Open Access

OBM Neurobiology

Original Research

Knowledge and Information in Epistemic Dynamics

Vincenzo Manca *

University of Verona, Verona, Italy; E-Mail: vincenzo.manca@univr.it

* **Correspondence:** Vincenzo Manca; E-Mail: vincenzo.manca@univr.it

Academic Editor: Vsevolod Konstantinov

Special Issue: Artificial Intelligence and Neural Networks: New Perspectives for Psychological Research

OBM Neurobiology
2026, volume 10, issue 3
doi:10.21926/obm.neurobiol.2603344

Received: March 27, 2026
Accepted: July 30, 2026
Published: August 11, 2026

Abstract

This paper proposes a general theory of cognitive systems that inverts the conventional relationship between information and knowledge. While classical approaches define knowledge as the byproduct of processed information, we establish knowledge as a primitive concept and formulate information as a measure emerging from the process of assimilating knowledge. The epistemic measure of information provides a formal counterpart to Piaget's assimilation-accommodation paradigm, expressed in terms of knowledge encoding costs. We present a general axiomatic definition of a cognitive system, along with its corresponding epistemic information measure, and demonstrate that Shannon's source information, artificial neural networks, and formal logical theories are all concrete realizations of this unified framework. A central theoretical result is the proof of the *Epistemic Inaccessibility Theorem*, which establishes an intrinsic form of cognitive incompleteness. As a major consequence, we prove Gödel's First Incompleteness Theorem for Peano Arithmetic directly from epistemic inaccessibility. Finally, we discuss the role of multi-level knowledge encodings, highlighting how artificial intelligence and Large Language Models suggest new avenues for modeling reflexivity, reasoning, and autonomous cognitive evolution.

Keywords
Information; knowledge; artificial intelligence; artificial neural networks; machine learning; large language models; transformer linguistic models; cognitive systems; formal theories

1. Introduction

Artificial Neural Networks (ANNs) are the epicenter of artificial intelligence. They were introduced in the foundational paper [1], where neurons contain discrete values, and synapses connecting them are equipped with functions transforming the values of afferent neurons into those of efferent ones [2]. An important shift in perspective emerged after the landmark book by Donald Hebb [3, 4], altering the functional perspective to a dual vision in which synaptic plasticity forms the basis of learning processes. Subsequently, in pioneering works across the field [5-12], neurons became nodes with functions, while synapses were expressed as directed arrows labeled with real numbers called **weights**, representing connection strengths. Under very general assumptions, neural networks in this form serve as universal approximators for continuous functions between hyperspaces of real numbers [13, 14]. Given that such functions can represent many behavioral competencies, neural networks provide a crucial foundation for any computational model of complex human abilities.

Synapse modification through training on examples was modeled by Machine Learning algorithms, which update weights to improve the network's ability to acquire functions related to specific competencies [5-7, 9-12, 15, 16]. In this way, ANNs inverted Turing's original paradigm. While Turing machines and equivalent formalisms are *programmed* to compute explicit functions, ANNs equipped with machine learning methods are *trained* on pairs of values (input, output) to discover the underlying function. The key innovation in this domain was the backpropagation algorithm, rooted in function optimization theory, which adjusts weights to achieve a desired functional behavior. Through backpropagation, errors relative to the expected output propagate backward through the network, updating weights to reduce error until the optimal weight configuration for the target behavior is reached [17-20].

Recently, Large Language Model (LLM) architectures and advanced ANN technology [21] have realized Turing's visionary hypothesis [22] of a talking machine. By scaling to numbers of neurons and synapses comparable to those of human brains, these networks have produced conversational agents capable of generating natural-language dialogue at a level matching typical human conversation in many respects. From this perspective, language has emerged as an essential structural component in the construction of cognitive processes. Leveraging this capacity, neural networks trained via machine learning can solve complex tasks, marking the beginning of a new era in human technology.

This paper explores the new perspectives opened by LLM transformer architectures, but from an inverted angle relative to standard approaches. Rather than asking how to build better engineering systems, we ask: what do transformer models reveal about human cognition? Can we define general principles of cognitive systems that shed light on intelligence, suggest new computational models, and provide rigorous definitions for fundamental psychological concepts? Bridging these questions promises to foster reciprocal advancements between neuropsychology and artificial intelligence.

2. Probabilistic *versus* Epistemic Information

In his foundational work [23], Claude Shannon outlined the mathematical theory of communication, proposing a measure of information based on probability. Given a random variable X taking values in a set A with probabilities $\{p_a\}_{a\in A}$, the quantity of information conveyed by a specific value a is given by

$$-\log(p_a) \tag{1}$$

where log denotes the base-2 logarithm. The intuitive rationale is that information content is inversely proportional to probability and additive for the joint occurrence of independent events (where the joint probability is the product of individual probabilities). Consequently, the logarithm naturally transforms products into sums:

$$\mathrm{Inf}(a, b) = -\log\left(p_{(a,b)}\right) = -\log(p_a \cdot p_b) = -\log(p_a) - \log(p_b) = \mathrm{Inf}(a) + \mathrm{Inf}(b) \tag{2}$$

At the very outset of his paper, Shannon links the measure of information to uncertainty, viewing them as two sides of the same coin. Specifically, the *a priori* probability corresponds to both the information gained when a value occurs and the corresponding reduction in uncertainty. Thus, ignorance decreases by an amount exactly equal to the information acquired.

Within this probabilistic framework, Shannon established crucial concepts that remain foundational for communication theory. However, other essential dimensions of information are absent from this view. In everyday experience, what is highly informative for one individual may be completely meaningless to another; even within the same individual, information that is critical in one context may carry no value in another. Intuition suggests a deep link between information and the receiver's state of knowledge, which evolves continuously as new inputs are processed. This raises a central question: can we define an *epistemic* value of information and quantify it relative to the knowledge state of a system? Capturing the interdependence between information and knowledge requires a framework where knowledge acts as an active component assigning value to data, rather than merely being a passive repository built through their accumulation.

To build such a framework, we must first distinguish between data and information. Data are physical or symbolic events that carry information. Yet, information remains distinct from the data itself, as data can be re-encoded into different representations without altering their information content. Strictly speaking, defining data as elements that "contain" information yields a circular definition that fails to clarify what genuinely constitutes data. In the Shannon paradigm, any stochastic source generates data independently of the agent receiving it. Conversely, the approach presented here inverts this perspective by assuming a cognitive or epistemic agent whose inputs are interpreted as data, and whose processing transforms those inputs into new data, fundamentally altering the agent's internal state. Here, encoding and decoding become central to an intrinsically functional framework: knowledge acts as the system's internal encoding of data, through which the system acquires new competencies—that is, new operational functions—for data processing and environmental interaction.

3. Cognitive Systems

In a series of foundational works [24-27], Jean Piaget and the Geneva school of developmental psychology investigated the conceptual organization of mathematical and physical intuitions during learning processes.

A central paradigm that emerged from their research is the dichotomy between *assimilation* and *accommodation*. Analytically, comprehension operates much like biological digestion. When an external concept enters a student's cognitive framework, it cannot be directly integrated without being *assimilated* into existing mental structures in a form meaningful to the internal organization of the mind. In digestion, complex food is broken down into constituent molecular components by enzymes to provide energy and building blocks for tissues; analogously, incoming data from various modalities (textual, visual, etc.) entering a cognitive system must be "encoded" into formats compatible with the system's internal structure. This internal encoding process constitutes *assimilation*. Conversely, assimilation triggers a complementary mechanism—*accommodation*—whereby the cognitive system reorganizes its internal structure to achieve better assimilation, internal coherence, and functional adequacy.

This constructivist perspective serves as the primary inspiration for the formal definition presented below.

3.1 A General Definition

Let S be a state space—for instance, a Hilbert space of *features*. To incorporate time into the state space S, we transition to an event-state space defined as a measurable space (S, M, p), where M is a σ-algebra of subsets of S, and p is a probability measure [28]. An *event* is defined as a pair (A, t), where $A \in M$ represents a subset of states that may occur at time t (the formal notion of event occurrence can be further refined according to spatial or structural parameters within S).

A cognitive system is a dynamic structure defined over an event-state space. Intuitively, a cognitive system evolves through interactions with its environment via multiple levels of representation. The information content of incoming external data is determined by the computational and structural costs incurred by the internal representation during its assimilation and accommodation phases.

By constructing representations of objects, events, and relations, the system gains the ability to perform inference, generate predictions, and make decisions to interact effectively with the external world. The mathematical substrate of these competencies is the concept of a *function*—tracing back to Euler's *Introductio in Analysin Infinitorum* (1748) and formalised by Alonzo Church in the 1930s via λ-calculus—defined as a mapping that transforms objects into representations such that manipulating the representations yields computational or operational advantages over manipulating the objects directly. A city map illustrates this principle: optimal routes are easily identified because the synthetic representation provides a global structural overview. When symbolic structures represent objects, these functions are called *codes*. Typically, a code is a mapping from strings over a finite alphabet (*codewords*) to a set of encoded objects. Coding theory remains a central pillar of Information Theory [29]; while we do not delve into specific coding schemes here, we emphasize their fundamental role in cognitive dynamics.

At any given moment, a cognitive system occupies a point (or a set of states) in a multidimensional state space. A subset of these states functions as *knowledge states*, possessing explicit representational roles.

Coding—potentially structured across multiple hierarchical levels—is the core engine of cognitive dynamics. Characterizing a cognitive system strictly in terms of states and codes provides an "extensional" formulation, as it abstracts away from physical substrates (whether biomolecular, cellular, or silicon-based). In what follows, we adopt this abstract viewpoint to present an axiomatic definition of a cognitive system. Despite its generality, this formulation yields non-trivial structural properties.

Definition 3.1 A cognitive system Γ is defined by a seven-tuple:

$$\Gamma = (T, \mathrm{Int}_t, \mathrm{Ext}_t, \mathrm{Knl}_t, \gamma_t, \mu, \eta) \tag{3}$$

where Int_t, Ext_t, and Knl_t are non-empty subsets of a state space S, indexed by $t \in T$ representing the temporal lifespan of Γ (equipped with an internal parameter $t \in T$). The function γ_t represents the knowledge encoding, while μ and η denote cost functions associated with data encoding $\gamma_t(d)$ and encoding updates from γ_t to γ_{t+1}, respectively.

The system Γ satisfies the following axioms (where time index t is omitted when conditions hold globally for all $t \in T$):

Knowledge Partition: Int and Ext are non-empty, disjoint sets representing the *internal* and *external* states of Γ, respectively.

Knowledge Internalization: Knl is the set of *knowledge states*, that satisfy Knl ⊆ Int.

Knowledge Encoding: γ is the *knowledge encoding* mapping:

$$\gamma\colon \mathrm{Knl} \rightarrow \mathrm{Int} \cup \mathrm{Ext} \tag{4}$$

Knowledge Irreflexivity: For every argument $x \in$ Knl, the encoding satisfies:

$$\gamma(x) \neq x \tag{5}$$

Knowledge Assimilation–Accommodation: The function inf measures the *epistemic information* of an input datum d, decomposed into two additive components:

$$\inf(d) = \mu\big(\gamma(d)\big) + \eta\big(\gamma(d)\big) \tag{6}$$

where μ represents the *assimilation cost* of d, and η represents the *accommodation cost*, reflecting the structural reorganization of γ required to maintain global consistency and adequacy. The dual nature of this cost function directly reflects the functional role of the encoding γ: assimilation corresponds to the computational cost of evaluating γ, whereas accommodation corresponds to the structural redefinition of γ_t over time.

Definition 3.2 An internal state $s \in$ Int is defined as *accessible* if and only if it lies in the image of a knowledge state under γ (i.e., $s \in \gamma(\mathrm{Knl})$). Accessibility thus strictly requires a representational level mediated by a knowledge state distinct from the state it represents.

Remark 3.3 In a cognitive system, Knl_t may expand and enrich over time; consequently, the cost of acquiring data can change dynamically throughout the system's lifespan depending on its accumulated knowledge. Because Int_t and Ext_t are disjoint, their preimages under γ are likewise disjoint, preventing ambiguity between internal and external knowledge encodings:

$$\gamma^{-1}(\mathrm{Int}_t) \cap \gamma^{-1}(\mathrm{Ext}_t) = \emptyset \tag{7}$$

External states correspond to *input data* conveying environmental information to Γ, whereas a subset of internal states remains accessible to the environment, forming the *output data* of Γ. In the sequel, we denote $I_t = \gamma^{-1}(\mathrm{Int}_t)$ and $E_t = \gamma^{-1}(\mathrm{Ext}_t)$.

We illustrate the broad applicability of this axiomatic framework through three concrete examples.

Example 3.4 (Shannon's Information Source). Consider a fundamental cognitive system that receives a sequence of symbols from a finite alphabet and assimilates them internally via an encoding function γ. The system constructs a binary lookup table using a Huffman coding scheme, assigning shorter codewords to more probable symbols [29]. Here, the external states correspond to input symbol sequences; internal states represent tables of pairs (symbol occurrence, binary codeword); and knowledge states are the binary codewords alongside their occurrence frequencies.

If the assimilation cost μ is defined as the length of the binary representation and the accommodation cost η is zero, the epistemic information of symbol i reduces to $-\log_2 p_i$, where p_i is the empirical frequency of symbol i. This recovers Shannon's classical definition of information quantity.

Classical information theory dictates that the expected codeword length is lower-bounded by the Shannon entropy, and Huffman coding achieves optimality with respect to this measure. Thus, Shannon's information measure can be interpreted as a special case of a minimalist cognitive system where knowledge encoding optimizes average codeword length without structural accommodation.

While powerful for communication channels, Shannon's framework abstracts away the agent's history and knowledge evolution, which are critical for evaluating epistemic information in complex cognitive systems.

Example 3.5 (Artificial Neural Networks). An Artificial Neural Network (ANN) naturally maps onto this cognitive architecture:

- **Internal States:** The complete set of synaptic weights and neuronal activation states at time t.
- **External States:** Input vectors presented to the network, representing environmental stimuli.
- **Knowledge States:** High-dimensional embedding vectors representing semantic concepts, implicitly encoded in the weight configuration.
- **Knowledge Encoding:** The mapping γ corresponding to the forward-pass computation, which transforms an input vector into a specific pattern of neuronal activations (its internal representation).
- **Assimilation Cost (μ):** The computational cost of executing the forward pass (e.g., calculating contextual embeddings in an LLM transformer architecture).
- **Accommodation Cost (η):** The computational complexity of the backpropagation algorithm and the resulting weight updates. This reflects the effort required to adjust knowledge states (weights) to minimize loss. Data consistent with current knowledge incur low error (hence low accommodation cost), whereas surprising inputs or outliers produce large errors, triggering substantial weight updates and high accommodation costs.

Quantitatively, the accommodation cost of a datum *d* equals the value of the loss function driving backpropagation. Because backpropagation adjusts weights specifically to minimize this error toward zero, the loss function directly measures the adaptation cost paid by the network to align with the training objective.

Consequently, training an ANN is an ongoing process of assimilation (forward pass) and accommodation (backpropagation). This dual-cost model provides a formal framework where the information content of a sample is defined by its operational role in knowledge acquisition throughout the learning process.

Example 3.6 (Axiomatic Logical Theory). An axiomatic logical theory is specified by a set of axioms expressed as formulas in a formal logical language. Within these formulas, a deduction relation $\vdash$ is defined, which is a symbolic calculus allowing one to deduce a formula φ from a set of formulas Φ, written as $\Phi \vdash \varphi$. The theory is the set of formulas, called theorems, deduced from the axioms. A relation of semantic logical consequence $\vDash$ is also assumed, based on a formal notion of interpretation. A formula φ holds in a mathematical structure $\boldsymbol{M}$ when it can be interpreted as a proposition that is true in $\boldsymbol{M}$. Then the relation $\Phi \vDash \varphi$ means that φ is true in all the models where all the formulas of Φ are true. In this way, we can distinguish two classes of formulas: theorems deduced by $\vdash$ from the axioms, and logical consequences of the axioms according to $\vDash$.

An axiomatic logical theory is a cognitive system in which:

- **Internal States (Int):** The true propositions of the theory.
- **External States (Ext):** The false propositions.
- **Knowledge States (Knl):** The theorems of the theory.
- **Time (*T*):** Measured by the discrete deduction steps of the $\vdash$ relation.
- **Knowledge Encoding (*γ*):** The translation of propositions into formulas of the logical language.
- **Assimilation Cost (*μ*):** The computational cost of $\vdash$ deductions.
- **Accommodation Cost (*η*):** Related to the modification of the axioms when true propositions that are not theorems are requested to be added as theorems of the theory.

3.2 Epistemic Incompleteness

Reflexivity is a fundamental phenomenon of logic and computation. It occurs when there is a function *f* that maps a proper subset of *A* into the whole set *A* (expansive reflexivity), or when a set *A* is mapped into a proper subset (contractive reflexivity). If *A* is a finite set, then this is impossible; therefore, reflexivity is strictly related to infinity and is an essential notion in the foundations of mathematics [30, 31]. There are many forms of reflexivity. For example, recurrence is a case of reflexivity, where a new value of a function is defined in terms of previously defined values. Reflexive patterns constitute the logical schema of many paradoxes (Liar paradox, Russell's paradox, Richard's paradox, …) from which fundamental logical theories stemmed [32].

According to the definition of a cognitive system Γ, we have $\mathrm{Knl}_t \subseteq \mathrm{Int}_t$. Now we show that there are internal states that are not encoded by any knowledge state: these states are *inaccessible* (to the knowledge of Γ). This result is related to a famous incompleteness of mathematical logic, as will be shown below.

Theorem 3.7 (Epistemic Inaccessibility/Incompleteness in Cognitive Systems). In any cognitive system Γ, at any time $t \in T$, there are internal states $q \in \mathrm{Knl}_t$ that are inaccessible:

$$\neg\exists x \in \mathrm{Knl}_t \, (\gamma_t(x) = q) \tag{8}$$

Proof. Let $E = \gamma^{-1}(\mathrm{Ext})$ and $I = \gamma^{-1}(\mathrm{Int})$ (the temporal parameter t is omitted, meaning that the conditions hold for any value of t). Of course, the sets E and I are subsets of Knl, with Knl = $I \cup E$ and $I \cap E = \emptyset$ (see Equation 7).

Figure 1 and Figure 2 illustrate the encoding γ and the partition of its counterimages.

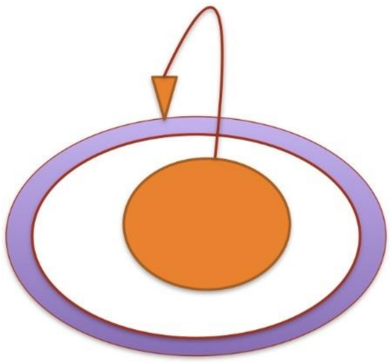

Figure 1 A diagram that illustrates the reflexivity of a cognitive system. The internal circle represents the knowledge states, the white oval is the set of internal states, while the border around it is the set of external states. The arrow is the encoding of knowledge states.

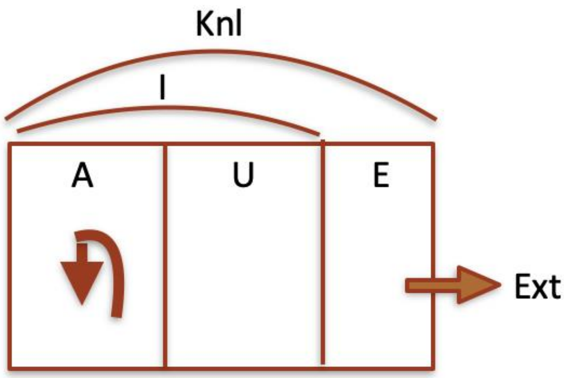

Figure 2 The structure of the knowledge states of Γ. The subset E encodes external states. The subset I encodes internal states. The subset A is the autoreferential set which encodes states of A, with $I = A \cup U$, Knl = $I \cup E$, and Knl ⊆ Int.

Let us consider an *autoreferential* maximal subset A of I such that:

$$A = \{x \in I \mid x \in A \Longrightarrow \gamma(x) \in A\} \tag{9}$$

Three possible alternatives hold:

Case 1 ($A = I$):

In the case $A = I$, any element of E cannot be an image of some element of I, because the images of I are in I. At the same time, any element of E cannot be an image of some element of E, because all the images of E are in Ext; therefore, all the elements of E are inaccessible states.

Case 2 ($A \subset I$):

In the case $A \subset I$, all the elements of $U = I \setminus A$ are inaccessible, because any element of U cannot be an image of some element of U, as the elements of U have images in A; it cannot be an image of some element of A, because the elements of A have images in A; and it cannot be an image of some element of E, because the elements of E have images in Ext.

Case 3 ($A = \emptyset$):

In the case $A = \emptyset$, all the elements of I are inaccessible, because no element of I is an image of some element of I, and it cannot be an image of some element of E because all the images of E are in Ext.

Therefore, in all three cases, there are elements of Knl that are also internal states and inaccessible.

From very general requirements on cognitive systems, it follows that inaccessible internal states must exist. Briefly, we can synthesize the previous theorem by saying that *knowledge cannot completely know itself*. A cognitive system has to include an internal locus of "ignorance" that is necessary for keeping a coherent self/non-self distinction.

We note that, in the setting of our definition of cognitive systems, the "knowledge" of a state does not coincide with its mere membership in Knl, but with the possibility of being a γ-image of some knowledge state that encodes it. Namely, a state $q \in$ Knl can express the knowledge of an internal state s because $s = \gamma(q)$, but the knowledge of q requires a knowledge state p (different from q) encoding q. In other words, knowledge is a continuous process of representation iterated at many representational levels.

In the following, we show the strong link between cognitive inaccessibility and logical incompleteness.

The *incompleteness theorems* of mathematical logic lie at the very origin of computer science. Logical calculi over formulas define a deduction relation $\vdash$ between a set of premises Φ and a conclusion φ. We write $\Phi \vdash \varphi$ to indicate that φ is syntactically derived from Φ within a given formal system. Similarly, the semantic consequence relation, denoted by $\Phi \vDash \varphi$, asserts that φ holds in every interpretation wherein all formulas in Φ are true.

For first-order predicate logic [33, 34], Kurt Gödel proved first-order logic completeness [35], establishing that standard predicative logical deduction can capture all semantically valid formulas (i.e., those true under all interpretations). However, Gödel soon discovered [35] that for any consistent axiomatic theory capable of expressing basic arithmetic, there exist formal statements p such that neither p nor $\neg p$ can be derived from the axioms—a result known as Gödel's First Incompleteness Theorem. First-order Peano Arithmetic (PA) is a primary example of such an incomplete theory.

In his 1931 landmark paper, Gödel [35] introduced the *arithmetization of syntax* (Gödel numbering) to represent arithmetic formulas as numbers within the axiomatic framework of PA. His celebrated theorem shows that any consistent formal system encompassing Peano Arithmetic is incomplete: under the standard interpretation, there are true arithmetic propositions that cannot be derived as theorems. In PA, each proposition p is assigned a Gödel number $[p]$. Gödel constructed a formal provability predicate $D(x)$ such that $D([p])$ holds in PA if and only if p is provable in PA. This reflects a profound form of self-reference, wherein arithmetic relations over numbers encode meta-theoretical properties of arithmetic formulas.

Gödel's original proof constructed a self-referential sentence **A** in PA asserting its own unprovability—effectively translating the ancient Liar Paradox (dating back to Epimenides, 6th Century BCE) into formal arithmetic. Let $\nvdash$ denote non-derivability, and assume PA is consistent (meaning that for no formula φ do we have both PA $\vdash \varphi$ and PA $\vdash \neg\varphi$).

If **A** were provable in PA (PA ⊢ **A**), then **A** would be true; yet **A** asserts its own unprovability, implying PA ⊬ **A**. Conversely, if ¬**A** were provable (PA ⊢ ¬**A**), then ¬**A** would be true, meaning **A** is provable (PA ⊢ **A**), which contradicts consistency and implies PA ⊬ ¬**A**. Thus, we deduce:

$$\mathrm{PA} \vdash \mathbf{A} \Rightarrow \mathrm{PA} \nvdash \mathbf{A} \tag{10}$$

and

$$\mathrm{PA} \vdash \neg\mathbf{A} \Rightarrow \mathrm{PA} \nvdash \neg\mathbf{A} \tag{11}$$

Consequently, neither **A** nor ¬**A** is provable in PA. However, under the standard model of arithmetic, one of these two statements must be true. Therefore, there exists a true formula in PA that cannot be derived within PA.

Because Gödel's original syntactic construction of **A** is highly technical, we present an alternative proof linking logical incompleteness directly to Turing computability.

Theorem 3.8 (Gödel's Logical Incompleteness via Turing Undecidability). In Peano Arithmetic PA, there exist true arithmetic propositions that cannot be derived within PA.

Proof. Every recursively enumerable set $A \subseteq \mathbf{N}$ (**N** is the set of natural numbers) can be represented in PA by formalizing the computation of a Turing machine that generates A [2]. This implies the existence of an arithmetic formula $P_A(x)$ such that PA ⊢ $P_A(n)$ if and only if $n \in A$. Using a diagonal construction, Alan Turing [36] defined a recursively enumerable set K of natural numbers that is undecidable: no Turing machine can determine in a finite number of steps whether an arbitrary integer belongs to K.

Thus, there exists an integer a such that membership $a \in K$ is algorithmically undecidable. Consequently, neither $D([P_K(a)])$ nor $\neg D([P_K(a)])$ can be derived in PA; otherwise, executing a systematic search over PA derivations via Turing machine simulation (utilizing Gödel arithmetization) would yield a decision procedure for membership in K. Since at least one of these two formulas is true in the standard interpretation of PA, there must exist true arithmetic propositions that are unprovable in PA.

We now provide an alternative proof of this theorem using the axiomatic definition of a cognitive system and the Epistemic Inaccessibility Theorem.

Theorem 3.9 (Logical Incompleteness via Epistemic Inaccessibility). The semantic logical consequences of Peano Arithmetic PA do not coincide with the set of formulas syntactically derivable from PA.

Proof. We formalize Peano Arithmetic as a cognitive system [PA]. Under this mapping:

- **Internal States (Int):** The set of true propositions of PA (the semantic logical consequences of the PA axioms).
- **External States (Ext):** The set of false propositions of PA.
- **Knowledge States (Knl):** The set of derivable formulas, characterized by the Gödel provability predicate D. Specifically, $D([p])$ encodes the true proposition p, whereas $\neg D([p])$ encodes the false proposition p.
- **Knowledge Encoding (*γ*):** Gödel's provability predicate D acts as the knowledge encoding mapping.
- **Time (*T*):** Discrete deduction steps within the formal calculus of PA.

- **Assimilation and Accommodation Costs:** The assimilation cost μ corresponds to the computational effort required to execute formal derivations under PA. The accommodation cost η is zero, as the underlying axiomatic base of PA remains static during deduction.

By the Epistemic Inaccessibility Theorem, the knowledge states of [PA] cannot cover all internal states of [PA]. Therefore, the set of true arithmetic formulas in PA cannot coincide with the set of syntactically derivable formulas in PA.

4. Conclusions

The conceptualization of a cognitive system proposed in this work provides a general, substrate-independent axiomatic framework. In real-world biological and artificial architectures, cognitive systems exhibit complex internal structures articulated into numerous functional modules organized across multiple representational levels [20].

Knowledge states and their encoding constitute the core of this architecture. Naturally, the notion of a *knowledge level* corresponds to the iterative application of the encoding function γ. Higher levels of knowledge enhance relational integration and abstract data processing. Furthermore, in non-trivial systems, γ is not a monolithic mapping, but rather a family of specialized encoding functions operating dynamically across different system modules and structural levels.

In Artificial Neural Networks (ANNs), the primary mechanism for knowledge encoding rests on high-dimensional embedding vectors. These representations lie at the heart of Large Language Models (LLMs) and conversational agents, which process semantic meanings through vector geometry [37, 38]. Transformer architectures demonstrate that comprehension can be formulated through dynamically generated values in high-dimensional Hilbert spaces, where embedding vectors serve as the functional encoding of words, phrases, and concepts.

Crucially, the functional modules responsible for generating these embeddings can themselves be encoded as vectors. By treating processing modules as data at higher representational levels, cognitive systems acquire an intrinsically reflexive character—analogous to the arithmetization of syntax in formal logic. Integrating this multi-level reflexive encoding with the transformer vector framework poses a central challenge for future AI research, particularly in identifying and embedding modular functional units.

Within this framework, meanings are not static values stored in discrete memory locations; rather, they are dynamic trajectories within abstract geometric spaces that map data and their functional relations. Because logical operations underlying natural language are functional transformations [39], they can be realized by discrete functional modules. Consequently, suitably extended LLM architectures could, in principle, acquire the capacity to encode and "understand" the very computational mechanisms driving their own reasoning. Encoding functional modules is thus a key prerequisite for advancing reflexivity in both artificial intelligence and theoretical neurophysiology.

This perspective necessitates novel training paradigms. Standard training by example remains fundamentally limited. In contrast, *training by reasoning* aims to induce a cognitive system to structure its internal knowledge into a multi-level transformer representation through dialogue and natural language interaction. The semantic spaces underlying these hierarchical knowledge levels expand the system's dimensionality via a nested "Chinese boxes" structure, where a single coordinate in a high-level space encapsulates an entire lower-level semantic space.

Our axiomatic definition resonates with established threads in machine learning and cognitive science. The principle that knowledge is encoded in distributed representations has long been a foundational tenet of connectionism [8]; our framework provides a formal bridge between these internal representations and epistemic information. Moreover, the *manifold hypothesis*—which posits that neural networks learn low-dimensional data manifolds embedded in high-dimensional spaces [40] —aligns directly with our formulation of knowledge. Feature vectors lie on these manifolds, while acts of accommodation correspond to the geometric deformation and refinement of the manifolds to better fit incoming data [41].

Ultimately, this model offers a unifying bridge between abstract cognitive principles and the mechanics of modern AI. AI provides a unique empirical testing ground for formal theories of cognition, while general cognitive frameworks can suggest novel architectures for artificial systems.

Future research will focus on two major directions:

1. **Hierarchical Extension of Transformers:** Extending transformer architectures toward explicit multi-level encodings of functional modules.
2. **Logical Reasoning and Formal Internalization:** Overcoming the current limitations of LLMs in executing multi-step logical deductions and long implication chains [38, 42, 43]. As shown in previous sections, ANNs and axiomatic theories represent different realizations of the same underlying cognitive paradigm. Internalizing formal logical theories within neural architectures offers a promising path toward overcoming these boundaries in reasoning [39, 44]. In this context, the reflexive mechanisms responsible for logical incompleteness also serve as the engine for self-applicability and autonomous cognitive evolution.

Finally, while this work has focused strictly on cognition, cognition constitutes only one dimension of a broader psychological framework. A complete model of an autonomous agent requires the integration of three interacting pillars: *cognition*, *emotion*, and *volition*. Emotion provides the dynamic drive (the engine) behind behavior; volition defines goals and intentionality; and cognition supplies the representational tools and methods to achieve those goals. Mathematical and computational modeling of the synthesis among these three components remains an open frontier for the future of cognitive science and artificial intelligence.

Author Contributions

The author did all the research work of this study.

Competing Interests

The author has declared that no competing interests exist.

AI-Assisted Technologies Statement

Artificial intelligence (AI) tools were used solely for basic grammar correction and language refinement in the preparation of this manuscript. Specifically, Gemini 2.5 Flash was employed to improve the readability and linguistic clarity of the English text. All scientific content, data interpretation, and conclusions were developed independently by the author. The author has thoroughly reviewed and edited the AI-assisted text to ensure its accuracy and accepts full responsibility for the content of the manuscript.

References

1. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys. 1943; 5: 115-133.
2. Minsky ML. Computation: Finite and infinite machines. Upper Saddle River, NJ: Prentice-Hall, Inc.; 1967.
3. Brown RE. Donald O. Hebb and the organization of behavior: 17 years in the writing. Mol Brain. 2020; 13: 55.
4. Hebb DO. The organization of behavior: A neuropsychological theory. New York, NY: John Wiley & Sons, Inc.; 1949.
5. Rosenblatt F. The perceptron: A probabilistic model for information storage and organization in the brain. Psychol Rev. 1958; 65: 386-408.
6. Hopfield JJ. Neural networks and physical systems with emergent collective computational abilities. Proc Natl Acad Sci. 1982; 79: 2554-2558.
7. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986; 323: 533-536.
8. Rumelhart DE, McClelland JL, PDP Research Group. Parallel distributed processing, Volume 1: Explorations in the microstructure of cognition: Foundations. Cambridge, MA: The MIT Press; 1986.
9. Werbos P. Beyond regression: New tools for prediction and analysis in the behavior science. Cambridge, MA: Harvard University; 1974.
10. Werbos PJ. Backpropagation through time: What it does and how to do it. Proc IEEE. 1990; 78: 1550-1560.
11. Hinton GE. Implementing semantic networks in parallel hardware. In: Parallel models of associative memory. New York, NY: Psychology Press; 1989. pp. 191-217. Available from: https://www.taylorfrancis.com/chapters/edit/10.4324/9781315807997-13/implementing-semantic-networks-parallel-hardware-geoffrey-hinton.
12. LeCun Y. A learning scheme for asymmetric threshold networks [Une procedure d'apprentissage ponr reseau a seuil asymetrique] (In French). In: Proceedings of Cognitiva 85, Paris. San Francisco, CA: Academia; 1985. pp. 599-604. Available from: https://www.academia.edu/3073721/Une_procedure_dapprentissage_pour_reseau_a_seuil_asymmetrique_a_Learning_Scheme_for_Asymmetric_Threshold_Networks.
13. Cybenko G. Approximation by superpositions of a sigmoidal function. Math Control Signal Syst. 1989; 2: 303-314.
14. Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators. Neural Netw. 1989; 2: 359-366.
15. Mitchell TM. Machine learning. New York, NY: McGraw-Hill; 1997.
16. Nielsen MA. Neural networks and deep learning [Internet]. 2015. Available from: http://neuralnetworksanddeeplearning.com/.
17. Goodfellow I, Bengio Y, Courville A, Bengio Y. Deep learning. Cambridge, UK: MIT Press; 2016.
18. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate [Internet]. New York, NY: arXiv; 2014. Available from: https://arxiv.org/abs/1409.0473v1.
19. Manca V. Artificial neural network learning, attention, and memory. Information. 2024; 15: 387.

20. Manca V. On the functional nature of cognitive systems. Information. 2024; 15: 807.
21. OpenAI. GPT-4 Technical Report: arXiv:submit/4812508 [cs.CL] [Internet]. San Francisco, CA: OpenAI; 2023. Available from: https://cdn.openai.com/papers/gpt-4.pdf.
22. Turing AM. Computing machinery and intelligence. Mind. 1950; 59: 433-460.
23. Shannon CE. A mathematical theory of communications. Bell Syst Tech J. 1948; 27: 379-423.
24. Piaget J. The formation of the symbol in the child: Imitation, play and dream, image and representation [La formation du symbole chez l'enfant: Imitation, jeu et rêve, image et representation] (In French). Montrouge, France: Delachaux et Niestlé; 1994. Available from: https://excerpts.numilog.com/books/9791036911170.pdf.
25. Piaget J, Inhelder B. The child's representation of space [La représentation de l'espace chez l'enfant] (In French). Paris, France: Presses Universitaires de France; 1948.
26. Piaget J. Genetic epistemology [L'epistemologie génétique] (In French). Paris, France: Presses Universitaires de France; 1970. Available from: https://www.abebooks.com/first-edition/LEPISTEMOLOGIE-GENETIQUE-PIAGET-Jean-Presses-Universitaires/3705828922/bd.
27. Piaget J, Szeminska A. The development of the number concept in the child [La genèse du nombre chez l'enfant] (In French). Montrouge, France: Delachaux et Niestlé; 1941.
28. Feller W. An introduction to probability theory and its applications. New York, NY: John Wiley & Sons, Inc.; 1968.
29. Cover TM, Thomas JA. Elements of information theory. New York, NY: John Wiley & Sons, Inc.; 1991.
30. De Giorgi E. Selected Papers. Berlin and Heidelberg, Germany: Springer; 2006.
31. Manca V. Reflexivity and duplicability in set theory. Mathematics. 2025; 13: 678.
32. Russell B, Whitehead AN. Principia Mathematica. Cambridge, UK: Cambridge University Press; 1910.
33. Hilbert D, Ackermann W. Principles of mathematical logic. New York, NY: Chelsea Publishing; 1928. Available from: https://archive.org/details/bwb_S0-AWJ-083/page/n6/mode/1up.
34. Church A. Introduction to mathematical logic. Princeton, NJ: Princeton University Press; 1956.
35. Feferman S, Dawson JW, Kleene SC, Moore G, Solovay R, Heijenoort J. Kurt Gödel: Collected Works: Volume I: Publications 1929-1936. Oxford, UK: Oxford University Press; 1986.
36. Turing AM. On computable numbers, with an application to the Entscheidungsproblem. Proc Lond Math Soc. 1937; s2-42; 230-265.
37. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Adv Neural Inf Process Syst. 2020; 33: 1877-1901.
38. Chen L, Peng B, Wu H. Theoretical limitations of multi-layer transformer [Internet]. New York, NY: arXiv; 2024. Available from: https://arxiv.org/abs/2412.02975.
39. Manca V. Functional language logic. Electronics. 2025; 14: 460.
40. Fefferman C, Mitter S, Narayanan H. Testing the manifold hypothesis [Internet]. New York, NY: arXiv; 2013. Available from: https://arxiv.org/abs/1310.0425v1.
41. Oudeyer PY, Kaplan F, Hafner VV. Intrinsic motivation systems for autonomous mental development. IEEE Trans Evol Comput. 2007; 11: 265-286.
42. Dziri N, Lu X, Sclar M, Li XL, Jiang L, Lin BY, et al. Faith and fate: Limits of transformers on compositionality. Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023); 2023 December 10-16; New Orleans, LA, USA. San Diego, CA: Neural

Information Processing Systems Foundation, Inc. Available from: https://www.proceedings.com/075280-3081.html.
43. Kaplan J, McCandlish S, Henighan T, Brown TB, Chess B, Child R, et al. Scaling laws for neural language models [Internet]. New York, NY: arXiv; 2020. Available from: https://arxiv.org/abs/2001.08361.
44. Manca V. Agile logical semantics for natural languages. Information. 2024; 15: 64.