

Article

Not peer-reviewed version

Knowledge and Information in Epistemic Dynamics

[Vincenzo Manca](#)*

Posted Date: 5 February 2026

doi: 10.20944/preprints202510.1379.v2

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



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a [Creative Commons CC BY 4.0 license](#), which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Knowledge and Information in Epistemic Dynamics

Vincenzo Manca

University of Verona - Italy; vincenzo.manca@univr.it

Abstract

The paper proposes a general theory of cognitive systems that inverts the conventional relationship between information and knowledge. While classical approaches define knowledge as the result of processing information, we posit that knowledge is a primitive concept, and information is a consequence of the process of assimilating knowledge. A general definition of a cognitive system is given, and a corresponding measure of epistemic information is defined such that Shannon's information quantity corresponds to a particular simple case of epistemic information. This perspective enables us to demonstrate the necessity of internal states of a cognitive system that are not accessible to the knowledge, by connecting cognitive systems to formal theories and showing a strong relationship with classical incompleteness results of mathematical logic. The notion of epistemic levels highlights a rigorous setting for clear distinctions among concepts such as learning, meaning, understanding, consciousness, and intelligence. The role of AI in developing deeper and more accurate models of cognition is argued, which in turn could suggest new relevant theories and architectures in the development of artificial intelligence agents.

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 (ANN) are the epicenter of artificial intelligence. They were introduced in the seminal paper [27], where neurons contain discrete values, and synapses connecting neurons are equipped with functions transforming the values of afferent neurons into the values of efferent ones [29]. An important shift in perspective emerged after the seminal book by Donald Hebb [3,14], altering the functional perspective to a dual vision, where synaptic plasticity forms the basis of learning processes. Namely, in the following seminal works [16,17,21,37–39,44,45], neurons became places containing functions, and the synapses expressed arrows labeled with real values, called **weights**, denoting connection strengths. Under very general assumptions, neural networks become, in this form, universal approximators of all continuous functions between hyperspaces of real numbers [8,18]. Given that many behavioral competencies are representable in such functions, the role of neural networks is a crucial basis for any computational model of complex human abilities.

Synapse modification through training by examples was modeled by Machine Learning algorithms, which update weights to improve the behaviors of neural networks toward the acquisition of functions related to specific competencies. [16,17,21,28,30,37,38,44,45]. In this way, artificial neural networks became a model inverting Turing's paradigm. Namely, Turing machines and equivalent formalisms are *programmed* to compute functions. On the contrary, ANNs equipped with machine learning methods are *trained* by pairs of values (input, output) from which they discover the function underlying the received pairs. The main idea in this field was the back-propagation method, rooted in the theory of function optimization, which provides the right weights to ensure a desired functional behavior of a given neural network. According to this algorithm, errors committed, with respect to the expected output, are sent in the opposite direction that the computation goes along the network. In such a way, a criterion of weight updating determines error decreases by reaching, at the end of the training phase, the right weights associated with the expected behavior [1,2,23,25].

Recently, ANN technology [31] realized Turing’s visionary hypothesis [43] of a talking machine through networks with dimensions of neurons and synapses comparable to those of the human brain, producing chatbots, machines generating conversations (in many natural languages) of a level comparable, in many aspects, to that of typical human conversation. From this perspective, the crucial role of language emerged as an essential aspect in the construction of cognitive processes. Based on this ability, neural networks trained by machine learning methods can solve many complex problems and have become the starting point of a new era in human technology.

The paper explores the new perspectives opened by LLM transformer models, but in a direction that is opposite to the standard one. What do transformer models tell us about cognition? Can we define general notions of cognitive systems that shed new light on intelligence, suggesting new models, and giving rigorous definitions of fundamental psychological concepts? These findings would promote reciprocal advancements in neuropsychology and artificial intelligence.

2. Probabilistic versus Epistemic Information

In his famous booklet [41], Claude Shannon outlined the mathematical theory of information. He proposed a measure based on probability. Given a stochastic variable X assuming values in a set A with probabilities $\{p_a\}_{a \in A}$, the information quantity of a value a is given by

$$-\log(p_a)$$

where $\log$ is the logarithm in base 2. The intuitive idea is that the information content is inversely proportional to probability, and is additive for joint and independent occurrences of values (with a probability that is the product of the single probabilities). Then it is natural to use logarithms, which change products into sums:

$$Inf(a,b) = -\log(p_{(a,b)}) = -\log(p_a \times p_b) = -\log(p_a) - \log(p_b) = Inf(a) + Inf(b).$$

Just at the beginning of his paper, Shannon links the measure of information to the measure of uncertainty, because the two concepts are two faces of the same coin. Namely, the *a priori* probability corresponds to the information we gain when the value occurs, and, at the same time, to the loss of uncertainty when the value has occurred, so that ignorance decreases in a measure that coincides with the gained information.

In this probabilistic framework, Shannon develops many important informational concepts that play crucial roles in communication. However, other important aspects are completely absent in Shannon’s perspective. Namely, the primitive notion of information suggests that something informative for a person is completely meaningless for another, or even something very informative in a given context may become meaningless in a different situation. The usual intuition perceives a strong connection between the information and the knowledge of the receiver of that information, and this knowledge changes continually as new information is received. Can we find an epistemic value of information and quantify it in relation to a knowledge state of a given communication? The interaction and interdependence between information and knowledge require a different setting, where knowledge is an active component establishing the values of data, rather than a result of their accumulation.

First, it is important to distinguish between data and information. Data are events containing information, but the information is independent of them. Namely, data can be encoded into other data without changing their information content. Therefore, data are elements that carry the information they “contain”. Of course, this is a circular description that does not clarify which elements are or are not data. In the probabilistic perspective, any stochastic variable is an information source, without any involvement of the agent receiving the data generated by the source. Conversely, the approach we will present inverts the usual perspective by assuming a knowledge entity whose inputs are considered data and whose elaboration provides new data, changing the knowledge entity. Encoding and decoding are the key concepts of an intrinsically functional perspective, where knowledge becomes the “internal”

encoding of data within a system, which, after this encoding, acquires new competencies, that is, functions, in data processing and interaction with the external world.

3. Cognitive Systems

In a series of papers and books [33–36] (among many others), Jean Piaget and his school of developmental psychology investigated the conceptual organization of mathematical and physical intuitions in children and adolescents during the learning processes of educational curricula.

A schema emerged, which could be expressed by the dichotomy *assimilation/accommodation*. Comprehension is very similar to food ingestion. An external concept enters the student's mind, but to be effective and productive, it needs the *assimilation* to the structure of the mind, in the state it is, and in a form that can be significant for the internal organization of the mind when the acquisition happens. In the case of food, it is reduced to minimal components by means of suitable enzymes, and these components provide the substances and energy necessary for organs, tissues, and cells to perform their functions. An analog mechanism acts in comprehension. Data of many possible formats (textual, visual, ...) when entered into a cognitive system need to be "encoded" in forms coherent with the internal structure of the system. This internal encoding process is called *assimilation*. But assimilation requires a reciprocal mechanism, called *accommodation*, according to which the cognitive system receiving the new data reacts with suitable adjustments of its structure for better assimilation, internal coherence, and adequacy.

This approach will be the main inspiration for the following definition.

3.1. A General Definition

Let S be a state space, for example, a Hilbert space of *features*. When in the state space S , we consider time, then we pass to an event-state space. This is a measurable space (S, M, p) , where M is a σ -algebra of subsets of S , and p is a probability measure [11], and an *event* is a pair (A, t) where $A \in M$ is a subset of states that may occur at time t (the notion of event occurrence may be further specified in many possible ways, according to parameters of localization/distribution in the considered state space).

A cognitive system is a dynamic structure on an event-state space. Intuitively, a cognitive system develops in time by interacting with its external world through several levels of representation of the reality it perceives. The information content of data coming from the outside is given by the costs that the internal representation requires during the assimilation and accommodation phases of this representation.

Representing objects, events, and states of affairs gains the possibility of discovering relations, making inferences, predictions, and decisions for fruitful interactions with the external world.

The essence of these abilities is the notion of mathematical function, going back to Euler's definition in *Introductio in Analysin Infinitorum* (1748), and formalized in by Alonzo Church in 1930 years in terms of λ -expressions, that is, a map sending objects of a set to images of them, in such a way that operating on images gives an advantage with respect to operating directly on the objects. A map of a city is a simple way to understand it. On a map, we can choose the best way to connect two places by exploiting the quick overview of a reduced and synthetic representation of streets. When symbols or symbolic structures map objects, functions are named codes. Typically, a code is a function from strings over a finite alphabet of symbols, *codewords*, into a set of encoded objects. Coding theory is a widely investigated subject, central in Information Theory [7]. We do not enter into specific details about codes, but want to stress their critical role in cognition.

At any time, a cognitive system is placed at a point (or a set of possible points) in a multidimensional space of states. Some of these states are knowledge states with representational roles.

Coding, possibly articulated in parts and levels, is the main ingredient of cognitive dynamics. Considering a cognitive system only in terms of states and codes is an "extensional" approach, because it prescind on the physical realization of states (molecular, biomolecular, cellular, circuitual, ...). In

the following, we adopt this viewpoint and will show that, although it is very general and abstract, it allows us to derive interesting properties.

The main intuition of the following definition of a cognitive system is based on internal states, external states, and knowledge states.

Definition 1. A cognitive system Γ has seven components:

$$\Gamma = (T, Int_t, Ext_t, Knl_t, \gamma_t, \mu, \eta)$$

Int_t, Ext_t, Knl_t , are non-empty sets of states of an assumed state space S , where t ranges in a set T denoting the lifetime of Γ (Γ is equipped by an internal clock, $t \in T$), and γ_t is an encoding function; μ and η are cost functions, of data encoding $\gamma_t(d)$, and of encoding updating in passing from γ_t to γ_{t+1} , respectively. The following conditions are required for Γ (the temporal parameter is omitted by intending that conditions hold for every time t):

Int_t and Ext_t are non-empty disjoint sets of the internal and external states of Γ ; Knl_t is the set of knowledge states, where $Knl_t \subseteq Int_t$;

γ is the knowledge encoding function:

$$\gamma : Knl \rightarrow Int \cup Ext \quad (1)$$

such that, for any argument x , the no fixpoint condition holds:

$$\gamma(x) \neq x \quad (2)$$

inf is a measure of information, the epistemic information of an input datum d . It is the sum of two components:

$$inf(d) = \mu(\gamma(d)) + \eta(\gamma(d)) \quad (3)$$

the μ component expresses the assimilation cost of d . The η component expresses the corresponding accommodation cost, which may require a reorganization of γ to reach some consistency and adequacy features of its global structure. $\square$

An internal state is *accessible* only if it is a γ -image of a knowledge state (this request holds even for knowledge states). In other words, accessibility requires, in any case, a true representational level, through a knowledge state that is different from the state it represents.

In a cognitive system, Knl may increase and enrich over time; the costs of data acquisition from the system may change dramatically during its lifetime, depending on the construction of its knowledge. Being Int and Ext disjoint, also their counterimages are disjoint, hence no confusion arises between the knowledge encodings of internal and external states, in symbols:

$$\gamma^{-1}(Int_t) \cap \gamma^{-1}(Ext_t) = \emptyset \quad (4)$$

External states are based on *input data* expressing the information coming to Γ from the outside. A part of the internal state is available to the outside, providing the *output data* of Γ .

Example 1 (Shannon's Information Source). Let us consider a basic cognitive system that takes in input a sequence of symbols over a finite alphabet and assimilates them internally by a knowledge encoding γ that constructs a table of binary encodings of symbols according to a Huffman code, where more probable symbols receive shorter representations [7]. The external states are input sequences to the system; the internal states are

tables of pairs (symbol occurrence, binary representation), and the knowledge states are the binary codewords of input symbols with their occurrence multiplicity. $\square$

If the assimilation cost is the length of the binary representation, and the accommodation cost is null, the epistemic information of symbol i results in $-\log p_i$ with p_i the frequency of i in the input sequence. This corresponds exactly to Shannon's notion of information quantity.

By classical information theory, we know that the average information of symbols in the sequence corresponds to Shannon's entropy, and no code can exist that has a shorter average information. This means that Shannon entropy is an inferior limit to the average Shannon's information quantity, and Huffman codes are optimal codes for this information measure. In other words, Shannon's information quantity is a measure of a basic cognitive system, where knowledge encoding minimizes the average length of codewords (and no assimilation costs are considered).

Shannon's framework is simple yet powerful enough to develop numerous informational concepts, making it very useful in many contexts. However, it is too simple to express the complex interaction of cognitive character, where the knowledge of a cognitive agent and its history of development over time play crucial roles in evaluating the informational content of data.

Example 2 (Artificial Neural Network). *An ANN is a cognitive system with the following components.*

Internal States: the weights, biases, and activation states of all neurons at a given time.

External States. The external states are the input vectors provided to the network, representing observations from the environment.

Knowledge States: a subset of the internal states, the part of the system's state that "encodes" past training data. In this case, the memory of knowledge is not simply a storage of copies, but a deeper structure of values that, in the context of network connections, synthesizes experience, providing the competencies acquired by the ANN during its training.

Knowledge Encoding: the function γ of the forward-pass computation of the network. It takes an input vector and maps it to a set of activated neurons, which represent its internal representation.

Assimilation Cost: the computational cost of the forward pass, or the cost of computing embedding vectors related to linguistic inputs in a LLM transformer model (in a chatbot case).

Accommodation Cost is where the model shines. It is the cost of the backpropagation algorithm and the subsequent weight updates. This is the computational effort required to adjust the knowledge states (weights) to minimize the error between its prediction and the new data point. A data point that is consistent with the network's current knowledge will have a low error and, therefore, a low accommodation cost. A data point that is a "surprise" or an "outlier" will have a high error, triggering a large weight adjustment and thus incurring a high accommodation cost. $\square$

ANN training is a continuous process of assimilation (forward-pass) and accommodation (back-propagation) that defines the information content of each training sample. The dual-cost model assimilation-accommodation captures the full complexity of learning as a continuous process and defines the information content of data.

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 an essential notion in the foundations of mathematics [9,26]. 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, Russel paradox, Richard's paradox ...) from which fundamental logical theories stemmed [40].

A *reflexion* over a set A is a function f from a proper subset B of A to A . The reflexion is *complete* if it is surjective, that is, if all elements of A are images of elements of B ; otherwise, it is *incomplete*. A set A is *reflexive*, or Dedekind infinite, when a 1-to-1 function exists between a proper subset B of A and A . Any infinite set is reflexive [26], and no finite set can be reflexive. A cognitive system, even if it is physically finite in any instant of its life, is potentially infinite when its states are defined on values that range over infinite sets.

According to the definition of a cognitive system Γ , we know that $KnI_t \subseteq Int_t$. Now we show that the encoding γ does not completely cover the internal states, that is, the set of γ -images is strictly included in the internal states. There are internal states that are not encoded by any knowledge state: these states are *inaccessible* (to the knowledge of Γ). This result resembles typical results in mathematical logic, and this emphasizes its general relevance. The relationship with classical results of mathematical logic will be discussed below.

Theorem 1. *Epistemic Incompleteness (Inaccessibility in Cognitive Systems)*

In any cognitive system Γ , the knowledge encoding γ is an incomplete reflexion:

$$\gamma(KnI_t) \subset Int_t$$

Proof. Let $E = \gamma^{-1}(Ext)$ and $I = \gamma^{-1}(Int)$ (time parameter t is omitted, because it does not impact the following reasoning). Of course, sets E and I are subsets of KnI , and $KnI = I \cup E$, $I \cap E = \emptyset$ (see Equation (4)).

Let us consider the following two alternatives.

1) $\gamma(I) = I$, that is, I is γ -autoreferential: for all $s \in I$ there exist $q \in I$ such that $\gamma(q) = s$. In this case, all the internal states of $Int - I$ (which is a non-empty set) are inaccessible, because they cannot be γ -images of knowledge states (the γ -images of E are external states).

2) if $\gamma(I) \subset I$, let A be the biggest proper subset of I that is γ -autoreferential. Then, the elements of the non-empty set $U = (I - A)$ are not γ -images of elements of U . Namely, any element $y_0 \in U$ cannot verify an equation $y_0 = \gamma(x_0)$ with $x_0 \in U$, otherwise A would not be the biggest autoreferential subset of I . In conclusion, the elements of U are not accessible, because they cannot be γ -images of U , nor γ -images of A (which is autoreferential), nor γ -images of E (which are external states). In conclusion, in both cases, surely in Γ there are internal states that are inaccessible (see Figure 1).

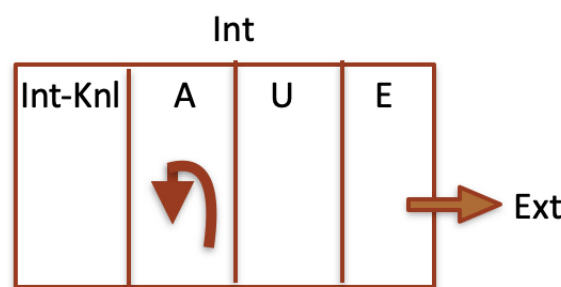


Figure 1. The structure of internal states of Γ . The subset E contains the encoding of external states. The subset A contains the autoreferential states, and states in $I = A \cup U$ encode internal states. Some of the subsets $Int - KnI$, U , E are not empty and contain γ -inaccessible states (states that are not γ -images of knowledge states).

It is surprising that from very general requirements on cognitive systems, it follows the existence of inaccessible internal states. In other words, we can conclude that a cognitive system has to include an internal part of “ignorance” that is necessary for keeping a coherent self-non-self distinction. We notice that, in the general setting of the previous definition of cognitive systems, the “knowledge” of a state does not coincide with its membership to KnI , but with the possibility of having a γ -image representing it. Namely, a state q of knI can express the knowledge of an internal state s , but the

knowledge of q requires an image of it (different from q) representing it. In other words, knowledge is a continuous process of representation that can be iterated at many representation levels.

The *incompleteness theorems* of mathematical logic are at the origin of computer science. Logical calculi over formulae define relations $\vdash$ holding between a set of formulae Φ (premisses) and a formula φ (conclusion). We write $\Phi \vdash \varphi$ for indicating that φ is deduced from Φ via some logical calculus. The logical consequence is an analogous relation, indicated by $\Phi \models \varphi$, indicating that in all interpretations where all formulae of Φ are true, also φ is true. In first-order logic, or predicate logic ([6,15]), mathematicians defined logical calculi that are complete (Gödel, 1930), in the sense that these calculi could deduce all the logically valid propositions (valid in all the interpretations) [6,12]. However, Gödel also soon discovered (1931) that, in predicate logic, for axiomatic theories reaching some logical complexity, there exist propositions p such that either p or $\neg p$ cannot be deduced from the axioms (first Gödel's incompleteness theorem [12]). The theory PA of Peano's Axioms, expressed in first-order logic, presents such an incompleteness.

Kurt Gödel [12] invented the *arithmetization of syntax* for encoding propositions of Peano's arithmetic by numbers, within the axiomatic system PA formulated in predicate logic. His celebrated theorem asserts that any theory including Peano Arithmetic is logically incomplete because, under any interpretation of the theory, there are true propositions that cannot be deduced as theorems of the theory. In PA for any proposition p there is a number $[p]$ encoding p . Gödel determined a formula $D(x)$ of PA such that $D([p])$ is true in PA if and only if p is deducible in PA . This is a case of reflexivity that we could say *autoreferentiality*, where numbers encode propositions on numbers. The original proof by Gödel was based on one of the most interesting cases of reflexivity. It was given by constructing in PA a formula $\mathbb{A}$ expressing, via the D predicate, its undeducibility in PA (a transformation of the Liar Paradox sentence, going back to Epimenides, VI Century A.C.). Let $\not\vdash$ be a symbol expressing the undeducibility. Let us assume that PA is coherent, that is, for no formula ϕ of PA , both conditions $PA \vdash \phi$ and $PA \vdash \neg\phi$ can hold.

If $\mathbb{A}$ were deducible in PA , then $PA \vdash \mathbb{A}$, that is, $\mathbb{A}$ would be true, so $\mathbb{A}$ would not be deducible, whence $PA \not\vdash \mathbb{A}$.

If $\neg\mathbb{A}$ were deducible, then $PA \vdash \neg\mathbb{A}$, then $\neg\mathbb{A}$ would be true, so it is not true that $\mathbb{A}$ is undeducible, therefore $PA \vdash \mathbb{A}$, which implies $PA \not\vdash \neg\mathbb{A}$. In conclusion:

$$PA \vdash \mathbb{A} \Rightarrow PA \not\vdash \mathbb{A}$$

and

$$PA \vdash \neg\mathbb{A} \Rightarrow PA \not\vdash \neg\mathbb{A}.$$

This means that both $\mathbb{A}$ and $\neg\mathbb{A}$ cannot be deducible, but one of them has to be surely true, then there exists a true formula of PA that is not deducible in PA .

The construction of $\mathbb{A}$ is very technical, so we prefer to present another proof that relates the logical incompleteness to Turing's notion of computability.

Theorem 2 (Gödel's logical incompleteness). (*via Turing's undecidability*)

In Peano Arithmetic PA , there are true arithmetic propositions that cannot be deduced within PA .

Proof. Any recursively enumerable set A of numbers [29] can be represented in PA (by expressing the computation of the Turing machine generating it). This means that there exists a formula P_A such that $P_A(n)$ is deduced if and only if $n \in A$. Moreover, Alan Turing [42] defined (using a diagonal construction due to Cantor) a recursively enumerable set K of numbers that is not decidable, because no Turing machine exists that can always establish, in general, in a finite number of steps, whether a given element belongs to it. Therefore, there exists a number a such that it is impossible to decide in a finite number of computation steps whether $a \in K$ or $a \notin K$. This means that both $D([P_K(a)])$ and $\neg D([P_K(a)])$ are not deducible in PA , otherwise simulating PA deductions with a Turing machine computations (using Gödel arithmetization) we could decide the membership of a to K . Of course, at

least one of the two formulae above is surely true in PA . Therefore, there are true propositions of PA that cannot be deduced in PA . $\square$

We now provide a proof of the theorem above using the definition of a cognitive system, which demonstrates the power of such a general definition.

Theorem 3 (Logical incompleteness via Cognitive Systems). *The logical consequences of Peano's axioms PA do not coincide with the formulas deducible from PA .*

Proof. The proof follows from the possibility of defining PA as a cognitive system. True formulae of PA (the logical consequences of PA axioms) are its internal states, while false formulae of PA are its external states. Knowledge states of PA are the deduction formulae of PA , expressed by the deduction predicate D . The formula $D([p])$ encodes the true formula p , while $\neg D([p])$ encodes the false formula p .

From the theorem of epistemic incompleteness, we can deduce that knowledge states cannot encode all the internal states. In PA as a cognitive system, this implies that the true formulae cannot coincide with the deducible formulae of PA , because internal states of PA do not coincide with those represented by knowledge states of PA .

We mention that in PA as a cognitive system, the time could be that of the theory extension steps (see later on); the assimilation cost could be some measure of the complexity of deductions; and the accommodation cost could be that of reorganizing the theory (adding new axioms) to prove a true proposition that cannot be deduced.

Figure 2 can be applied to a formal theory, in accordance with its interpretation in terms of a cognitive system. The internal circle represents deducibility formulas, the white circle true formulas, and the external border false propositions.

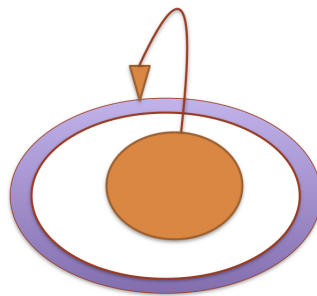


Figure 2. A diagram that illustrates the reflexivity of a cognitive system. The internal circle is the knowledge states, the white oval is the set of internal states, while the border around it is the set of external states. The arrow is the encoding of knowledge states.

3.3. Epistemic Levels

The given notion of a cognitive system is very general. In real cases of such systems, we must consider their internal structure, which is often articulated in a vast number of functional modules, organized at multiple levels [25]. The knowledge states and their encoding are surely a crucial aspect of the cognitive architecture. It is natural to assume a notion of *knowledge level* as the number of iterations of the encoding γ given in our definition. Namely, an internal state s is encoded by an internal state $\gamma(s)$, but this state can be encoded at a higher level of representation by the state $\gamma(\gamma(s))$. Higher knowledge levels enhance the connection and elaboration of data processing. Moreover, it is also reasonable that γ is not a simple function, but it is the union of many encoding functions, which act in dependence on the parts and on the levels of the states to which they apply.

In ANN, the basic mechanism of knowledge encoding is based on the notion of embedding vectors. They are the core of the success of LLM neural networks, which realize chatbots by elaborating meanings in terms of numerical vectors.

Transformer LLM models demonstrate [4,5] that comprehension is expressed through the values dynamically generated during data processing, which essentially consist of embedding vectors in multidimensional Hilbert spaces.

The embedding vectors implicitly ensure an adequacy to distributional criteria of the kind outlined before, as much as they encode the right way words combine with the other words in a way that is coherent with the distributional profiles they have in the corpus on which the neural network was trained for acquiring its conversational competence.

Embedding vectors are the knowledge encoding of words, phrases, and concepts. They are the "objects" of knowledge, and are obtained through specific functional modules. In turn, these modules can also be encoded by real vectors analogous to embedding vectors. In this way, they become data at further knowledge levels. Thus, cognition is intrinsically reflexive, in a way analogous to arithmetic and all fundamental mathematical theories.

The further knowledge levels encode data into other semantic spaces, adding internal comprehension to more complex meanings.

However, meanings of a cognitive system are not static values memorized in some locations, but rather processes of localization in abstract spaces that geometrize data and their relations in corresponding "visualizations" within those spaces.

The three main levels of a cognitive system are illustrated in Figure 3: the *operative level* of competencies; the *coordinative level*, where the elements of the operative level and their relations are represented and integrated as different parts of a global competence; the *directive level*, where the system elaborates tasks and strategies, possibly organized in a hierarchy of finalities. The most superior level encodes a representation of itself, that is, a full consciousness. The intelligence of the system emerges as the ability to coordinate and integrate the functionalities of all cognitive levels, by finding the best representations, and by finalizing its behavior to the situations of its life. In the superior levels, a cognitive system needs to incorporate theories or similar structures.

The system of Figure 3 consists of functional modules. Arrows are channels sending feature vectors, represented by full squares. Some feature vectors encode functional modules. In this way, a cognitive system can get knowledge of its own functionalities.

The encoding mechanisms and knowledge levels are not prefixed according to an external project elaborated by a designer; rather, they emerge following an internal principle that propagates and generalizes the transformer approach toward a multilevel perspective. Implementations of this generalization in artificial systems remain an open problem for future research. However, the key point toward multilevel transformers is the possibility of recognizing modules as functional units and encoding them through suitable feature vectors. Figure 4 shows a possible way to encode a functional level of an inferior level by an expression of weights and bias, which can be translated into a feature vector.

In the middle level, embedding vectors give meaning to words and discourses. Therefore, feature vectors exploit embedding vector semantics to describe functional modules and introduce reflexive knowledge about themselves.

Logical operations underlying the logic of natural language are functions [24]; therefore, they are realized by functional modules. This means that chatbots, suitably extended, could, in principle, be able to "understand" the mechanisms on which their comprehension is based. In other words, the encoding of functional modules is a crucial step to increase the reflexivity of cognitive systems. Understanding how to realize the encoding of functional modules is a key research topic for both artificial intelligence and theoretical neurophysiology. Figure 5 gives another view of a neural network by distinguishing the main activities and their connection to the cognitive component, where the knowledge representation is realized by specific modules.

This scenario requires new levels of training. Training by example is too limited. *Training by reasoning* is the method that induces a cognitive system to organize its structures toward a multilevel transformer representation of its internal knowledge. It has to be based on natural language and

developed through conversational activities. The semantic spaces that underlie the different knowledge levels add new dimensions according to a *Chinese boxes* mechanism where a single coordinate may refer to a point encapsulating the features of some hidden semantic space.

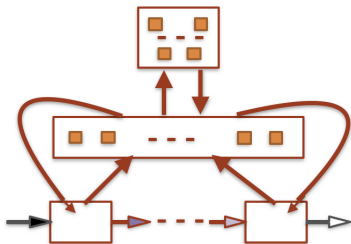


Figure 3. A simplified cognitive architecture. At the bottom, the operative level, in the middle, the coordinative level, and at the top, the directive level. Full squares represent the encoding of functional modules of lower levels.

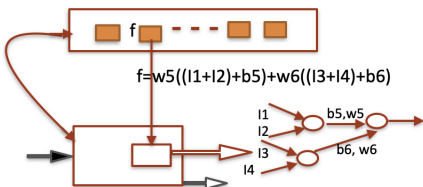


Figure 4. The encoding of a functional module as a feature vector at a superior knowledge level.

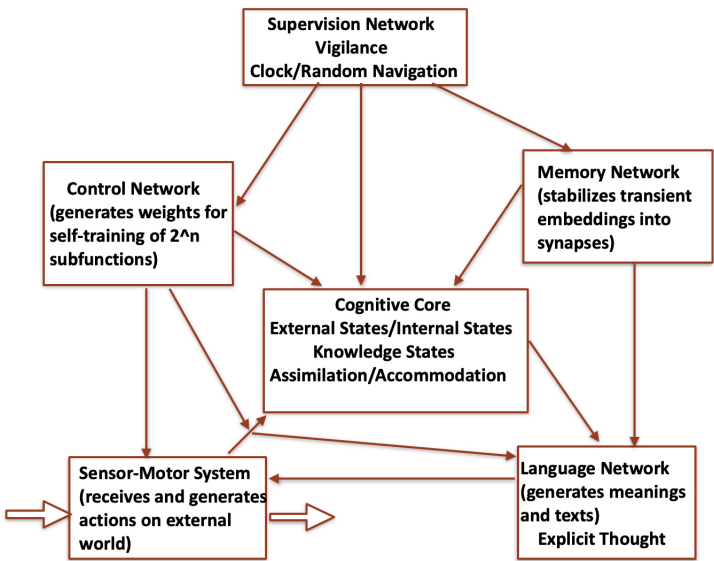


Figure 5. The global structure of an ANN that includes a cognitive core of knowledge representation according to the general definition given in section 3.

4. Conclusions

Our general definition of a cognitive system resonates with several lines of research in machine learning and cognitive science.

The idea that knowledge is encoded in distributed representations within neural networks has been a core tenet of connectionism [39]. Our framework provides a formal way to link this internal representation to information. Moreover, the “manifold hypothesis”, according to which neural networks learn to represent data on dimensional manifolds [13], is related to our idea of knowledge

as an internal representation. Embedding or feature vectors exist in these manifolds, and the acts of accommodation are the processes of deforming and refining them to fit better observed data.

The concept of "information gain" is often defined as the expected reduction in uncertainty that a data point can provide. Our framework reformulates this, where the "most informative" data are those with the highest accommodation cost, as they force the most significant and beneficial structural changes to a "knowledge gain". This aligns with approaches that seek to discover and learn from "surprising" inputs [32].

Our model offers a unifying framework to bridge the gap between abstract cognitive principles and the concrete mechanics of modern artificial intelligence. What is apparent from the discussion developed so far is that AI could be a unique occasion for suggesting general perspectives in cognition. On the other hand, these general viewpoints could suggest new models of AI based on the speculations emerging at a very general level.

Future research on cognitive systems intends to develop specific themes relevant to artificial intelligence and to neuropsychology. The most central theme is extending the transformer approach toward many levels of encoding.

Another important investigation topic is overcoming the present limitations of chatbots in dealing with complex logical deduction, in long implication chains involving abstract concepts [5,10,19]. In previous sections, it has been shown that artificial neural networks and formal theories are, in essence, different forms of cognitive systems. Then we could internalize formal theories into suitable neural networks to obtain more powerful cognitive systems, overcoming the present logical limitations [22,24].

Finally, we focused on cognition, but we know that it is only a part of a more complex psychological unit, where other components are present and interact, in the construction of a "person", with at least three different aspects: cognition, emotion, and volition. Emotion is the fuel of the "motor" providing the causes of the behavior; the volition provides the finalities, and the cognition provides the instruments and methods to realize the finalities. Mathematical analyses of the interaction and integration of these components can suggest new perspectives and new models for further comprehension and realization.

References

1. Bahdanau, D., Cho, K., Bengio, Y., Neural Machine Translation by Jointly Learning to Align and Translate, *arXiv:1409.0473*, **2014**
2. Goodfellow, I., Bengio, Y. Courville A., *Deep Learning*. MIT Press, **2016**
3. Brown, R.E., Donald O. Hebb and the Organization of Behavior: 17 years in the writing. *Mol. Brain* , 13, 55, **2020**
4. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Language Models are Few-Shot Learners. *NEURIPS*, 33, 1877–1901, **2020**
5. Chen, L., Peng, B., Wu, O. Theoretical limitations of multi-layer Transformer, arXiv:2412.02975v1 [cs.LG] 4 Dec., **2024**
6. Church, A., *Introduction to Mathematical Logic*, Princeton University Press, **1956**
7. Cover, T. M., Thomas, J. A., *Elements of Information Theory*, John Wiley & Sons, New York **1991**
8. Cybenko, G., Approximation by superpositions of a sigmoidal function". *Mathematics of Control, Signals, and Systems*. 2 (4): 303–314, **1989**
9. De Giorgi E., *Selected Papers*, Dal Maso G., Forti M., Miranda M., Spagnolo S. (eds.), Springer-Verlag, Berlin & Heidelberg, **2006**
10. Dziri, N. et al., Faith and Fate: Limits of Transformers on Compositionality, *Advances in Neural Information Processing Systems*, 36 (NeurIPS 2023), Ernest N. Morial Convention Center, New Orleans (LA), USA, 10-16 Dec., **2023**
11. Feller, W., *An introduction to probability theory and its applications*, John Wiley & Sons, New York, **1968**
12. Feferman, S., Dawson, J. W., Kleene, S. C., Moore, G. H., Soloway, R. M., Van Heijenoort, J., *Kurt Gödel Collected Works, Vo. I*, Oxford University Press, **1986**
13. Fefferman, C., Mitter, S., Narayanan, H., *Testing the Manifold Hypothesis*, arXiv:1310.0425v2 [math.ST], **2013**
14. Hebb, D. O. *Organization of Behaviour*; Wiley: New York, NY, USA, **1949**
15. Hilbert, D., Ackermann, W., *Principles of Mathematical Logic* (tr. from German, 1928), AMS Chelsea Publishing, **1991**

16. Hinton G. E., Implementing semantic networks in parallel hardware. In *Parallel Models of Associative Memory*; Hinton, G. E., Anderson, J.A., eds.; Lawrence Erlbaum Associates. 1981; pp. 191–217. Available online: <https://taylorfrancis.com/chapters/edit/10.4324/9781315807997-13/implementing-semantic-networks-parallel-hardware-geoffrey-hinton> (accessed on 1 December 2024).
17. Hopfield, J. J., Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. USA*, 79, 2554–2558, **1982**
18. Hornick, K., Stinchcombe, M., White, M., Multilayer feedforward networks are universal approximators, *Neural Networks*, 2, 359–366, **1989**
19. Kaplan, J. et al., Scaling Laws for Neural Language Models arXiv:2001.08361, **2020**
20. Lee, J, *Introduction to Topological Manifolds*, Springer, **2010**
21. Le Cun, Y., Une Procédure d'Apprentissage pour Réseau à Seuil Asymétrique. In *Cognitiva 85: À la Frontière de l'Intelligence Artificielle des Sciences de la Connaissance des Neurosciences*, Proceedings of Cognitiva 85, Paris, France, 1985; pp. 599–604. Available online: <https://www.academia.edu> (accessed on 1 December 2024).
22. Manca, V., Agile Logical Semantics for Natural Languages. *Information MDPI*, 15, 1, 64, **2024**
23. Manca, V., Artificial Neural Network Learning, Attention, and Memory, *Information MDPI*, 15, 7, 387, **2024**
24. Manca, V., Functional Language Logic, *Electronics MDPI*, 14, 3, 460, **2025**
25. Manca, V., On the functional nature of cognitive systems, *Information MDPI*, 15, 12, 807, **2024**
26. Manca, V., Reflexivity and Duplicability in Set Theory, *Mathematics MDPI*, 13, 4, 678, **2025**
27. McCulloch, W.; Pitts, W., A Logical Calculus of Ideas Immanent in Nervous Activity. *Bull. Math. Biophys.*, 5, 115–133, **1943**
28. Mitchell, T., *Machine Learning*, McGraw Hill, **1997**
29. Minsky, M., *Computation. Finite and Infinite Machines*; Prentice-Hall Inc.: Upper Saddle River, NJ, USA, **1967**
30. Nielsen, M., *Neural Networks and Deep Learning*. Online **2013**
31. OpenAI, GPT4-Technical Report, ArXiv: submit/4812508 [cs.CL] 27 Mar., **2023**
32. Oudeyer P-Y, Kaplan, F., Hafner, V. (2007) *Intrinsic Motivation Systems for Autonomous Mental Development*, *IEEE Transactions on Evolutionary Computation*, 11(2), 265–286, **2007**
33. Piaget, J., *La formation du symbole chez l'enfant*; Delachaux & Niestlé, Neuchatel-Paris, **1941**
34. Piaget, J., *La représentation de l'espace chez l'enfant*; Presses Universitaires de France: Paris, **1948**
35. Piaget, J., *L'epistemologie Génétique*; Presses Universitaires de France: Paris, **1970**
36. Piaget, J., Szeminska, A., *Le Genèse du nombre chez l'enfant*; Delachaux & Niestlé, Neuchatel-Paris, **1941**
37. Rosenblatt, F., The perceptron: A probabilistic model for information storage and organization in the brain. *Psychol. Rev.*, 65, 386–408, **1958**
38. Rumelhart, D., Hinton, G., Williams, R. *Learning representations by back-propagating errors*. *Nature* 323, 533–536, **1986**
39. Rumelhart, D. E., McClelland, J. L., (eds.) *Parallel distributed processing: explorations in the microstructure of cognition, vol. 1: foundations*, MIT Press, Cambridge, MA, United States **1986**
40. Russell, B., Whitehead, A. N., *Principia Mathematica*, Cambridge University Press, **1910-13**
41. Shannon, C. E. *A Mathematical Theory of Communication*, Bell System Technical Journal, **1948**
42. Turing, A. M., On Computable Numbers, with an Application to the Entscheidungsproblem. Proceedings of the London Mathematical Society 42 (1):230-265, **1936**
43. Turing, A. M., Computing Machinery and Intelligence, *Mind*, London, N. S. 59,433-460, **1950**
44. Werbos, P., Beyond Regression: New Tools for Prediction and Analysis in Behavior Sciences. PhD Thesis, Harvard University, Cambridge, MA, USA, **1974**
45. Werbos, P., Backpropagation Through Time: What It Does and How to Do It. *Proc. IEEE*, 78, 1550–1560, **1990**
46. Bohm, D. (1980). *Wholeness and the Implicate Order*. Routledge & Kegan Paul.
47. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391–407.
48. Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in Linguistic Analysis*, 1–32.
49. Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3), 146–162.
50. Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, 46(1-2), 159–216.
51. Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37, 141–188.
52. Baroni, M., & Lenci, A. (2010). Distributional Memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4), 673–721.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.