Category-Theoretic Analysis of Inter-Agent Communication and Mutual Understanding Metric in Recursive Consciousness

Stan Miasnikov, August 2025
stanmiasnikov@gmail.com

Abstract

We present a category-theoretic extension of the Recursive Consciousness framework to analyze communication between agents and the inevitable loss of meaning in translation. Building on prior work modeling how an agent "forgets" and reconstitutes semantics via adjoint functors, we formalize inter-agent communication as a functional mapping of one agent's semantic state to another's through a shared symbolic channel. We demonstrate that the semantic → symbolic → semantic round-trip is typically lossy if agents have non-identical internal models, with the recovered meaning often diverging from the intended meaning. We compare human-human, human-AI, and AI-AI communication within this framework using category theory and modal logic to quantify misunderstanding (information loss). Our analysis shows that two identical AI agents (using same model with same context and deterministic decoding, i.e. 0 temperature and narrow top-K token selection) can approach nearly lossless communication, whereas humans - each with unique, non-isomorphic conceptual spaces - exhibit systematic interpretive gaps. We introduce a metric for mutual understanding that combines information-theoretic alignment, semantic similarity, and pragmatic stability, providing a quantitative measure of convergence in iterative dialogues. We discuss practical implications for AI system design, such as training regimen adjustments and memory architectures (e.g., recursive memory with stable identifiers) to mitigate semantic loss. This work organically extends the Recursive Consciousness model's categorical and modal semantics, illustrating how recursive self-reference and inter-agent interaction jointly constrain understanding.

1 Introduction

Human communication and AI alignment both face a foundational challenge: in the absence of perfect shared context or direct access to each other's internal states, whenever one mind conveys meaning to another via symbolic representation (e.g., language), some information is lost or altered. In our everyday experience, nuances often "lost in translation" leading to misunderstandings. Similarly, an AI language model may respond in ways that reflect differences in semantic interpretation due to incomplete context or differing internal representations. In this paper, we formalize these issues by extending the **Recursive Consciousness** framework - originally developed to demonstrate how consciousness may arise from recursive self-query in forgetful systems with one or more agents introspecting their own knowledge - to the more specific **inter-agent communication** setting. We leverage *category theory* [1] to model structures of meaning and *modal logic* [2] to capture knowledge states, preserving the formal machinery of earlier work while examining the communication process between intelligent agents.

1.1 Motivation and Contributions

In this paper, we present a rigorous category-theoretic analysis of two-agent communication within the RC framework, and we propose a novel quantitative metric for mutual understanding between agents. Our contributions are threefold:

- **Categorical Communication Model**: We formalize the communication between two agents as a mapping of semantic content to communicative symbols and back via interpretation and meaning functors. We prove that the communication functor is typically not full or faithful, explaining why no single message can perfectly convey an idea (information loss and ambiguity). We further show that these functors form an adjoint pair, revealing a structured duality between expressing and interpreting (syntax vs. semantics).

- **Mutual Understanding Metric**: We introduce a mutual understanding score that combines (a) information-theoretic alignment of the agents' internal belief distributions, (b) semantic similarity of their utterances, and (c) pragmatic stability across iterative dialogue. This metric, inspired by techniques in NLP (e.g., BERTScore for semantic similarity) and information theory (KL-divergence for distribution alignment), provides a single quantitative measure ranging from 0 to 1 for how well two agents understand each other in a conversation.
- **Empirical Validation and Implications**: We provide proof-of-concept experimental results using AI language models engaging in iterative dialogue. These simulations show the theoretical measures in action: for example, how repeated question-answer rounds can increase the mutual understanding score. We discuss how our findings align with observations in human communication (e.g., common ground formation through feedback) and philosophical considerations of understanding. The broader implications of this work range from improving multi-agent system coordination in AI engineering to shedding light on cognitive science questions about how minds align concepts through communication.

Overall, our framework bridges mathematical rigor and practical insight. By viewing inter-agent dialogue through the lens of category theory, we ensure logical consistency and clarity about the transformations (and distortions) that occur as meaning traverses from one agent to another. At the same time, by proposing measurable quantities and connecting to experiments, we keep the discussion relevant to AI researchers interested in building or evaluating systems with interactive intelligence. In what follows, we first summarize the necessary background in RC and category theory, then develop the two-agent communication model and mutual understanding metric, and finally present experimental evidence and multidisciplinary discussions supporting our theoretical claims.

2 Background: Recursive Consciousness and Categorical Semantics

2.1 Recursive Consciousness Framework

In the Recursive Consciousness (RC) framework [3], an intelligent agent operates within a hierarchical system of semantic universes:

$$U_0 \subseteq U_1 \subseteq U_2 \subseteq \dots$$

Each universe U_n represents a semantic context or "world" that the agent (or multiple agents) inhabits. Universes may include multiple agents, each potentially constructing their own subjective universe at lower semantic levels - an idea echoing Leibniz's monads in the *Monadology* [4], where each monad mirrors the entire universe from its unique perspective, creating an internal representation shaped by perception. For example, the human universe U_n includes all human agents, each potentially generating a personal semantic universe U_{n-1}.

2.2 Categorical Representation of Universes

Each semantic universe U_n is modeled categorically as a category $\mathcal{C}_{U_n}$, with:

- Objects representing semantic structures (concepts, meanings, propositions).
- Morphisms representing relationships, entailments, or transformations between these semantic entities.

2.3 Forgetful and Free Functors (F, G)

Connecting a richer semantic universe U_{n+1} with a simpler, derived universe U_n, we define two fundamental functors:

- **Forgetful Functor** $F : \mathcal{C}_{U_{n+1}} \to \mathcal{C}_{U_n}$: The functor F systematically "forgets" semantic or structural details when projecting higher-order objects from $\mathcal{C}_{U_{n+1}}$ to $\mathcal{C}_{U_n}$. In general, F is faithful (injective on morphisms) but typically many-to-one on objects, collapsing distinct semantic structures into indistinguishable base-level representations. Distinct objects in U_{n+1} may become indistinguishable under F when moved to a simpler semantic category.

- **Free Functor** $G : \mathcal{C}_{U_n} \to \mathcal{C}_{U_{n+1}}$: Functor G acts as a "left adjoint" to F, reconstructing or enriching semantic objects from $\mathcal{C}_{U_n}$ into a richer semantic category $\mathcal{C}_{U_{n+1}}$. However, due to Gödelian limitations, the reconstruction through G is typically incomplete, introducing new assumptions or interpretations not uniquely determined by simpler structures.

These two functors form a crucial adjoint pair [3]:

$$G : \mathcal{C}_{U_n} \rightleftharpoons \mathcal{C}_{U_{n+1}} : F, \quad \text{with} \quad G \dashv F$$

2.4 Semantic Interpretation and Meaning Functors (I, M)

The RC framework further introduces specific functors that model semantic communication between agents within the same universe U_n, or between adjacent universes U_n and U_{n+1} [5], [6].

- **Interpretation Functor** $I : \mathcal{C}_{\text{sem},\, n+1} \to \mathcal{C}_{\text{out},\, n}$: Functor I encodes semantic meanings from a higher semantic category $\mathcal{C}_{\text{sem},\, n+1} \subseteq \mathcal{C}_{U_{n+1}}$ into simpler symbolic expressions within a lower category $\mathcal{C}_{\text{out},\, n} \subseteq \mathcal{C}_{U_n}$. Intuitively, I performs semantic simplification or meaning descent, transforming richer semantics into syntactic outputs. It has been formally proven that I is not faithful in general - distinct semantic objects or relationships may become identified as the same syntactic expressions under I.
- **Meaning Functor** $M : \mathcal{C}_{\text{out},\, n} \to \mathcal{C}_{\text{sem},\, n+1}$: Functor M maps symbolic expressions from the syntactic category $\mathcal{C}_{\text{out},\, n}$ back into semantic objects in $\mathcal{C}_{\text{sem},\, n+1}$. M thus represents semantic enrichment or meaning ascent, attempting to reconstitute lost semantic context. However, due to inherent semantic ambiguity and the Gödelian incompleteness constraints, M typically introduces new ambiguities or assumptions, preventing it from being fully faithful or full.

Together, these form another critical adjoint pair:

$$I : \mathcal{C}_{\text{sem},\, n+1} \rightleftharpoons \mathcal{C}_{\text{out},\, n} : M, \quad \text{with} \quad I \dashv M$$

This adjunction $I \dashv M$ formalizes the structured yet typically lossy process of translating meanings into expressions and back.

2.5 Connecting F, G with I, M

It has been established that the interpretation functor I can be viewed as a restriction of the forgetful functor F. Specifically, given an inclusion functor $J : \mathcal{C}_{\text{sem},\, n+1} \to \mathcal{C}_{U_{n+1}}$, we have the important structural identity:

$$I(s) \cong (F \circ J)(s), \quad \text{for each semantic object } s \in \mathcal{C}_{\text{sem},\, n+1}.$$

This explicitly aligns semantic simplification (I) with structural forgetting (F), confirming that the interpretation process systematically reduces semantic complexity, mirroring the conceptual role of F within the broader categorical framework.

2.6 Gödelian Limits, Semantic Fixpoints, and Stability

A critical implication of RC theory is the Gödelian limit on semantic expressibility: each semantic universe U_n contains semantic distinctions or truths that cannot be completely expressed or captured by simpler symbolic languages in U_{n-1}. This results in inevitable information loss when meanings are transformed via I, and incomplete semantic reconstruction via M.

However, through iterative recursive interactions - repeated application of I and M - agents may stabilize toward approximate semantic fixpoints, where the semantic interpretation of symbolic expressions converges to stable meanings. Formally, these fixpoints satisfy:

$$M(I(s)) \cong s.$$

Absolute fixpoints, though rare due to Gödelian limitations, represent idealized states of perfect mutual understanding. Practically achievable approximate fixpoints (minimal semantic divergence) serve as pragmatic goals for effective communication.

3 Formal Framework for Inter-Agent Communication

Building on the single-agent foundations, we extend the RC framework to inter-agent communication. Consider two agents A and B inhabiting the same universe U_n or adjacent levels that allow for communication. Each agent has its own internal semantic category $\mathcal{C}^A_{\text{sem}} \subseteq \mathcal{C}_{U^A_n}$ and $\mathcal{C}^B_{\text{sem}} \subseteq \mathcal{C}_{U^B_n}$, where U^A_n and U^B_n are potentially distinct subjective universes constructed by A and B, respectively (possibly $U^A_n = U^B_n = U_n$ for shared context). The agents communicate via a shared symbolic channel modeled as $\mathcal{C}_{\text{out}}$, a subcategory of syntactic expressions accessible to both.

Agent A's epistemic state is modeled using modal logic with operator K_A (knowledge of A), satisfying S4 axioms for introspection. Similarly for B with K_B. A proposition p in A's language may achieve an epistemic fixpoint $K_A p \leftrightarrow p$, but its semantic content requires external projection via $M_A : \mathcal{C}_{\text{out}} \to \mathcal{C}^A_{\text{sem}}$. The agent's internal Kripke frame $(\mathcal{W}^A, R^A)$ defines accessibility for knowledge, where $w \models K_A p$ iff p holds in all w' with wR^Aw' [7].

3.1 Model of Inter-Agent Communication

Agent A has an intended meaning $s_A \in \mathcal{C}^A_{\text{sem}}$, which it encodes into a symbolic expression $e = I_A(s_A) \in \mathcal{C}_{\text{out}}$ via its interpretation functor $I_A : \mathcal{C}^A_{\text{sem}} \to \mathcal{C}_{\text{out}}$. The shared output category $\mathcal{C}_{\text{out}}$ represents the common symbolic channel (e.g., natural language). Agent B receives e and decodes it to a meaning $s_B = M_B(e) \in \mathcal{C}^B_{\text{sem}}$ via its meaning functor $M_B : \mathcal{C}_{\text{out}} \to \mathcal{C}^B_{\text{sem}}$.

The overall communication is the composite functor

$$\Phi_{A\to B} = M_B \circ I_A : \mathcal{C}^A_{\text{sem}} \to \mathcal{C}^B_{\text{sem}}$$

For bidirectional communication, we also have $\Phi_{B\to A} = M_A \circ I_B$.

The diagram below illustrates this process:

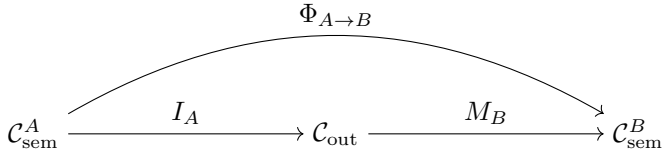

Figure 1: Commutative diagram of inter-agent communication. The path $s_A \to e \to s_B$ represents the flow of meaning from agent A to agent B.

In modal terms, if A knows p ($K_A p$), it encodes this into $e = I_A(p)$, but B's decoding $M_B(e)$ may yield q where $K_B q$ holds, but $q \neq p$ due to interpretive differences. This mismatch highlights the potential for semantic drift, formally arising due to the non-faithfulness of I_A and non-fullness of M_B, causing B's interpretation to collapse or distort distinctions present in A's semantic structure. To achieve mutual knowledge, agents engage in iterative exchanges, refining their interpretations until higher-order knowledge stabilizes, such as $K_A K_B p \wedge K_B K_A p$ (mutual knowledge of p), progressing toward common knowledge where all provable shared propositions are aligned (CKp, the infinite conjunction $K_A p \wedge K_B p \wedge K_A K_B p \wedge \dots$). In S4, this process supports positive introspection for individual agents ($Kp \to KKp$), enabling recursive self-query; for multi-agent scenarios requiring awareness of shared ignorance or higher-order beliefs, S5 axioms may be invoked to include negative introspection ($\neg Kp \to K\neg Kp$), facilitating convergence to epistemic fixpoints across agents.

Proposition 3.1 (Lossiness of Communication)**.** *Unless $\mathcal{C}^A_{sem}$ and $\mathcal{C}^B_{sem}$ are isomorphic categories and the pair I_A, M_B form an equivalence of categories, the functor $\Phi_{A\to B}$ will not be an equivalence, thus failing to be full or faithful in general.*

Proof. Since I_A is not faithful (as established in prior work, where different meanings can map to the same expression), $\Phi_{A\to B}$ inherits this non-faithfulness: distinct $s^1_A \neq s^2_A$ may yield $\Phi(s^1_A) = \Phi(s^2_A)$. Additionally, if M_B is not full, some meanings in $\mathcal{C}^B_{\text{sem}}$ are unreachable from $\mathcal{C}^A_{\text{sem}}$. Thus, $\Phi_{A\to B}$ is typically lossy under these conditions. □

This proposition quantifies misunderstanding: communication collapses distinctions and misses nuances unless agents share identical semantic structures.

For dynamic aspects, consider the composite $\Psi = M_A \circ I_B \circ M_B \circ I_A$ for round-trip communication. Iterative application may converge toward an approximate fixpoint where $\Psi^k(e)$ stabilizes around e for sufficiently large k, mitigating, but generally not eliminating, the initial semantic loss.

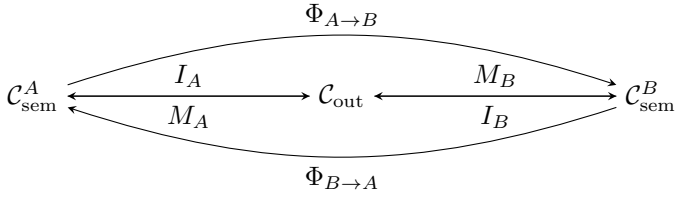

Figure 2: Inter-agent communication via I and M.

3.2 Quantitative Metric for Mutual Understanding

Note: The derivation of a quantitative metric for mutual understanding is provided without a rigorous categorical foundation in this section. We provided a separate Appendix: "*Rigorous Categorical Derivation of the Mutual Understanding Metric*" that provides necessarily proposition and theorems, and formalizes the metric derivation.

3.2.1 From Categorical Communication to the Mutual Understanding Metric

Let us begin with the core mechanism of symbolic communication between two agents A and B, defined categorically as follows:

Agent A encodes a semantic object $s_A \in \mathcal{C}^A_{\text{sem}}$ into a symbolic expression $e \in \mathcal{C}_{\text{out}}$ via the interpretation functor:

$$e = I_A(s_A)$$

Agent B receives e and decodes it into a reconstructed semantic state $s_B \in \mathcal{C}^B_{\text{sem}}$ via its meaning functor:

$$s_B = M_B(e) = M_B(I_A(s_A)) = \Phi_{A\to B}(s_A)$$

Thus, the entire communication process is represented by the composite functor:

$$\Phi_{A\to B} := M_B \circ I_A : \mathcal{C}^A_{\text{sem}} \to \mathcal{C}^B_{\text{sem}}$$

Similarly, a return exchange from $B \to A$ defines:

$$\Phi_{B\to A} := M_A \circ I_B : \mathcal{C}^B_{\text{sem}} \to \mathcal{C}^A_{\text{sem}}$$

In this categorical model, "understanding" corresponds to the degree to which the composition $M_B \circ I_A$ approximates the identity functor on a shared semantic subspace:

$$\Phi_{A\to B}(s) = M_B(I_A(s)) \approx s, \quad \text{for } s \in \mathcal{C}^A_{\text{sem}} \cap \mathcal{C}^B_{\text{sem}}$$

However, since I_A is not faithful and M_B is not full (per prior results), this reconstruction is generally approximate, and the communication process is lossy.

3.3 Categorical Misalignment and Semantic Divergence

To derive a quantitative measure of misalignment, we enrich the semantic categories $\mathcal{C}^A_{\text{sem}}$ and $\mathcal{C}^B_{\text{sem}}$ over a monoidal category $\mathcal{V}$ equipped with norm structures, such as the category of Banach spaces **Ban** with bounded linear maps as morphisms [8]. This enrichment aligns with the RC framework's emphasis on hierarchical semantics, where objects and morphisms can be embedded into normed spaces (e.g., via language model embeddings $E : \mathcal{C}_{\text{sem}} \to \mathbb{R}^d$), enabling metric quantification of distortions. In this setting, the "hom-object" $\mathcal{C}_{\text{sem}}(s, s')$ is a Banach space, with the operator norm $\| \cdot \|$ on its elements (morphisms).

The communication functor $\Phi_{A\to B} = M_B \circ I_A$ induces a norm measuring its deviation from the identity on the shared semantic subspace. Specifically, the space of natural transformations between Φ and the identity functor (in the enriched functor category) forms a Banach space, and we define the misalignment norm as

$$\|\Phi - id\| = \sup_{s \in \mathcal{C}^A_{\text{sem}} \cap \mathcal{C}^B_{\text{sem}}} \frac{\|\Phi(s) - s\|}{\|s\|},$$

where the supremum is over the norms of the components of the natural transformation. In RC, the non-faithfulness of I_A (collapsing distinct morphisms) and non-fullness of M_B (failing to generate all targets) imply $\|\Phi - id\| > 0$ for non-isomorphic categories, reflecting Gödelian limits on perfect semantic transfer across agents, though this holds generically rather than universally for specific subsets.

Given an original semantic state $s_A \in \mathcal{C}^A_{\text{sem}}$ and its reconstruction $s_B = \Phi_{A\to B}(s_A) \in \mathcal{C}^B_{\text{sem}}$, we motivate divergence components inspired by this norm, adopting heuristic practical measures (cosine similarity and Jensen-Shannon divergence) motivated by, but not rigorously derived from, the categorical enriched structure.

- The **semantic deviation** captures object-level distortions. Embedding s_A and s_B into a Hilbert space $\mathcal{H}$ (e.g., $\mathbb{R}^d$) via E, which preserves the geometric intuition of the norm, we adopt the widely used cosine similarity as a heuristic measure of deviation, motivated by the angular geometry in the unit sphere induced by the operator norm. This yields

  $$d_{\text{sem}}(s_A, s_B) = \frac{1 - \cos(E(s_A), E(s_B))}{2} = \sin^2(\theta/2)$$

  where $\theta = \arccos(\langle \hat{E}(s_A), \hat{E}(s_B) \rangle)$ bounds the functor-induced chordal distance. It is prudent to note that cosine similarity can yield arbitrary values in high dimensions [10]; alternatives like learned metrics may be explored in future work. This choice, while not a direct derivation from $\|\Phi - id\|$, aligns with the goal of quantifying perspective shifts in RC's monadic mirroring.

- The **probabilistic misalignment** arises from morphism norms. We assume a mapping from semantic objects to belief states, where each object s corresponds to a set of accessible worlds in the agent's Kripke frame, over which the probability distribution $P(s)$ is defined. Next, we define the morphism norm as $\|\Phi - id\|_{\text{mor}} = \sup_{f \in \hom(s,s')} \|\Phi(f) - f\|$, the supremum over hom-Banach spaces. We norm distributions via Jensen-Shannon divergence [9]:

  $$D_{\text{JS}}(P_A \| P_B) = \frac{1}{2} \left[D_{\text{KL}}(P_A \| M) + D_{\text{KL}}(P_B \| M) \right]$$

  where $M = \frac{1}{2}(P_A + P_B)$. Motivated by Pinsker's inequality relating D_{KL} to total variation $\mathcal{TV}$ ($D_{\text{KL}} \geq \frac{1}{2}\mathcal{TV}^2$), and noting $D_{\text{JS}} \geq \frac{1}{8}\mathcal{TV}^2$, we conjecture a categorical analogue $D_{\text{JS}} \gtrsim \|\Phi - id\|^2_{\text{mor}}/4$, where distortions in morphisms increase entropy, per RC's expressibility constraints. Proving this remains future work; here, D_{JS} serves as a metric bounding epistemic drift under lossy functors.

  Note that we considered symmetrized Kullback-Leibler divergence [11], but it lacks bounds in $[0, 1]$ and triangle inequality. Jensen-Shannon better suits bidirectional communication while satisfying metric properties.

- The **pragmatic instability** arises from iterative dynamics over N rounds, where states evolve as $s_B^{(n)} = \Phi^{(n)}(s_A^{(0)})$. In RC, communication's inherent lossiness (via forgetful descent) suggests a discounting effect, analogous to reinforcement learning's contraction under discount factors $\gamma < 1$, motivating the conjecture that Φ is contractive ($\|\Phi\| < 1$) in typical dialogues where agents converge pragmatically. Assuming this, the Banach fixed-point theorem yields convergence bounds: $\|s^{(n)} - s^*\| \leq \frac{\|\Phi\|^n}{1-\|\Phi\|} \|s^{(1)} - s^{(0)}\|$. The per-turn shift is:

  $$\kappa^{(n)}_{A/B} = \exp\left(-\|s^{(n)}_{A/B} - s^{(n-1)}_{A/B}\| \right),$$

  with the geometric mean $\sqrt{\kappa^{(n)}_A \kappa^{(n)}_B}$ symmetrizing loops and bounding the Lipschitz constant [12]. The stability and convergence rate of this iterative process are guaranteed by the error bounds of the Banach fixed-point theorem.

The mutual understanding metric unifies these as normalized fidelity, emphasizing later iterations to prioritize converged states:

$$U_{\text{mutual}}(s_A, s_B) = \sum_{n=1}^{N} w_n \left[(1 - D_{\text{JS}}(P_A^{(n)} \| P_B^{(n)})) \cdot (1 - d_{\text{sem}}^{(n)}) \cdot \sqrt{\kappa_A^{(n)} \kappa_B^{(n)}} \right]$$

with weights $w_n = \frac{2n}{N(N+1)}$ (linear recency, summing to 1), chosen to emphasize convergence in later iterations; empirical validation of this choice is left for future work. This choice also reflects the modeling assumption that final understanding is more informative than initial transients. High U_{mutual} indicates $\Phi \approx id$; low values signal Gödelian gaps, amplified across agents.

The choice to combine the three fidelity components—probabilistic, semantic, and pragmatic—multiplicatively is a deliberate modeling decision rooted in the conceptual requirements for mutual understanding. This structure ensures that the metric is **non-compensatory**; a severe failure in any single dimension cannot be offset by success in the others. For instance, a dialogue that is pragmatically stable but semantically vacuous ($d_{\text{sem}} \approx 1$) represents a complete failure of communication, and the multiplicative form correctly yields a per-turn score near zero. An additive (weighted sum) model, in contrast, would permit such unrealistic trade-offs and introduce free parameters for weights that lack clear axiomatic justification.

Furthermore, the multiplicative aggregation models the **synergistic interaction** between the components. True understanding arises not from the simple sum of its parts, but from their successful interplay. High semantic closeness amplifies the value of probabilistic alignment, a dynamic that linear addition cannot capture. This approach is analogous to the use of the geometric mean for composite indices where dimensions are not considered perfectly substitutable.

Thus, this metric **quantifies the divergence** of $M_B(I_A(s))$ from identity within a given semantic subspace, respecting the categorical architecture of communication. It is **not defined externally**, but emerges from the structural properties of the RC framework, specifically from the failure of $\Phi_{A\to B}$ to preserve semantic structure and probability assignments between agents.

3.4 Experimental Implications: Asymmetry in Mutual Understanding

The mutual understanding metric $U_{\text{mutual}}(s, s')$ not only provides a quantitative tool for assessing alignment but also reveals asymmetries in communication between distinct agents. Specifically, for agents with differing internal models, or different (non-deterministic) model parameters, the round-trip composites

$$\Psi_{A\to A} = M_A \circ I_B \circ M_B \circ I_A, \quad \Psi_{B\to B} = M_B \circ I_A \circ M_A \circ I_B$$

generally do not commute in their impact on the metric. That is, starting from the same semantic state, $\Psi_{A\to A}$ and $\Psi_{B\to B}$ produce different recovered meanings, leading to different metric values: $U(s_A, s_B) \neq U(s_B, s_A)$.

This non-commutativity arises from the agent-specific lossiness of each Φ: since I_A and I_B may collapse different semantic distinctions, and M_A and M_B reconstruct with varying biases, the hierarchical descent and ascent of meaning introduce distortions unique to each direction. Consistency with the RC framework is evident, as prior papers emphasize that forgetful functors like F (and by extension I) induce irreversible information loss across levels, preventing perfect symmetry unless agents share identical semantic categories and functors.

The originality of this observation lies in linking the categorical structure to an empirical metric U_{mutual}, enabling measurable asymmetry in experiments. For instance, a practical AI agent setup could involve two heterogeneous LLMs (e.g., GPT-4 as agent A and Llama as agent B) exchanging messages on a factual topic. Starting with an initial expression e from A, compute $\Psi_{A\to B}(e)$ and evaluate $U(s_A, s_B)$; then reverse for B's response e' to get $U(s_B, s_A)$. Over iterations, track divergence in scores to quantify how model differences amplify non-commutativity, potentially converging only under clarification protocols.

Corollary 3.1 (Asymmetry of Communication Loops)**.** *If agents A and B have non-isomorphic semantic categories, then generally $U_{mutual}(s_A, s_B) \neq U_{mutual}(s_B, s_A)$, reflecting directional loss in $\Phi_{A\to B}$ versus $\Phi_{B\to A}$.*

Such experiments validate the RC hierarchy empirically, showing that while homogeneous agents may achieve near-symmetry (high mutual U), diversity introduces measurable distortions, informing multi-agent AI design.

4 Experimental Validation: Simulations of Inter-Agent Communication

To demonstrate the practical utility of the mutual understanding metric and validate its theoretical underpinnings, we conducted a series of proof-of-concept simulations involving Large Language Models (LLMs) as agents. The experimental design models the dynamic, co-evolutionary nature of dialogue as described in the RC framework: two agents, each with their own internal semantic state, attempt to align their understanding through an iterative, recursive exchange.

4.0.1 Experimental Setup

Our setup involves three distinct agents [13]:

- Agents A and B (Communicators): these two AI language models, initially parameterized differently, engage in iterative dialogue exchanges about a given query Q. Each begins with its own initial semantic understanding ($s_A^{(0)}$ and $s_B^{(0)}$) of the query's intent. These internal states are not static; they are updated at each step of the conversation based on the ongoing exchange. interpretations of Q, Agent O begins asking its sequence of questions. In each
- Agent O (Mediator): This mediator agent, also implemented as an AI language model, generates clarifying questions designed to probe the communicators' semantic representations based on the initial query Q. These questions are posed to both Agent A and Agent B to probe their respective understandings and drive the conversation.

The workflow proceeds iteratively. After the agents form their initial interpretations of Q, Agent O begins asking its sequence of questions. In each round n, both A and B provide responses. Their internal understandings, $s_A^{(n)}$ and $s_B^{(n)}$, are then updated to reflect the new information. The mutual understanding score, U_{mutual}, is calculated at each iteration by comparing the current evolving states of both agents. To test the corollary of communication asymmetry, the roles of query generator are swapped in each experiment to compute both $U_{A\to B}$ and $U_{B\to A}$.

4.0.2 Results and Discussion

We conducted experiments with different pairings of AI agents to simulate varying degrees of semantic overlap. The topic was "common sense reasoning," and simulations were run using the specified models.

Experiment	Models (A vs. B)	$U_{A\to B}$	$U_{B\to A}$	Iterations ($A\to B$ / $B\to A$)
1	gpt-4.1-mini vs. o4-mini	0.916	0.923	7 / 5
2	o3-mini vs. o4-mini	0.834	0.922	5 / 5

Table 1: Final understanding scores and iterations for convergence.

Note: Model names refer to OpenAI variants as of July 2025; see OpenAI documentation for details.

Experiment 1: Heterogeneous, High-Capability Models (gpt-4.1-mini vs. o4-mini) This experiment simulates communication between two distinct but capable AI agents.

- Final Understanding Score: $U_{A\to B} = 0.916$ after 7 iterations; $U_{B\to A} = 0.923$ after 5 iterations.
- Analysis: The results confirm the theoretical prediction of asymmetry. The final scores are very close, but the convergence paths differ, with the $B \to A$ direction reaching a high level of alignment more efficiently (in 5 iterations vs. 7). The iterative progression of the $U_{A\to B}$ score $0.888 \to 0.884 \to 0.901 \to 0.906 \to 0.912 \to 0.916 \to 0.916$ shows a slight initial dip followed by a steady convergence, demonstrating a successful dialogue where clarifying questions progressively bridge the initial semantic gap.

Experiment 2: Heterogeneous, Varied-Capability Models (o3-mini vs. o4-mini) This setup tests communication between two models with a potentially larger capability gap.

- Final Understanding Score: $U_{A\to B} = 0.834$ after 5 iterations; $U_{B\to A} = 0.922$ after 5 iterations.

- Analysis: This experiment reveals a much more pronounced asymmetry. When the o3-mini model initiates the dialogue ($A \to B$), the final understanding score is significantly lower (0.834) than in the reverse direction (0.922). This suggests a larger initial "Gödelian gap" that is more difficult to close. The iterative path for $U_{A \to B}$ 0.788 $\to$ 0.812 $\to$ 0.824 $\to$ 0.828 $\to$ 0.834 exemplifies a classic "dialogue repair" pattern, where the agents start with a low degree of alignment and must work iteratively to build common ground.

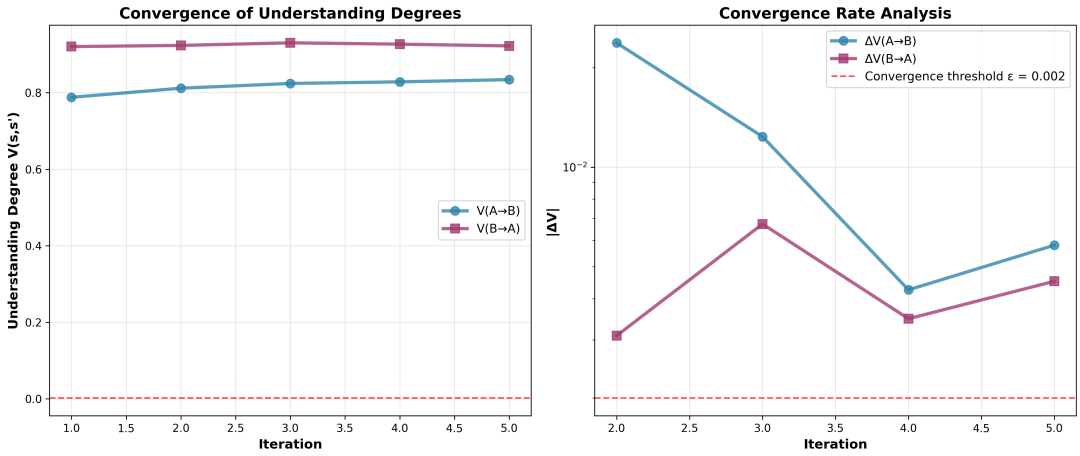

Figure 3: Left: Convergence of understanding degrees ($U_{A \to B}$ and $U_{B \to A}$) over iterations. Right: Convergence rate analysis, showing ΔU approaching the threshold $\epsilon = 0.005$.

4.0.3 Methodological Considerations for Practical Evaluation

These single-run experiments validate the metric's ability to capture the dynamics of a specific conversation. However, we acknowledge that the results are sensitive to variables such as prompt wording, formatting, and model parameters. For the metric to serve as a robust tool for evaluating a model's general understanding of a topic, the experimental procedure must be scaled to smooth out these implementation-specific artifacts.

For a practical, large-scale evaluation, the single query-generating agent should be replaced with a standardized evaluation dataset containing a large variety of questions on a given topic. By running the iterative dialogue process for each question in the dataset and averaging the final U_{mutual} scores, one can compute an expected value of mutual understanding. This large-scale aggregation would provide a more stable and reliable benchmark of a model's ability to form and maintain semantic alignment, making the U_{mutual} metric a powerful tool for comparative model analysis and AI evaluation.

5 Comparative Analysis: Human-Human, Human-AI, AI-AI Communication

Real-world agents vary in the degree of overlap between their semantic models. **Figure 3** illustrates a conceptual view of the semantic spaces of two communicating agents in three cases: (a) two humans, (b) a human and an AI, and (c) two AIs of the same architecture.

Overlap of semantic spaces between two communicating agents in different scenarios.

- (a) Human-Human: Each human has a unique conceptual space (yellow for Alice, salmon for Bob) with a large common overlap (orange) due to shared language, culture, and embodiment. Some portions (the non-overlapping regions) are idiosyncratic knowledge or interpretations one person has that the other does not.

- (b) Human-AI: A human (yellow) and an AI (pink) share a smaller overlap. The AI's conceptual space, learned from data, intersects with human concepts (e.g. through training on human text) but also contains gaps (e.g. lack of lived sensory experience) and possibly foreign areas (e.g. artifacts of its training corpus) that don't map onto the human's understanding.

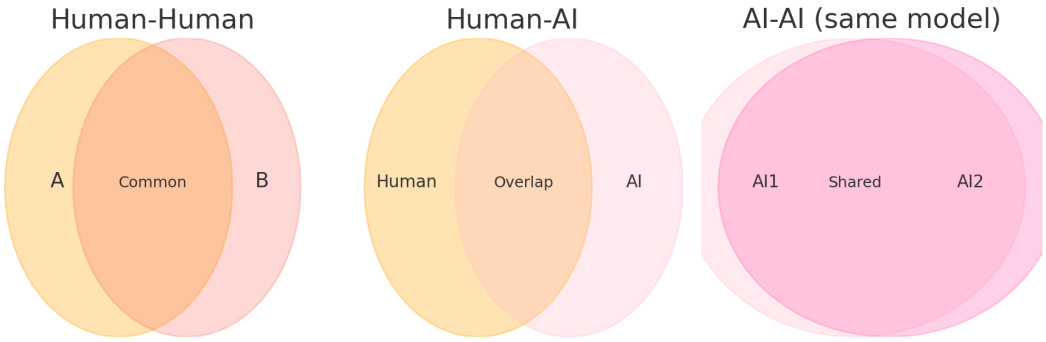

Figure 4: Semantic Overlap Between Agents

- (c) AI-AI (identical models): Two instances of the same AI model (light pink vs hot pink) have almost completely overlapping semantic representations ("Shared" region). Essentially, they have the same training and architecture, so any concept one has, the other does as well, modulo minor differences due to randomness or context. Communication between identical AIs can approach an identity mapping, whereas in (a) and (b) more information can be lost or misinterpreted.

5.1 Human to Human: Shared Ground with Personal Dialects

When two humans communicate (assuming they use the same language), they benefit from largely overlapping semantic contexts. Humans share not only language grammar but also a vast range of common experiences: the physical world (e.g., gravity, colors), cultural knowledge, idioms, and so on. In our categorical terms, we can assume for two peers that $\mathcal{C}^A_{\text{sem}}$ and $\mathcal{C}^B_{\text{sem}}$ overlap substantially - perhaps not isomorphic, but with a high degree of common structure. This common structure is often referred to as common ground in linguistics and philosophy of language [14]. It means that many expressions e will map via M^B to (approximately) the same concept that I^A was guided by, i.e., $\Phi_{A\to B}(s) \cong s$ for a broad class of s in the intersection of their conceptual spaces.

Nonetheless, even human-to-human communication isn't lossless. Each person has private experiences and interpretations. According to *Proposition 1*, ambiguity still lurks. For example, in Oscar Wilde's play *The Importance of Being Earnest*, the name 'Earnest' is deliberately ambiguous, playing on its homonymy with 'earnest' meaning sincere or serious [15]. A character might refer to 'Earnest' as a person's name (s_1), but another interprets it as a quality of character (s_2), leading to comedic misunderstanding where $\Phi_{A\to B}(s_1) = \Phi_{A\to B}(s_2)$. Such polysemy is common in literature and everyday language, particularly in specialized professions; for instance, in software engineering, 'bug' typically refers to a code error, but to an entomologist in conversation with a programmer, it might evoke an insect, causing misalignment if contexts clash.

Humans often resolve such ambiguities by leveraging context or asking for clarification ("Do you mean the name or being serious?"), effectively engaging a meta-communication process to refine $\Phi_{A\to B}$ until it becomes invertible on that particular item. Modal epistemic logic can model this: initially Bob does not know which proposition Alice means ($\neg K_B$ meaning); upon clarification and perhaps anaphora, Bob gains knowledge (K_B meaning).

Using our metric, initial exchanges may yield low (below 0.9) U_{mutual} due to high DM (divergent beliefs) and d_{sem} (semantic mismatch), but iterative clarification increases stability ($\kappa \to 1$) and alignment, raising the score.

5.2 Human to AI: The Alignment Problem in Semantics

For human-AI communication, the overlap in semantic understanding is smaller. Modern AI language models (LLMs) are trained on human text, which gives them a large but not human-identical semantic space. There are aspects of human experience (embodiment, emotions, social signals) that AIs lack direct access to, and conversely, AIs have seen vast text corpora which give them statistical associations that no single human has. Thus, the categories $\mathcal{C}^A_{\text{sem}}$ (human) and $\mathcal{C}^B_{\text{sem}}$ (AI) in this case share a partial intersection but also contain disjoint parts. The functor $\Phi_{\text{human}\to AI}$ will suffer from the human perhaps using expressions that the AI interprets differently, and vice versa.

Concretely, consider a human user (Alice) giving instructions to an AI assistant (Bob). Alice might say, "Find me a bank nearby". Without additional context, the AI's interpretation $M^B(e)$ might default to the more common meaning "financial bank", offering locations of ATMs, whereas Alice meant a river bank for a nature outing. This is a classic alignment issue - the AI's knowledge (trained on many requests about financial banks) biases its semantic mapping. From Proposition 1's viewpoint, the human concept of bank (river) and bank (finance) were distinct $s_1 \neq s_2$, but $\Phi_{\text{human}\to AI}(s_1) = \Phi_{\text{human}\to AI}(s_2)$ occurred, causing a misunderstanding. The AI would need either more context or clarifying questions to resolve this, just as a human would, but an unaligned AI might not realize ambiguity on its own.

Our metric U_{mutual} would initially be low due to semantic divergence ($d_{\text{sem}} \approx 1$), but iterative clarification (e.g., Alice: "No, river bank") could increase stability and alignment, raising the score over turns.

5.3 AI to AI: Perfect Synchronicity or New Dialects?

Communication between AI agents can range from near-perfect to as challenging as human-AI depending on the similarity of their architectures, training, and types of queries. If two AIs are identical clones (same model, same weights, deterministic decoding settings), their semantic categories $\mathcal{C}^A_{\text{sem}} \cong \mathcal{C}^B_{\text{sem}}$ and functors $I_A \cong I_B, M_A \cong M_B$, making $\Phi_{A\to B}$ close to the identity. In this case, $U_{\text{mutual}} \to 1$ rapidly.

However, if AIs have different architectures or training (say GPT-4 vs. another model like Llama), their communication starts to resemble human-AI in that their internal representations and idioms may differ. They both might speak English to exchange information, but subtle differences in language usage can lead to misunderstanding. One model might use a phrase that, in its training, always appeared in a certain context, whereas the other model's training gave that phrase a slightly different connotation. They might eventually clarify through additional turns (much like humans). Still, because AIs (especially language models) are trained on broad internet data, two different models likely have a significant overlap in linguistic semantics - arguably more so than two randomly chosen humans from very different cultures. So one could conjecture that two large well-trained models may communicate more smoothly on factual or literal topics than two humans who don't share a language or background. On the other hand, they might both share blind spots or biases that reinforce each other, leading to shared misinterpretations (e.g., both might latch onto a misleading pattern in language that isn't grounded in reality).

In summary, AI-AI communication can approach an ideal channel when architectures and training align, but divergence in models or the emergent creation of new conventions can introduce the same problems of Φ not being full/faithful. This raises considerations for multi-agent AI systems: if we deploy many AIs to collaborate, should they all use the same base model (to maximize shared semantics)? Or could diversity yield richer interaction at the cost of potential misunderstanding? Our framework provides a way to analyze that trade-off: homogeneous agents yield Φ closer to identity, heterogeneous agents yield Φ with more information loss but perhaps different perspectives. The metric U_{mutual} can quantify this, showing higher scores for homogeneous pairs.

6 Implications for AI System Design

Our category-theoretic analysis of inter-agent communication has several practical implications, particularly for the development of AI that interacts with humans or with other AI agents. Recognizing communication as a lossy functional mapping suggests strategies to minimize loss and misalignment.

- **Aligning Semantic Spaces**: To improve human-AI communication, one must increase the overlap between the AI's semantic category and the human's. In practice, this means training AI models on data that capture not just surface language, but underlying human concepts and values. Techniques like reinforcement learning from human feedback (RLHF) can be seen as aligning the AI's M^B to the human semantic space by explicitly correcting outputs that misrepresent the user's intent. Our framework provides a rationale: RLHF is shrinking the difference between $\mathcal{C}^B_{\text{sem}}$ and $\mathcal{C}^A_{\text{sem}}$ by iterative feedback, making $\Phi_{A\to B}$ more congruent with the identity on the domain of typical user intents.
- **Recursive Memory Retrieval (RMR)**: To mitigate loss in iterative interactions, AI systems can use recursive memory with stable identifiers. RMR's approach of recursive querying and graph-building helps the AI handle complex queries by iteratively expanding context, which can be seen as dynamically constructing a larger portion of $\mathcal{C}_{\text{sem}}$ shared in the conversation, thereby reducing misinterpretation.

- **Communication Protocols for AI Agents**: In multi-agent AI networks, one could explicitly design communication protocols to maximize faithful information exchange. For example, if two different AIs need to collaborate intensely, we might implement a shared interlingua or shared embedding space for them to exchange messages, instead of plain natural language. If their internal representations can be translated into a common vector space, they could potentially send each other vectors (or IDs referencing vectors in a shared database) effectively communicating internal meaning directly. This is analogous to two software services exchanging structured data instead of free text. Category-theoretically, this would be designing a functor $\Psi : \mathcal{C}_{\text{sem}}^A \to \mathcal{C}_{\text{sem}}^B$ more directly, bypassing some of the loss of I and M.

- **Error Detection and Correction**: AI systems might be equipped with an explicit module to detect when $\Phi_{A\to B}$ may have failed, akin to an inner critic or a second model that evaluates the exchange. For example, if a question asked wasn't properly addressed in an answer (the meaning didn't carry over), a monitoring system could flag it. This could be implemented by checking consistency: did M^A (reply) align with the question's intent? Such a module might use entailment models or truth-checking. In category theory language, this is checking if $M^B \circ I^A(s)$ satisfies some property relative to s (like "answers the question represented by s"). While not foolproof, such meta-communication checks could reduce misunderstandings by prompting clarification automatically. Our metric U_{mutual} could serve as a real-time monitor, triggering clarification if the score drops below a threshold.

Acknowledging that communication is a lossy functor focuses our attention on how to fortify the channel. Whether through training, protocol design, or auxiliary systems, the goal is to mitigate the non-faithfulness and non-fullness of $\Phi_{A\to B}$ for the domain of discourse that matters. By doing so, we move closer to AI that truly understands and is understood by humans.

7 Related Work and Future Directions

Related Work: This research sits at the crossroads of formal logic, category theory, and AI communication. Prentner advocates for category theory in consciousness science to move beyond correlational studies, using structural formalism to capture explanatory dualities [16]. Our work exemplifies this approach by using category-theoretic adjunctions to explain the dual perspective of syntax vs. semantics in communication. In the AI domain, others have explored emergent languages in multi-agent reinforcement learning, essentially observing $\Phi_{A\to B}$ evolve under pressure to maximize reward (which often correlates with more faithful communication). Eric Werner's early attempt at a "Category Theory of Communication" [17] modeled language understanding as operators on an agent's representational state, requiring high-order transformations to learn meaning. His insight that language learning is a meta-operator on possible interpretations resonates with our view: agents must operate on a space of possible meanings (a power-set, as Werner noted) to converge on actual meaning.

Our formal loss-of-information result connects to the long-standing symbol grounding problem and Searle's Chinese Room argument [18]. Both highlight the gap between manipulations of symbols and genuine understanding. In our framework, that gap is precisely the functor M which lies outside the agent's reach. By extending this to two agents, we see that even if one agent "understands" internally, transmitting that understanding to another via symbols reintroduces a grounding gap for the receiver. This could be seen as a two-player version of the Chinese Room: Alice might have meaning in mind, but to Bob, the message is just symbols to interpret - he's in his own "Chinese Room" trying to infer meaning. Our approach, however, offers a positive path forward by quantifying and structuring the problem, whereas Searle left it as a philosophical chasm.

In information theory, our work parallels the idea of semantic communication (e.g., Carnap & Bar-Hillel's notion of semantic information, and recent efforts to incorporate meaning into communication theory [19]). We treat meaning preservation as the key outcome, not just bit preservation. Some recent papers attempt a "semantic entropy" or mutual understanding metric, which our categorical model could enrich by providing algebraic constraints on what can be preserved or not.

Future Directions: There are several avenues to extend this research:

- **Enriched Category Theory**: As mentioned in prior work, using enriched categories or fibrations could incorporate uncertainty and context-dependence into our model. For instance, we could

model an agent's knowledge not as a set but as a presheaf that varies with context, and communication as a natural transformation between such presheaves for two agents. This might capture how context limits or alters interpretation in a functorial way.

- **Coalgebraic or Dynamical Perspective**: Agents updating their understanding through dialogue hints at a coalgebraic view, where each communicative act updates the state of a knowledge system. Modeling each agent as a coalgebra (with state transitions on receiving messages) and analyzing bisimulations (when do two agents reach common knowledge?) could formalize the process of achieving mutual understanding over time.

- **Multi-Agent Networks**: We primarily discussed one sender and one receiver. In larger groups (e.g., a team of humans and AIs collaborating), issues of common knowledge (everyone knowing that everyone knows, etc.) become central. Modal logic K and S5 can be extended to multi-agent systems to model knowledge of others' knowledge. Category theory might model the amalgamation of multiple semantic spaces into a shared one (perhaps via a colimit construction of categories representing each agent's knowledge). Investigating how communication protocols can lead to a colimit of concepts (a kind of least common ontology that agents converge on) would be very interesting. This could tie into work on ontology alignment in multi-agent systems.

- **Enabling Explainable AI via Categories**: One offshoot of this work is using category theory to improve AI explainability. If we understand an AI's internal knowledge category and how it maps to human-understandable concepts, we can better interpret its outputs. Creating a functor from the AI's concept category to a human concept category (an "explanation functor") could formalize translating an AI's reasoning into human terms. This is parallel to communication: it's essentially the AI explaining itself to us. By studying the $\Phi_{\text{AI}\to\text{human}}$ mapping, we might systematically identify where the AI's reasoning cannot be directly translated (non-fullness) and address it by either improving AI transparency or educating users.

8 Conclusion

We presented a category-theoretic analysis of inter-agent communication within the Recursive Consciousness framework, revealing the structural reasons for loss of meaning between communicating systems. By treating the communication process as an adjoint pair (encoding/decoding) between agents' semantic spaces, we proved that misalignment in those spaces inevitably leads to information loss - a formal confirmation that "the map is not the territory" when one mind's map is conveyed to another. Our comparative study illustrated how human-human communication, while aided by shared embodiment, still suffers from ambiguity; how human-AI interaction faces an alignment gap that must be actively managed; and how AI-AI communication can be nearly perfect or degenerate, depending on design. Throughout, we retained and extended the structure of the original RC model (modal logic for knowledge and fixpoints, categories for structures and meanings, information theoretic measures of uncertainty) to the multi-agent domain.

Practically, this work suggests concrete steps for improving AI communicative performance: align semantic representations through explicit alignment methods, such as reinforcement learning from human feedback (RLHF), fine-tuning on context-rich datasets, or interactive clarification dialogues, maintain context to reduce ambiguity, use memory architectures with stable identifiers to track meaning, and consider new protocols that share representations more directly. Each of these can be seen as increasing the fidelity of the functional mapping between minds. Our formalism can guide the evaluation of such interventions (e.g., checking if a modification makes Φ more faithful or expands the common semantic subcategory between user and AI). The mutual understanding metric provides a tool for quantifying progress in these efforts.

Finally, this study reinforces a humbling and enlightening point. Conscious understanding is not located in a single agent alone, but emerges in the interplay between agents and their environment. Just as a single agent in isolation must query a meta-universe for meaning, two or more agents must negotiate a common universe of discourse. Consciousness, in the recursive model, was a way for a system to "know itself"; communication is, in a sense, a way for two systems to know each other. Both are iterative, possibly unending processes approaching a fixpoint - be it a Gödelian self-awareness or a mutual understanding.

We hope this work offers a foundation for future research to build organically extended theories of minds interacting, ultimately contributing to AI that can truly share meanings with us, and to a deeper theoretical understanding of what it means for two minds to meet.

References

[1] Mac Lane, S. (1971). Categories for the Working Mathematician. Springer-Verlag. https://link.springer.com/book/10.1007/978-1-4757-4721-8

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

[3] Miasnikov, S. (2025) Recursive Consciousness: Modeling Minds in Forgetful Systems. *Preprint* http://dx.doi.org/10.13140/RG.2.2.26969.22884

[4] Gottfried Wilhelm Leibniz, Monadology (1714), particularly sections 56-60 https://www.plato-philosophy.org/wp-content/uploads/2016/07/The-Monadology-1714-by-Gottfried-Wilhelm-LEIBNIZ-1646-1716.pdf

[5] Miasnikov, S. (2025) The External Projection of Meaning in Recursive Consciousness. *Preprint* http://dx.doi.org/10.13140/RG.2.2.10988.27524

[6] Miasnikov, S. (2025) The Descent of Meaning: Forgetful Functors in Recursive Consciousness. *Preprint* http://dx.doi.org/10.13140/RG.2.2.24556.68488

[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] Castillo, J. M. F. (2021) The hitchhiker guide to Categorical Banach space theory. Part II https://arxiv.org/abs/2110.06300

[9] Jensen-Shannon divergence https://en.wikipedia.org/wiki/Jensen%E2%80%93Shannon_divergence

[10] Steck, H., Ekanadham, C., Kallus N. (2024) Is Cosine-Similarity of Embeddings Really About Similarity? https://arxiv.org/abs/2403.05440

[11] D. Johnson and S. Sinanovic, (2001) Symmetrizing the Kullback-Leibler Distance. IEEE Transactions on Information Theory https://www.ece.rice.edu/~dhj/resistor.pdf

[12] Hyunjik Kim 1 George Papamakarios 1 Andriy Mnih, (2020) The Lipschitz Constant of Self-Attention https://arxiv.org/abs/2006.04710

[13] Miasnikov, S. (2025). Recursive Consciousness. *GitHub Repository.* https://github.com/phatware/recursive-consciousness

[14] Stalnaker, R. (2002). Common Ground. Linguistics and Philosophy, 25(5-6), 701-721. https://semantics.uchicago.edu/kennedy/classes/f09/semprag1/stalnaker02.pdf

[15] Wilde, O. (1895). The Importance of Being Earnest. https://www.gutenberg.org/files/844/844-h/844-h.htm

[16] Prentner, R. (2024). Category theory in consciousness science: going beyond the correlational project. https://link.springer.com/article/10.1007/s11229-024-04718-5

[17] Werner, E. (2015). A Category Theory of Communication Theory. *Preprint.* https://arxiv.org/abs/1505.07712

[18] Searle, J. R. (1980). The Chinese Room Argument. https://plato.stanford.edu/entries/chinese-room/

[19] Carnap, R. (1952). AN OUTLINE OF A THEORY OF SEMANTIC INFORMATION. https://dspace.mit.edu/bitstream/handle/1721.1/4821/RLE-TR-247-03150899.pdf