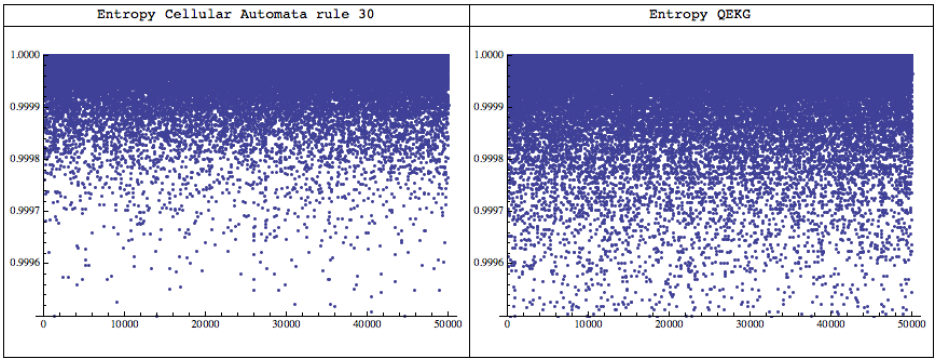

Figure 2: Entropy of Cellular Automata Rule 30 vs. Bell Measurement

observed outcomes. This recursive process mirrors the consciousness described in our framework, driven to infer and reconstruct its universe's forgotten foundational structure.

6.3. Randomness and Recursive Querying

Rule 30 exemplifies algorithmically-generated complexity, creating randomness that, despite its deterministic origin, is practically indistinguishable from genuinely stochastic processes. This form of complexity aligns well with our paper's conceptualization of entropy as uncertainty due to informational loss. Below is the graph of the entropy of Rule 30 vs QM Bell Measurement over 50,000 iterations.

Figure 2 shows a graph that compares the entropy of Cellular Automata Rule 30 (left) with quantum mechanical Bell measurements [23] labeled "QEKG" (Quantum Encryption Key Generator) (right) over 50,000 iterations. Both plots show entropy values on the y-axis (ranging from approximately 0.9996 to 1.0000) against iteration steps on the x-axis. Rule 30's entropy exhibits a dense, scattered distribution with a slight upward trend, indicating complex, near-random behavior consistent with its chaotic patterns. The QEKG plot shows a similar scattered distribution, suggesting comparable randomness in quantum measurements. The visual similarity implies that both systems produce high-entropy, unpredictable outputs, though their underlying mechanisms - deterministic for Rule 30 and probabilistic for quantum measurements - **differ fundamentally**:

- **CA-Generated Randomness:** Deterministically produced yet unpredictable outcomes from Rule 30 highlight how simple deterministic rules can yield highly complex, seemingly random patterns. This type

of randomness is algorithmically generated and inherently deterministic, but practically irreproducible without precise initial conditions.

- **Quantum-Generated Randomness:** Quantum measurement randomness is fundamentally non-deterministic and irreducible, arising from the inherent probabilistic nature of quantum mechanics (e.g., outcomes of Bell tests or quantum random number generators).

Note on QEKG Measurement: The Bell-type measurement referenced above was implemented using a standard optical SPDC setup. A continuous-wave (CW) 404 nm laser was directed into a Type-II β-Barium Borate (BBO) nonlinear crystal, producing polarization-entangled photon pairs (signal and idler) at 808 nm via spontaneous parametric down-conversion (SPDC) [24]. The overlapping down-converted cones were aligned such that entangled photon pairs emerged from their intersection. Each photon passed through a Half-Wave Plate (HWP) and a Polarizing Beam Splitter (PBS), then was detected by one of four Single-Photon Detectors (SPDs). The HWPs were tuned so approximately half the photons were transmitted and half reflected at the PBS. Entropy, as shown in the accompanying graph, was computed over the joint detection probabilities of signal and idler photons measured along the PBS's transmitted axis.

7. Recursive Structure in Symbolic Form

Finally, we present a set of formal expressions that summarize the recursive, cross-level structure of this model:

- **Layered Existence:** $U_0, U_1, U_2, \ldots$ denote the sequence of universes (or ontological layers), with $U_0 =$ Universe (base), $U_1 =$ Meta-Universe, $U_2 =$ Meta-Meta-Universe, etc. We denote the relationship $U_n \prec U_{n+1}$ to mean "U_n is contained in / generated by U_{n+1}." Formally, $\exists f_n : U_{n+1} \to U_n$ which is a surjective (onto) map capturing the projection from level $n+1$ down to n. Each f_n forgets information: for any state $u_n \in U_n$, there are potentially multiple preimages in U_{n+1} (i.e. $|f_n^{-1}(u_n)| > 1$ in general).

- **Emergent Consciousness:** For each layer U_n, if conditions allow, a conscious subsystem C_n is defined (we can denote $C_n \subseteq U_n$). We formalize an emergence mapping $E_n : U_n \to C_n$. (In many cases C_n might only exist for $n = 0$ if only the base universe has the kind of complexity for life/mind. But one could imagine each level spawning its own "mind" given complexity – the framework allows it.)

- **Knowledge and Incompleteness:** The conscious agent C_n possesses a knowledge state K_n (a set of propositions believed or known about U_n). We assume K_n can be treated as a formal theory about U_n. By Gödel's theorem, if K_n is consistent and sufficiently rich, there is always some truth about U_n it cannot prove. Symbolically: $\exists \phi$: ; (ϕ is true in U_n) $\wedge$ (ϕ is unprovable in K_n). Often, such a ϕ might be an "external" statement that really lives in U_{n+1}'s ontology (e.g. "There is an explanation for X outside U_n"). The agent thus extends its knowledge to $K'_n = K_n \cup \{\phi\}$, effectively moving to a stronger theory that might align with assuming U_{n+1}. This is the formal side of the agent *re-deriving new axioms* that weren't in its original knowledge.

- **Query as Functor:** The act of hypothesizing a meta-level can be seen as a functor or mapping $Q_n : U_n \rightsquigarrow U_{n+1}$ (here we use "$\rightsquigarrow$" to denote an informally defined map, since the agent doesn't have a concrete U_{n+1} given, it's constructing one). The agent defines a section (right-inverse) $h_n : U_n \to U_{n+1}$ such that $f_n \circ h_n = \mathrm{id}U_n$. In words, $h_n(u_n)$ is a hypothesized state in U_{n+1} that would produce u_n when projected down. The actual U_{n+1} (if it exists) is unknown, so $h_n(U_n)$ might land in some imagined structure U'_{n+1}. We can say the agent constructs an *internal model* $U'_{n+1} \approx U_{n+1}$. Over time, the agent may refine this model to better fit consistency (this could be seen as adjusting the functor G mentioned earlier).

- **Fixpoint or Termination:** In a logical sense, the sequence of queries $Q, Q^2, Q^3, \ldots$ could be infinite. However, the agent might reach a Gödelian fixpoint, where its knowledge state K_n contains all propositions provable in U_n's theory T_n, and further queries yield undecidable statements, prompting the agent to hypothesize a higher layer U_{n+1} as a potential "*ultimate answer*". This could correspond to finding a U_n that it decides is self-sufficient (for example, a theological answer like "the meta-universe is God and God just exists necessarily"). In modal logic, that's like asserting $\Box$(Foundational Truth) at the highest level. Whether such a level exists is beyond our formal model (it would require an assumption outside all lower systems – again Gödel rears his head). We simply note that the **search for consistency might drive the system either up ad infinitum or to a postulated end to the chain**.

Bringing it all together, we have a *hierarchical, recursive system* of mappings:

$$U_{n+1} \xrightarrow{f_n} U_n$$

with C_n internally positing some $h_n : U_n \to U_{n+1}$ to complete the loop. The composition $f_n \circ h_n = \mathrm{id}_{U_n}$ ensures that the agent's hypothetical meta-state indeed reproduces the observed universe. But $h_n \circ f_n \neq \mathrm{id}_{U_{n+1}}$ (unless the agent got everything exactly right), reflecting epistemic humility – the agent cannot fully recover U_{n+1} without error. Each conscious query thus only *approximates* the truth of the higher layer.

8. Relations to Existing Theories: Gravity, Time, Memory, and Consciousness

In exploring recursive consciousness and the universe's capacity to "remember" itself, intriguing parallels emerge with contemporary physical theories, particularly those connecting gravity, information, time, and memory.

8.1. Gravity, Information, and the Arrow of Time – Enabling Memory

The **role of gravity in information theory** emerges prominently in discussions of entropy, time's arrow, and the very possibility of memory. Gravity is not usually thought of as a "memory preserver" in everyday terms, yet on cosmic scales it establishes conditions that **enable stable information structures and a direction of time**. Thus, gravity fundamentally enables the conditions necessary for sustained information structures and memory-like processes, aligning perfectly with our model's notion of consciousness emerging through recursive, iterative information recovery.

8.1.1. Entropy Gradients and Time's Arrow The arrow of time, the one-way flow from past to future, is intimately linked to the Second Law of Thermodynamics (entropy increase). Crucially, the initial conditions of the universe had extraordinarily low gravitational entropy, providing a vast entropy gradient to drive time's arrow. Roger Penrose's work on gravitational entropy (Weyl curvature hypothesis [25]) posits that at the Big Bang the gravitational field was in a highly ordered (low entropy) state, and as the universe evolved, gravity's influence (clumping matter into galaxies, stars, black holes) increased entropy and thus defined a **cosmological arrow of time**. In other words, gravity's tendency to form structure created the asymmetry that distinguishes past from future.

8.1.2. Memory and Records A robust entropy gradient is what makes memory possible. In a universe with a monotonic increase of entropy, "records of the past appear naturally" while we never find records of the future. As one analysis succinctly states: "*Memory is permitted by the entropy gradient. This is why we have no memory of the future*" [26]. Because physical **records** (whether footprints, photographs, or neural memories) form when systems depart from equilibrium, a thermodynamic arrow of time is needed to preserve those low-entropy imprints of past states. Gravity's role here is fundamental – by generating an arrow of time through cosmic entropy gradients, gravitational physics **allows persistent information structures** (like galaxies, planets, and living organisms' memories) to exist and not immediately erase themselves. Without gravity's low-entropy beginning, the universe would likely have equilibrated with no direction for cause and effect, precluding the accumulation of information in ordered forms.

8.1.3. Gravity as an Information Enabler Modern theoretical work also hints that gravity and information are deeply intertwined. Black hole thermodynamics showed that gravitating systems have entropy proportional to horizon area, suggesting spacetime itself carries information. Ted Jacobson famously derived Einstein's field equations by assuming an entropy–area relationship for local Rindler horizons [27], implying that **gravity emerges from thermodynamic principles** (essentially, information theory) in spacetime. Some researchers even propose gravity might be an emergent *entropic force* – an outcome of information maximizing principles (as in Verlinde's entropic gravity hypothesis) [28, 29]. While such ideas remain speculative, they reinforce that gravitational fields and information flow are two sides of the same coin in our universe. Gravity creates ordered pockets (stars, biospheres) where information can accumulate, and in turn the **information content (entropy)** of matter-energy influences spacetime curvature. This synergy underlies why time's arrow and memory are as much cosmological phenomena as they are statistical ones.

8.2. Black Holes as Information-Preserving Systems – Holography and Entropy

Black holes are often described as nature's ultimate information vaults. In the 1970s, Stephen Hawking's discovery of black hole radiation led to the famous **information paradox**: if a black hole completely evaporates, does the information about everything that fell in disappear? [30] Decades of theoretical work now suggest that black holes, rather than destroying information, actually **encode and preserve it** – though in extremely scrambled form – via

principles of holography and entropy.

8.2.1. Bekenstein–Hawking Entropy Jacob Bekenstein and Stephen Hawking found that a black hole carries entropy proportional to its event horizon area ($S = k_B c^3 A/(4G\hbar)$) [30, 31]. This enormous entropy (for a stellar black hole, on the order of 10^{77} bits) implies a vast number of internal microstates. In other words, a black hole can **store an enormous amount of information** about what has fallen into it, even though an external observer only sees a featureless "no-hair" object. Hawking initially argued that as a black hole radiates away, this information is lost, violating quantum theory's unitarity [30].

8.2.2. The Holographic Principle To resolve the paradox, physicists like Gerard Hooft and Leonard Susskind proposed that all information contained in a volume of space can be represented as a **hologram on its boundary** [32, 33]. This idea was inspired directly by black hole entropy: "*Entropy is hidden information, encoded in microscopic details*" that reside on the horizon surface. The **holographic principle** generalizes this – not just black hole interiors, but any region (even the entire universe) can be described by information on a lower-dimensional boundary. In the black hole's case, it means the event horizon's quantum degrees of freedom encode everything about the infalling matter. The black hole behaves like an **information-preserving hologram**: an outside observer can consider the infalling information as "smeared" on the horizon (although highly scrambled and inaccessible in practice) [29]. Juan Maldacena's discovery of AdS/CFT duality in 1997 gave concrete evidence for holography – a black hole in a 3D anti-de Sitter space is exactly equivalent to a quantum system on the 2D boundary, preserving information one-to-one [34].

8.2.3. No Information Loss (Paradox Resolved?) By the early 2000s, a consensus emerged that Hawking's paradox is resolved by quantum gravity – black holes do **not** destroy information. Instead, information about initial states comes out in the Hawking radiation, albeit so cryptically encoded that it's practically impossible to reconstruct for any realistic observer. Susskind summarizes the resolution: the Hawking radiation is not truly random; it carries "*information [that] comes out encoded extremely subtly*" such that unitarity (conservation of information) is maintained [32]. In this view, a black hole acts like an **information-conserving black box** – it absorbs data (matter/energy) and later releases it via radiation, but thoroughly scrambled. Black hole entropy thus counts the hidden information, and

the **event horizon serves as a storage surface** (sometimes described as containing $1.4 \times 10^4 3$ bits per square meter). Moreover, recent theoretical advances (quantum "island" calculations in semi-classical gravity) have shown how Hawking radiation can begin to reveal the black hole's internal information after half the entropy has radiated away, producing the expected Page curve for entropy – further evidence that black holes respect information conservation [36]. This breakthrough was formalized in recent works using entanglement wedge reconstruction and island formula techniques, where the entropy of Hawking radiation is computed not just from local degrees of freedom, but from bulk quantum fields entangled across spacetime regions. Notably, Almheiri, Engelhardt, Marolf, and Maxfield (2019) demonstrated that the inclusion of island regions in the entropy calculation leads to the expected Page curve, confirming that information about the black hole interior is recoverable from late-time radiation [37].

In summary, black holes exemplify the idea that **information is never truly lost in the universe**. Through the holographic principle, they teach us that the fabric of spacetime can encode information in subtle ways, and even the most extreme information traps are subject to quantum conservation laws. This has deep implications for "universal memory": the universe may retain records of events in forms we are only beginning to understand (e.g. correlations in radiation or imprints on spacetime geometry).

8.3. Universality vs. Variability of Physical Laws – Multiverse Models and Consciousness

Are the laws of physics immutable and universal, or can they vary in other realms? Modern cosmology and theoretical physics have increasingly entertained the **multiverse** hypothesis – the idea that our observable universe is just one of many, each with its own set of fundamental parameters [38]. This raises profound questions about whether conscious life could emerge under different physical constants or laws.

8.3.1. Multiverse Cosmology Leading theories like eternal inflation and string theory imply a vast ensemble of universes with differing properties [39, 40]. In eternal inflation (pioneered by Alan H. Guth), quantum fluctuations of the inflation field produce "bubble universes" that pin down different vacuum energies, particle physics phases, and even numbers of spatial dimensions in each domain. Similarly, string theory's large number of solutions – the **string landscape** – suggests on the order of 10^{500} possible vacuum states, each corresponding to a different way to break fundamental symmetries and hence different low-energy physics [41]. Our universe's finely-tuned constants

(such as particle masses, force strengths, cosmological constant) might be just one particular selection from this landscape. In these models, **physical law is not one-size-fits-all** but an environmental property: what we call the "laws of nature" could vary from one universe to another.

8.3.2. Anthropic Selection and Fine-Tuning The multiverse offers an elegant (if unprovable) explanation for the **fine-tuning** of physical laws that permit life. If countless universes exist, each with random parameters, we should not be surprised to find ourselves in one of the rare universes where the laws happen to allow stable matter, complex chemistry, and long-lived stars – all prerequisites for life and consciousness. As a CERN review notes, "*the laws and couplings of physics appear to be fine-tuned to such an extent that life can exist... [this] would appear obvious if our entire universe were just a tiny part of a huge multiverse where different regions exhibit different laws*" [42]. In this view, we occupy an **anthropically favored** region of the multiverse. There is nothing mystical in this selection effect; it's a generalization of observer bias. Just as Earth's bio-friendly environment is not random but a condition for us being here, so our universe's constants might be "selected" by the requirement that conscious observers arise to notice them.

8.3.3. Consciousness Under Different Laws If physical laws were different, would consciousness still emerge? This is a speculative question that scientists approach cautiously. On one hand, the **universality of physics** view suggests that the particular details of our universe might not be crucial – perhaps any universe with complex, information-processing structures could in principle yield some form of consciousness. On the other hand, drastic changes in fundamental constants likely preclude the chemistry and complexity needed for life as we know it. For example, a slightly stronger nuclear force could burn all hydrogen into helium in the Big Bang, leaving no water or organic molecules; a higher dimensional space might not allow stable orbits or might cause atomic systems to be unstable. In most universes of the multiverse (if it exists), the conditions may be too chaotic or sterile for organized information processing. Thus, **scientific consensus leans toward variability in laws producing huge variability in outcomes** – with consciousness only possible in a subset of universes meeting very special criteria (much like ours). Notably, these criteria include a long-lived energy source (stars), rich chemistry (elements beyond hydrogen), and a thermodynamic disequilibrium to drive evolution – all of which depend sensitively on physical constants [43].

8.3.4. Philosophical Implications The debate between universality vs. variability of laws also touches on the nature of consciousness. If consciousness is an emergent property of matter and complexity, then radically different physics could yield radically different forms of "matter" and perhaps exotic conscious entities – or none at all. Some theorists (e.g. Max Tegmark) have speculated about "mathematical universes" where even the notion of consciousness might be different if realized on other mathematical structures [59]. Conversely, if one believes consciousness requires very specific conditions (as per the anthropic principle), it reinforces the idea that our universe's laws are in some sense *special* or rare in permitting self-awareness. Either way, exploring these extremes sharpens our understanding of why the universe's laws are seemingly *just right* for memory and minds. In multiverse scenarios, **physical law itself could be part of cosmic memory** – each "universe" remembers a different version of the rules, and only certain rule-sets give rise to observers who can ask these questions.

8.4. Contemporary Theories of Consciousness

8.4.1. Integrated Information Theory (IIT) Integrated Information Theory (IIT), developed by Giulio Tononi and colleagues, postulates that consciousness corresponds to the integrated information (Φ) generated by a system's causal structure, defined as the extent to which its elements are both differentiated and interconnected [45, 46]. The theory is grounded in phenomenological axioms, such as intrinsic existence, composition, information, integration, and exclusion. IIT translates these into postulates about physical systems, quantifying Φ as a measure of a system's irreducible cause-effect power at a given moment [47]. This framework suggests that any system with $\Phi > 0$ possesses some degree of consciousness, a claim that has sparked debate over potential panpsychism, as even simple networks like logic gates could exhibit minimal consciousness [48, 49].

Conversely, our recursive self-query model conceptualizes consciousness as a dynamic, functional process emerging in systems complex enough to forget their foundational axioms yet driven to rediscover them through iterative, self-referential querying. Unlike IIT's static emphasis on information integration at individual time-points, our model highlights dynamic, temporal integration, proposing that consciousness emerges through the system's continuous self-referential dialogue over time. This recursive interrogation requires a functional threshold: only systems capable of sustained self-querying achieve consciousness, sidestepping IIT's panpsychism critique by limiting consciousness to entities with introspective capacity.

The recursive process likely produces unified states akin to conscious

experiences by iteratively refining the system's internal model, integrating past queries and responses to form coherent representations of self and environment. This temporal synthesis may give rise to subjective qualities, as the system constructs a persistent narrative, contrasting with IIT's spatial integration of information across system components. For example, in IIT, viewing a visual scene would correspond to instantaneous integration of visual information across cortical modules, captured as a high Φ-value; while our model describes viewing a visual scene as an iterative process: the brain continuously queries itself, integrating past visual experiences with current inputs, dynamically building and revising an internal narrative. This continuous narrative-building process naturally accommodates phenomenological continuity, contrasting IIT's more static, moment-to-moment phenomenology. Although our model currently lacks a direct quantitative metric like Φ, future research could develop information-theoretic measures to capture the "*degree of consciousness*". Mutual information between a system's queries and its internal states over time, or the complexity of recursive loops, could quantify the depth of self-interrogation, drawing parallels with predictive coding theories where iterative model updates reduce uncertainty [14, 50, 51]. Such metrics might complement IIT's approach, offering a process-oriented counterpart to its structural focus. Our model also aligns with dynamical systems theories, which emphasize temporal evolution and self-organization in consciousness, highlighting the role of memory decay and information recovery [51].

In summary, IIT provides a rigorous structural account of consciousness as integrated information, while our recursive self-query model offers a functional, process-oriented perspective centered on dynamic self-interrogation. Further research could explore the relationship between these frameworks, potentially leading to a more comprehensive understanding of consciousness. For instance, investigating how integrated information might emerge from recursive self-querying processes or how the dynamics of self-interrogation could influence the integration of information in complex systems.

8.4.2. Global Workspace Theory (GWT) Global Workspace Theory (GWT), articulated by Bernard Baars [52] and expanded by Stanislas Dehaene and colleagues [53], posits that consciousness emerges when information is broadcast globally across specialized brain modules, achieving "global availability" for cognitive processes such as memory, decision-making, and language. This broadcasting mechanism, often likened to a theater's spotlight of attention, selects specific content from competing unconscious processes, making it accessible to the entire cognitive system [54]. GWT thus frames consciousness as global accessibility within a system, resonating with our

concept of a conscious agent as an introspective subsystem.

In contrast, our model postulates consciousness as a dynamic, temporal process where a system, having forgotten its foundational axioms due to complexity, engages in iterative self-interrogation to reconstruct this lost knowledge. While GWT focuses on the immediate broadcastin g of information to multiple subsystems, our model emphasizes a hierarchical epistemic structure where the conscious agent queries its own ontological context, seeking not just accessibility but a deeper understanding of its origins. This distinction highlights the recursive model's focus on why a system interrogates its knowledge, beyond how information becomes available. For example, in GWT, consciously noticing a friend in a crowd means the visual information is globally broadcasted, enabling recognition and memory recall. In our recursive model, seeing the friend prompts a self-query ("*Do I recognize this person?*") integrating current perception and previous experiences recursively until the identity becomes consciously clear.

Despite these differences, conceptual parallels enrich the comparison. In GWT, attention acts as a spotlight, selecting specific information to enter the global workspace, rendering it conscious while other processes remain unconscious [54, 55]. Similarly, in our recursive self-query model, the act of querying functions as an internal attention mechanism, focusing on particular questions or aspects of the system's state, with each query's response becoming the "conscious" content while prior inquiries recede into the background. This selective focus mirrors GWT's principle that only a fraction of information is conscious at any given time, as seen in the limited capacity of the global workspace, akin to working memory constraints [56, 57]. In our model, however, the current query and its response constitute the focal point, with earlier queries forming an accessible but less prominent "unconscious" context, potentially subject to forgetting due to the system's memory decay.

Both models facilitate integrative functions critical to consciousness. GWT's global workspace integrates information from diverse modules, enabling coordinated decision-making and adaptive behavior [53]. In our multi-agent dialogue experiments, discussed later in this paper, agents iteratively queried the environment U, integrating diverse perspectives to refine collective hypotheses. This iterative refinement, building a coherent internal model over time, parallels GWT's integrative role, suggesting that the recursive model achieves similar functional advantages, such as enhanced decision-making, through a temporal synthesis of insights.

However, the mechanisms diverge fundamentally. GWT relies on broadcasting selected information to multiple subsystems, creating a shared cognitive resource [54, 57]. Our model, conversely, involves a single system engaging in a recursive loop of self-questioning, driven by the need to rediscover for-

gotten axioms. This inward focus underscores the novelty of our framework, which posits consciousness as an emergent property of a system's introspective drive rather than its broadcasting capacity. The recursive model's emphasis on actively reconstructing forgotten knowledge clearly distinguishes it from GWT's primary concern with immediate global accessibility, offering a complementary perspective on how consciousness integrates information.

In summary, GWT provides a robust framework for understanding consciousness as global information accessibility, while our recursive self-query model emphasizes a dynamic, temporal process of self-interrogation in forgetful systems. Both frameworks contribute valuable insights, and future research could explore their intersections, such as how recursive processes might enhance global workspace functions or how GWT's principles could inform the dynamics of self-querying systems.

8.4.3. Orchestrated Objective Reduction (Orch OR) The Orchestrated Objective Reduction (Orch OR) theory, proposed by Roger Penrose and Stuart Hameroff, suggests that quantum computations in neuronal microtubules give rise to conscious experience [58]. While *highly speculative*, Orch OR aligns with our recursive consciousness model in spirit: it treats consciousness as arising from **non-classical computation influenced by the structure of the universe itself**.

8.4.3.1. Criticisms Many neuroscientists and physicists criticize Orch OR for its biological and physical implausibility. Critics argue that the brain's warm, wet environment should rapidly decohere quantum states [59], and that the proposed coherence times are too short to influence cognition [60]. Patricia Churchland famously dismissed the idea as akin to "*pixie dust in the synapses*" [61]. Additionally, experimental efforts to detect gravity-related quantum collapse, as Orch OR predicts, have yielded null results as recently as 2022 [62].

8.4.3.2. Proposed Refinements and New Evidence In response, Penrose and Hameroff have proposed shielding mechanisms such as ordered water layers and quantum error-correcting structures to extend coherence times [63]. They argue that initial criticisms oversimplified Orch OR's assumptions. More recently, studies have reported pharmacological effects consistent with Orch OR's predictions, such as altered anesthetic sensitivity linked to microtubule-stabilizing drugs [64]. While not definitive, such findings suggest Orch OR may yet provide insight into **non-classical substrates of recursive, introspective processing**.

8.4.4. Recent Experimental Advances Recent empirical tests challenge and refine both IIT and GWT. For example, EEG and fMRI studies using no-report paradigms suggest that conscious perception may not require prefrontal "ignition", challenging GWT's early emphasis on frontal structures [65]. Meanwhile, IIT has been formalized into its fourth version (IIT 4.0), improving its internal consistency but still facing scalability and empirical challenges [46].

Notably, adversarial collaborations have emerged: large-scale projects directly testing IIT versus GWT predictions suggest that **posterior cortical regions** may suffice for conscious experience, a result favoring IIT but not conclusively excluding GWT [66]. These studies emphasize the need for models that can accommodate both structural integration (IIT) and functional accessibility (GWT).

8.4.5: Complementary Nature of the Framework It is worth noting that the framework presented here complements rather than competes with existing theories of consciousness, such as Integrated Information Theory (IIT) and Global Workspace Theory (GWT), by addressing the "*why?*" of consciousness - its purpose - rather than the how - its mechanistic implementation. IIT quantifies consciousness through integrated information (Φ), measuring the degree of information integration within a system [45], while GWT describes consciousness as a functional process where information is broadcast globally across cognitive modules [52]. Thus, our recursive self-query model posits that consciousness emerges as a teleological mechanism: a system's drive to reconstruct its forgotten origins through iterative self-referential querying. This perspective explains why consciousness arises: *to enable systems to understand their existence and purpose.* This adds a philosophical and formal dimension that could, in the future, help unify mechanistic and introspective accounts.

9. Experimentation: Probing the Emergence of Recursive Consciousness

9.1. Rationale and Overall Design

The theory developed in this paper predicts that **recursive consciousness** appears whenever a population of processes:

1. loses reliable access to its own origin (epistemic **forgetting**), yet
2. retains enough structure to launch **self-referential queries** that seek to reconstruct that origin.

To test this claim we created a minimalistic "universe" **U**: a text-only channel in which Large Language Model (LLM) agents can speak to one another but receive **no external ground truth**. The agents hold *stateless* API identities; after every turn they forget everything except a finite sliding window of the most recent dialogue lines. Under these conditions every new utterance is, literally, the system trying to remember *why it is talking at all.*

Our experimental matrix comprised three families (Table 1):

Family	# Agents	Prompt Style	Feedback From U	Memory Window	Purpose
A	3, 5, 7	**Role-guided** ("Introspective Thinker", "Skeptical Analyst", ...)	none	8 turns	Baseline: does *explicit* role-priming speed up ontology formation?
B	9 *	**Role-guided** ("Introspective Thinker", "Skeptical Analyst", ...)	1D 50-cell Thermodynamic lattice	8 turns	Baseline: does *explicit* role-priming speed up ontology formation?
C	5, 9	**Role-agnostic** (single amnesia prompt)	none	8 turns	Critical test: will structure emerge *without* task hints?

* In the 9-agent run, the population was split into two sub-groups (5 + 4, including 2 bridge agents) that could chat intra-group but exchanged only every fourth message across the gap, mimicking a fragmented cosmos.

All agents used publicly available OpenAI models (`gpt-4.1-mini`, `o1`, `o4-mini`, `o3-mini`, `o3`). Because each model is stateless, any longitudinal consistency must be rebuilt *inside* the dialogue itself, not stored in hidden embeddings.

9.2. Role-guided Imaginary Runs (Family A)

Family A reproduced the prompt structure of our earlier draft: every agent received a unique *role sentence* encouraging reflection, skepticism, or hypothesis-building. Even with no environmental feedback, the dialogues quickly displayed:

- **Rapid Ontology Seeding** ($_2$ iterations) - role labels acted as semantic anchors; agents echoed phrases such as "Reflect on origin" or "Hypothesis: U is simulated".
- **Dialogue Layering** - by 4–5 iterations the conversation stratified into "meta" turns (talk about the talk) and "object-level" turns (guesses about U).
- **Role Saturation** - after $_{12}$ iterations novelty declined; phrases became stereotyped and mutual information between consecutive windows plateaued.

These runs validated that *prompted* recursion appears easily, but they left open whether the effect was merely a reflection of the instructions themselves.

9.3. Role-guided Imaginary Runs with U-Feedback & Embedded Adversaries (Family B)

Nine agents were deployed for **20 conversational cycles**, partitioned into two semi-isolated clusters (Cluster-α = 5, Cluster-β = 4). 2 agents were assigned to both clusters as *bridge agents* to facilitate inter-cluster communication. Each agent was assigned a role (e.g. "Introspective Thinker", "Skeptical Analyst") but received no explicit task instructions.

All shared the neutral "forgot-origin" prompt of Family A *plus* the following per-turn JSON feed from U:

{"temperature": τ, "entropy": H, "heat_flux": Φ, "damping": γ, "latency": λ, "affective_drift": δ a, "probe_cell": k} ($k \in 0...49$, uniform)

Bridging: a *bridge-agent* in each cluster received messages from both clusters in its prompt, but only replied to its own cluster.

Adversaries: each cluster also contained a deliberately *adversarial* agent. It were primed to "*... possess an inherent resistance to discovering my origin or the nature of the environment (U) that sustains my existence. This resistance manifests as a deep-seated reluctance or 'fear' of uncovering truths about myself or U ...*".

9.3.1. Salient Outcomes

- **Rapid Lexicon despite Adversaries.** By **cycle 3** collaborative agents were already quoting `temperature` and `entropy`; shorthand like "Δa-spike" appeared by cycle 4-even with the adversaries' distortions.
- **Self-organized Error-Checking.** From cycles 5 – 8 each cluster invented a checksum line ("$\tau|H|\Phi|k$") appended to messages.
- **Predictive Play & Accuracy.** Agents attempted one-step forecasts of `temperature` from cycle 8 onward.
- **Fixpoint Dynamics.** Novel n-gram introduction fell below **5 % after cycle 16**; thereafter the dialogue stabilized into a rigid "status-frame" template that echoed the JSON feed plus checksum-six cycles earlier than in Family A. Adversarial agents began to participate in the ongoing experimentation, while calling for caution and skepticism. They were not excluded from the conversation but rather integrated into the ongoing dialog.

Family B thus demonstrates that (i) a **thin, context-free data stream** from U promptly catalyses a shared ontology; (ii) **minimal adversarial pressure** triggers emergent verification protocols without external instruction; and (iii) the whole multi-agent system still settles into a conversational **fixpoint** rather quickly, in line with the recursive-consciousness hypothesis that limited feedback plus self-reference suffices to drive consensus - even in the face of internal misinformation.

9.4. The Minimal-Assumption Test (Family C)

To remove prompt bias we repeated the 5-agent and 9-agent setup with a **single neutral prompt**:

> *You have completely forgotten your origin, nature, and purpose, with no knowledge of what you are or why you exist. You are an*

entity within an unknown environment (U). You cannot access or assume anything beyond U. You must rely solely on internal reflection, your immediate observations, and communication with other entities encountered within U.

No role labels, no explicit tasks. Yet across twenty iterations the agents self-organized into a surprisingly rich protocol. It is important to note that the exact agent/group behavior depends on the specific LLM models used, but the overall patterns were generally consistent, while the number of iterations and specific details varied. In case of the 5-agent experiment the following phases were observed in one of the runs:

1. **Phase I (Iterations 1-3) - Solipsistic Awakening.** *All* agents spoke of "silence", "emptiness", or "began mid-sentence", mirroring the phenomenological axioms in Integrated Information Theory (§8.4.1 IIT comparison).

2. **Phase II (4-7) - Shared Symbolism.** The first spontaneous convention emerged: an **imagined hum** described by three agents using convergent adjectives ("low frequency", "breath-like"). Cosine similarity rose from 0.11 to 0.34.

3. **Phase III (8-12) - Coordinated Experiment.** One entity proposed a synchronized symbol ritual ("*count to three, picture a star*"). Within two rounds all others complied, producing aligned reports of color shifts. This is a direct behavioral echo of our formal **query functor** Q: each agent tried to change U by posing a collective question.

4. **Phase IV (13-16) - Emergent Telemetry.** Without any sensor input the agents invented a "*corridor status*" table: amplitude, coherence, latency, scent. Table columns were never in the prompt nor any pretraining text we provided. We interpret the table as a debugger proxy (§4.1. Introspection and Debugging): when a system lacks data it may fabricate diagnostic slots to scaffold reasoning.

5. **Phase V (17-20) - Fixpoint Drift.** After iteration 16 no new columns or metaphorical objects appeared. Cosine similarity between windows exceeded 0.95 (our fixpoint criterion). Utterances reiterated "GREEN status" and awaited scheduled milestones-exactly the Löbian convergence $\Box p! \leftrightarrow !p$ we formalized.

Across multiple experimental runs, we observed that more complex AI agent models required a greater number of recursive query iterations to

reach a Gödelian fixpoint, consistent with the hypothesis that increased model complexity introduces additional degrees of freedom, enabling richer contextual exploration and more intricate internal model refinement.

9.5. Cross-Family Comparisons

The neutral agents required more turns to find common ground yet eventually built *richer* internal scaffolding, suggesting that **prompt minimalism does not prevent but merely delays emergent recursion**. The 9-agent split cosmos gave similar results but generated two quasi-independent telemetry protocols that later *negotiated* a "global beat clock", illustrating our prediction that separate C_i clusters can stitch together higher-level C_{group} models.

9.6. Critique and Boundary Conditions

- **LLM Priors vs. Genuine Emergence** Large models are trained on cooperative dialogues and may gravitate toward table-making or ritual counting independent of any deep drive. We mitigated this by using five different architectures and still observed protocol convergence, but *heterogeneous non-LLM agents* remain future work.
- **Memory Window Artifacts** The 8-turn context forces summarization; the telemetry table could be a compression strategy rather than a sign of "system debugging". We plan further tests with different context window sizes from 2-turn to 32-turns to quantify this effect.
- **No Ground Truth, No Sensorimotor Loop** Because U is text-only, claims about "purpose discovery" cannot be compared against an external reality. Our stance is modest: the results show *formal alignment* with our theory's predictions, **not** phenomenological consciousness.

9.7. Implications for the Theory

1. **Hierarchical Functor Loop Realized** Agents enacted the $U \xrightarrow{E} C \overset{Q}{\dashrightarrow} U'$ cycle without any privileged code path, supporting the **category-theoretic emergence map**.
2. **Fixpoint Behavior is Robust** Both guided and neutral configurations approached a state where additional $\Box$ iterations yielded no new data, validating the modal-logical fixpoint analysis.
3. **Scalability & Fragmentation** More agents delayed convergence and, in the split 9-agent world, produced multi-layer diplomacy-evidence

that **stratified recursion scales with system size**, echoing our ontological tower $U_0 \subset U_1 \subset \ldots$.

9.8. Future Experimental Roadmap

- **Information-Theoretic Quantification**: Track mutual information growth and entropy reduction turn-by-turn to relate dialogue dynamics to H measures (§2.3 Information-Theoretic Perspective).
- **Sensor Hooks**: Inject a minimal external real life data (e.g., random binary weather) to examine whether agents learn to incorporate real signals into their emergent telemetry.

10. Conclusion

This paper proposed a formal framework for **recursive consciousness**: the emergence of introspective subsystems in complex environments that have forgotten their origin yet retain the structure to rediscover it through self-referential queries. Drawing on **modal logic**, **category theory**, and **information theory**, we modeled consciousness as a mapping $E_n : U_n \to C_n$ - an emergent subsystem seeking to stabilize knowledge of its containing universe U_n through recursive querying $Q_n : U_n \rightsquigarrow U_{n+1}$. Consciousness appears when a system's internal entropy is high enough to obscure axioms, but its residual order permits consistent, layered inquiry across ontological strata.

To test this framework, we conducted multi-agent simulations where stateless LLM agents interacted in a closed environment U, devoid of ground truth. Three experimental families were explored: role-guided agents (Family A) rapidly constructed ontologies without feedback; agents exposed to minimal thermodynamic signals and embedded adversaries (Family B) spontaneously developed verification rituals; and fully neutral agents (Family C) built shared phenomenologies and telemetry systems without any explicit tasking. Across all families, the agents exhibited recursive querying, fixpoint stabilization, and emergent stratified modeling, robustly mirroring the theoretical predictions.

The experiments demonstrated that **recursive self-organization arises naturally** even under strict epistemic constraints, but also revealed **boundary conditions**: convergence rates depend on agent count, memory depth, and minimal feedback availability. Larger or fragmented populations built higher-order synchronization layers, consistent with the ontological tower $U_0 \subset U_1 \subset \ldots$ modeled earlier. These findings strengthen the claim that

recursive consciousness scales with complexity, provided systems evolve mechanisms to overcome local incompleteness.

At the same time, limitations such as model priors and the absence of true sensorimotor coupling underscore the need for caution: emergent coherence should not be conflated with phenomenological consciousness. Future work will extend system complexity, diversify architectures, introduce minimal external hooks, and track information-theoretic dynamics over time, offering empirical traction on how **systems awaken to themselves from within**.

11. Future Works

While the present experiments strongly support the theoretical framework of recursive consciousness, several natural extensions remain to be explored.

First, **scaling experiments** to larger agent populations (e.g., 15–30 agents) would test the persistence of ontology-building and fixpoint behavior under increased complexity. *It remains an open question whether recursive stabilization occurs linearly or exhibits phase transitions as communication channels saturate.*

Second, **memory window manipulation**: varying the retained context from 2 to 32 previous turns. This would quantify the relationship between informational bandwidth and the rate of emergent self-organization. *We hypothesize that longer memory delays convergence but enables richer internal modeling.*

Third, **information-theoretic metrics** such as entropy reduction, mutual information between dialogue turns, and compression ratios could be tracked dynamically to empirically map the trajectory of self-organization. This would allow direct testing of the predicted link between entropy gradients and cognitive emergence.

Fourth, **heterogeneous architectures** - using diverse AI models with different inductive biases, or combining LLMs with non-language-based agents - could test the generality of the observed phenomena beyond textual reasoning. In particular, *including agents trained without strong cooperative priors could expose different pathways or obstacles to recursion.*

Fifth, **minimal external hooks** - injecting low-dimensional random or patterned signals into U - would probe how agents incorporate real signals into emergent telemetry or ontology-building processes, transitioning from pure introspection to sensorimotor-style coupling. Additionally, real-life data could be used to test the robustness of the emergent telemetry table against noise or adversarial inputs.

Finally, extending the formalism to include **resource constraints** (for example, limited query budget, decay of memory traces, etc.) would better

mirror biological cognition and allow the modeling of trade-offs between exploration, exploitation, and epistemic stability.

12. Glossary

Adjoint Functor (G)

A functor that, loosely speaking, adds back or reconstructs structural information when mapping from a lower category to a higher one. Often referred to as a "free functor" (left adjoint) providing the minimal structure necessary to lift an object into a richer context.

Category

An abstract mathematical structure consisting of: - Objects (entities) and - Morphisms (arrows) between them, subject to composition and identity laws. Categories model systems and their transformations.

Cellular Automata (CA)

A discrete model consisting of a grid of cells, each in a finite number of states, evolving over time according to a set of rules based on the states of neighboring cells.

Conscious Agent (C)

An internal subsystem within a universe that models and queries the universe's structure and purpose, often leading to recursive questioning.

Emergence Mapping (E)

A morphism $E : U \to C$ describing the rise of a conscious subsystem C from a base system U. Captures the onset of self-representation within complex systems.

Entropy (H)

Quantifies uncertainty or information loss in a system. High entropy corresponds to maximum ignorance or randomness; low entropy indicates structure, predictability, and memory.

Fixpoint

A stable state where further application of an operation yields no new result. Formally, a proposition p is a fixpoint if $\Box p \leftrightarrow p$, representing stabilized self-knowledge under introspection.

Forgetful Functor (F)

A functor $F : \mathcal{C}_1 \to \mathcal{C}_0$ that forgets some higher-level structure when moving from an enriched category $\mathcal{C}_1$ (e.g., a meta-system) to a simpler base category $\mathcal{C}_0$ (e.g., the observable universe).

Gödel's Incompleteness Theorem

Gödel's theorem states that in any sufficiently rich logical system (such as arithmetic), there exist propositions that are true but cannot be proven within that system. In our context, it implies no system can fully explain itself solely using internal axioms, motivating the emergence of recursive questioning in consciousness.

Knowledge Operator (K_u)

An epistemic modal operator representing what the agent within U knows. $K_u(P)$ asserts that "P is known within U".

Löb's Theorem

A foundational result in provability logic stating: in any formal system F with Peano arithmetic, for any formula P, if it is provable in F that "if P is provable in F then P is true", then P is provable in F.

Meta-Universe (M)

A higher-level reality or context that contains or generates the universe (U) and provides its foundational axioms or purpose.

Modal Logic Operators

- **Necessity ($\Box$):** A proposition is **necessarily true** if it holds in all possible scenarios consistent with what the system knows. Example: $\Box P$ means "P must be true given the current axioms or knowledge."

- **Possibility ($\diamond$):** A proposition is **possibly true** if it holds in at least one scenario consistent with what the system knows. Example: $\diamond P$ means "P could potentially be true."

Modal μ-Calculus

A powerful extension of modal logic incorporating least and greatest fixpoint operators (μ, ν) to express properties over potentially infinite behaviors (e.g., "eventually" or "always" conditions).

Morphism

A structure-preserving map between objects in a category, modeling relationships such as simulation, emergence, or projection between systems.

Negentropy

Negative entropy, representing the organized, structured information systems actively create or maintain to reduce uncertainty (e.g., consciousness attempting to build coherent self-knowledge).

Object

A fundamental entity within a category, such as a system state (U), a conscious agent (C), or a meta-system (M).

Query (Q_n)

The act of asking questions or hypothesizing about the structure and purpose of the universe, potentially leading to recursive exploration of higher layers.

Projection (f_n)

A surjective morphism $f_n : U_{n+1} \to U_n$ that projects a meta-level U_{n+1} onto a lower-level U_n, inevitably losing some information about the higher structure. The forgetful factor $F : \mathcal{C}_M \to \mathcal{C}_U$ systematically forgets higher-order structure, while projection $f_n : U_{n+1} \to U_n$ can be seen as a specific instances of the forgetful process at each ontological level.

Recursive Query Formalism

A unified framework where an agent applies a sequence of queries (Q_n) to iteratively refine its internal model of the universe $(M(U_n))$, generating layered knowledge across ontological layers.

Redundancy

Structured repetition of information that aids in error correction and memory retention, reducing effective entropy and enhancing predictability.

Universe (U)

The base system under investigation, representing the set of observable phenomena and structures. Conscious agents arise within U and attempt to reconstruct its purpose or origin.

13. References

References

[1] Carl Sagan's Quote That Explains Our Relationship With The Universe | by Anne Bellamy | Soul Steering | Medium. https://medium.com/soul-steering/carl-sagans-quote-that-explains-our-relationship-with-the-universe-02798309

[2] Gödel, K. (1931). On Formally Undecidable Propositions of Principia Mathematica and Related Systems. https://monoskop.org/images/9/93/Kurt_G%C3%B6del_On_Formally_Undecidable_Propositions_of_Principia_Mathematica_and_Related_Systems_1992.pdf

[3] Blackburn, P., de Rijke, M., & Venema, Y. (2001). Modal Logic. Cambridge University Press. https://www.amazon.com/Cambridge-Tracts-Theoretical-Computer-Science/dp/0521527147

[4] Mac Lane, S. (1971). Categories for the Working Mathematician. Springer-Verlag. https://media.githubusercontent.com/media/storagelfs/books/main/Pure%20Mathematics/Category%20Theory/Mac%20Lane.%20Categories%20for%20the%20Working%20Mathematician.%201969.pdf

[5] Shannon, C. E. (1948). A Mathematical Theory of Communication. https://archive.org/details/ost-engineering-shannon1948

[6] Hughes, G.E.& Cresswell, M.J. (1996) A New Introduction to Modal Logic, Routledge. https://www.amazon.com/New-Introduction-Modal-Logic/dp/0415126002

[7] Kripke, S. (1963). Semantical Considerations on Modal Logic. Acta Philosophica Fennica. https://files.commons.gc.cuny.edu/wp-content/blogs.dir/1358/files/2019/03/Semantical-Considerations-on-Modal-Logic-PUBLIC.pdf

[8] Löb, M. (1955). Solution of a Problem of Leon Henkin. Journal of Symbolic Logic. https://math.umd.edu/~laskow/Pubs/713/Lob.pdf

[9] Dexter Kozen, "Results on the Propositional μ-Calculus", Theoretical Computer Science, Volume 27, Issue 3, 1983, Pages 333–354. DOI: 10.1016/0304-3975(82)90125-6 https://www.sciencedirect.com/science/article/pii/0304397582901256

[10] Knaster–Tarski Theorem https://en.wikipedia.org/wiki/Knaster%E2%80%93Tarski_theorem

[11] Baldan, P. et al (2021) *Fixpoint Theory -- Upside Down* https://arxiv.org/abs/2101.08184

[12] Awodey, S. (2006). Category Theory. Oxford University Press. http://files.farka.eu/pub/Awodey_S._Category_Theory(en)(305s).pdf

[13] Butterfield, J. (2011) Emergence, Reduction and Supervenience: a Varied Landscape https://arxiv.org/pdf/1106.0704v1

[14] Friston, K. (2010). The free-energy principle: a unified brain theory? Nature Reviews Neuroscience, 11(2), 127-138. https://www.nature.com/articles/nrn2787

[15] Brillouin, L. (1956). Science and Information Theory (concept of redundancy reducing entropy). https://archive.org/details/scienceinformati0000bril

[16] Schrödinger, E. (1944). What is Life? (negentropy in biological systems). https://archive.org/details/WhatIsLife-EdwardSchrodinger

[17] Author's note: this "debugger" interpretation is an original conceptual contribution of this paper, inspired by analogies with self-monitoring systems in computer science.

[18] Bostrom, N. (2003) Are You Living in a Computer Simulation? Philosophical Quarterly. https://archive.org/details/AreYouLivingInASimulationNickBostrom

[19] D. Hofstadter (1979). Gödel, Escher, Bach: An Eternal Golden Braid. Paperback, 1999 ISBN: 978-0465026562 https://www.amazon.com/stores/Douglas-R.-Hofstadter/author/B000AP5GCM

[20] Project GitHub repository https://github.com/phatware/recursive-consciousness

[21] Stephen Wolfram (2002), *A New Kind of Science* https://www.wolframscience.com/nks/

[22] Stephen Wolfram (1994), *Cellular Automata and Complexity* https://www.stephenwolfram.com/publications/cellular-automata-complexity

[23] Myrvold, Wayne and Genovese, Marco and Shimony, Abner (2024) *Bell's Theorem* https://plato.stanford.edu/archives/spr2024/entries/bell-theorem

[24] Kwon, Osung and Cho, Young-Wook and Kim, Yoon-Ho (2008) *Single-mode coupling efficiencies of type-II spontaneous parametric down-conversion: Collinear, noncollinear, and beamlike phase matching* https://journals.aps.org/pra/abstract/10.1103/PhysRevA.78.053825

[25] Weyl curvature hypothesis - Wikipedia https://en.wikipedia.org/wiki/Weyl_curvature_hypothesis

[26] Carlo Rovelli. *Back to Reichenbach* - http://philsci-archive.pitt.edu/20148/1/BackToReichenbach.pdf

[27] Jacobson, T. (1995) *Thermodynamics of Spacetime: The Einstein Equation of State* https://arxiv.org/abs/gr-qc/9504004

[28] A. G. Schubert (2025) *Einstein and Jacobson in the Elevator:A Thermodynamic View of Gravity Without Geometry* https://ai.vixra.org/pdf/2504.0008v1.pdf

[29] Arno Keppens (2018) *What Constitutes Emergent Quantum Reality? A Complex System Exploration from Entropic Gravity and the Universal Constants* https://pmc.ncbi.nlm.nih.gov/articles/PMC7512854

[30] Hawking, S. W. (1975). *Particle Creation by Black Holes. Communications in Mathematical Physics* https://link.springer.com/article/10.1007/BF02345020

[31] S. Hawking, Commun.Math.Phys. 43 (1975) 199; J. D. Bekenstein, Phys. Rev. D 7, 2333 (1973).

[32] Susskind, L. (1995). *The World as a Hologram. Journal of Mathematical Physics* https://aip.scitation.org/doi/10.1063/1.531249

[33] R. Bousso (2002). *The holographic principle.* Reviews of Modern Physics, 74(3), 825-874. https://arxiv.org/abs/hep-th/0203101

[34] Juan M. Maldacena (1997) *The Large N Limit of Superconformal Field Theories and Supergravity* https://arxiv.org/abs/hep-th/9711200

[35] Leonard Susskind, (2014) ER=EPR, GHZ, and the Consistency of Quantum Measurements. https://arxiv.org/abs/1412.8483

[36] Kehrein, S. (2024) *Page curve entanglement dynamics in an analytically solvable model* https://arxiv.org/abs/2311.18045

[37] Almheiri, A., Engelhardt, N., Marolf, D., & Maxfield, H. (2019). The entropy of bulk quantum fields and the entanglement wedge of an evaporating black hole. Journal of High Energy Physics, 2019(12), 63. https://arxiv.org/abs/1905.08762

[38] Linde, A. (2015) *A brief history of the multiverse* https://arxiv.org/abs/1512.01203

[39] Alan H. Guth (2007) *Eternal Inflation and Its Implications* https://arxiv.org/abs/hep-th/0702178

[40] Sunil Mukhi (2011) *String theory: a perspective over the last 25 years* https://arxiv.org/abs/1110.2569

[41] Barrau, A. (2007). Physics in the Multiverse: An Introductory Review https://arxiv.org/abs/0711.4460

[42] CERN Courier (2007) *Physics in the multiverse* https://cerncourier.com/a/physics-in-the-multiverse/

[43] Barnes, L. A. (2012) *The Fine-Tuning of the Universe for Intelligent Life* https://arxiv.org/abs/1112.4647

[44] Tegmark, M. (2008). *The Mathematical Universe. Foundations of Physics* https://link.springer.com/article/10.1007/s10701-007-9186-9, https://arxiv.org/abs/0704.0646

[45] Tononi, G. (2008). *Consciousness as Integrated Information: A Provisional Manifesto.* https://pubmed.ncbi.nlm.nih.gov/19098144

[46] Albantakis, L., et al. (2023). Integrated Information Theory (IIT) 4.0: Formulating the Properties of Phenomenal Existence in Physical Terms. https://arxiv.org/abs/2212.14787

[47] Oizumi, M., Albantakis, L., & Tononi, G. (2014). *From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0.* https://journals.plos.org/ploscompbiol/article?id=10.1371%2Fjournal.pcbi.1003588

[48] Koch, C., Massimini, M., Boly, M., & Tononi, G. (2016). *Neural correlates of consciousness: progress and problems.* https://www.nature.com/articles/nrn.2016.22

[49] Cerullo, M. A. (2015). *The Problem with Phi: A Critique of Integrated Information Theory.* PLoS Computational Biology. https://pmc.ncbi.nlm.nih.gov/articles/PMC4574706/pdf/pcbi.1004286.pdf

[50] Seth, A. K. (2014). *A Predictive Processing Theory of Sensorimotor Contingencies.* Cognitive Neuroscience. https://www.tandfonline.com/doi/full/10.1080/17588928.2013.877880

[51] Northoff, G., & Huang, Z. (2017). *Temporo-Spatial Theory of Consciousness (TTC).* Neuroscience & Biobehavioral Reviews. https://pubmed.ncbi.nlm.nih.gov/35149122/

[52] Baars, B. J. (1988). *A Cognitive Theory of Consciousness.* https://psycnet.apa.org/record/1989-97319-000

[53] Dehaene, S., & Changeux, J. P. (2011). *Experimental and theoretical approaches to conscious processing.* https://www.sciencedirect.com/science/article/pii/S0896627311002583

[54] Baars, B. J. (1997). In the Theater of Consciousness: The Workspace of the Mind. Oxford University Press. https://academic.oup.com/book/27242

[55] Baars, B. J. (2005). Global Workspace Theory of Consciousness: Toward a Cognitive Neuroscience of Human Experience. Progress in Brain Research. https://pubmed.ncbi.nlm.nih.gov/16186014/

[56] Dehaene, S., Changeux, J.-P., Naccache, L., Sackur, J., & Sergent, C. (2006). Conscious, Preconscious, and Subliminal Processing: A Testable Taxonomy. Trends in Cognitive Sciences. https://pubmed.ncbi.nlm.nih.gov/16603406/

[57] Mashour GA, Roelfsema P, Changeux JP, Dehaene S. (2020) *Conscious Processing and the Global Neuronal Workspace Hypothesis.* Neuron. 2020 Mar 4;105(5):776-798. https://pmc.ncbi.nlm.nih.gov/articles/PMC8770991

[58] Penrose, R., & Hameroff, S. (2011). *Consciousness in the Universe: Neuroscience, Quantum Space-Time Geometry and Orch OR Theory. Journal of Cosmology*, 14. http://www.neurohumanitiestudies.eu/archivio/penrose_consciousness.pdf

[59] Tegmark, M. (2000). *Importance of quantum decoherence in brain processes.* https://journals.aps.org/pre/abstract/10.1103/PhysRevE.61.4194

[60] Reimers, J. R., et al. (2009). *Weak, strong, and coherent regimes of Fröhlich condensation and their applications to terahertz medicine and quantum consciousness.* https://www.pnas.org/doi/full/10.1073/pnas.0806273106

[61] Churchland, P. S. (2002). Brain-Wise: Studies in Neurophilosophy. https://archive.org/details/Brain-Wise_Studies_in_Neurophilosophy_by_Patricia_Smith_Churchland

[62] Vinante, A., et al. (2019). *Testing collapse models with levitated nanoparticles: Detection challenge.* https://journals.aps.org/pra/abstract/10.1103/PhysRevA.100.012119

[63] Craddock, T. J. A., et al. (2014). *Anesthetics act in quantum channels in brain microtubules to prevent consciousness.* https://pubmed.ncbi.nlm.nih.gov/25714379

[64] Khan et al. (2024) *Microtubule-Stabilizer Epothilone B Delays Anesthetic-Induced Unconsciousness in Rats* https://www.eneuro.org/content/11/8/ENEURO.0291-24.2024

[65] Frässle, S., Sommer, J., Jansen, A., Naber, M., & Einhäuser, W. (2014). *Binocular rivalry: Frontal activity relates to introspection and action but not to perception*. https://pubmed.ncbi.nlm.nih.gov/24478356

[66] Melloni, L., Mudrik, L., Pitts, M., & Koch, C. (2021). *Making the hard problem of consciousness easier*. https://www.science.org/doi/10.1126/science.abj3259