Commentary

No Consciousness? No Meaning (and no AGI)!

Marco Masi[1]

1. Independent researcher

The recent developments in artificial intelligence (AI), particularly in light of the impressive capabilities of transformer-based Large Language Models (LLMs), have reignited the discussion in cognitive science regarding whether computational devices could possess semantic understanding or whether they are merely mimicking human intelligence. Recent research has highlighted limitations in LLMs' reasoning, suggesting that the gap between mere symbol manipulation (syntax) and deeper understanding (semantics) remains wide open. While LLMs overcome certain aspects of the symbol grounding problem through human feedback, they still lack true semantic understanding, struggling with common-sense reasoning and abstract thinking. This paper argues that while adding sensory inputs and embodying AI through sensorimotor integration with the environment might enhance its ability to connect symbols to real-world meaning, this alone would not close the gap between syntax and semantics. True meaning-making also requires a connection to subjective experience, which current AI lacks. The path to AGI must address the fundamental relationship between symbol manipulation, data processing, pattern matching, and probabilistic best guesses with true knowledge that requires conscious experience. A transition from AI to AGI can occur only if it possesses conscious experience, which is closely tied to semantic understanding. Recognition of this connection could furnish new insights into longstanding practical and philosophical questions for theories in biology and cognitive science and provide more meaningful tests of intelligence than the Turing test.

Correspondence: papers@team.qeios.com — Qeios will forward to the authors

Introduction

Theories of language and meaning are not new. One could begin with Ferdinand de Saussure's fundamental insights about the meaning of signs and words as arising from their relationships with

other signs and words. The structure of language is based on these relationships and differences, without which meaning could not emerge. This approach laid the foundation for modern structural linguistics and semiotics.

Later, Ludwig Wittgenstein, in his masterpiece of logical positivism, *Tractatus Logico-Philosophicus*, emphasized that meaningful statements must be rooted in the clarity of concepts within a logical system of propositions. He described language as mirroring the world of facts (not things), developing a "picture theory" of language as a description of our experiences of these facts. The later Wittgenstein departed from the Tractatus in his *Philosophical Investigations*, realizing that language derives its meaning not only from logical structures but also from its contextual dependency. Language mirroring the world and facts is specific to social and practical contexts, and the meanings it conveys must have inherent flexibility and vagueness.

Noam Chomsky's theory of "transformational-generative grammar" championed the view that language is a product of the mind, an idea based on a biolinguistic conception, in which the human ability to use complex forms of language is pre-wired in the brain, in a "language acquisition device." This premise led him to hypothesize the existence of universal grammar, with core syntactic linguistic knowledge being genetically inherited.

Saussure's, Wittgenstein's, and Chomsky's theories are just a few examples of the many competing theories that have emerged over the years. However, a definitive answer to the question of the nature of language and meaning remains elusive.

Similar questions arose in the field of IT. Shortly before Wittgenstein published his Investigations, Claude Shannon, the modern father of information theory, was contemplating the notion of information. What does the word "information" mean? The etymology of words is often more insightful than our contemporary understanding of them. The word "information" derives from the Latin "informare," which suggests that something has been formed or shaped–by molding, carving, or puncturing an object into a pattern or by modifying its physical, internal, or external state. It is about forming or changing something, making it a medium for symbols that convey a message, expressing our thoughts to someone who can understand those symbols and refer to them meaningfully. Whether it is letters and words on paper or bits in a computer, information is always about forming patterns or modifying internal states in objects.

On the other hand, the information we receive also "forms" and "shapes" ideas in our minds. There is, therefore, an important distinction to be made between physical information and semantic information.

Physical information conveys a message in the form of symbols, signs, and tokens, and in communication theory, it is quantified by the Shannon information measure. However, a sequence of symbols and its quantification in bits are, in and of themselves, not meaningful unless a mind apprehends them and "collapses" them into a meaningful semantic whole. Physical information has no meaning whatsoever if there is no mind receiving the message and translating it into coherent thoughts.

Shannon was acutely aware of this distinction. In his seminal 1948 paper, "A Mathematical Theory of Communication," he pointed out: "*The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message [a sequence of discrete symbols] selected at another point. Frequently the messages have meaning; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem*"[1].

In physics and statistical theory, several other definitions of information exist. However, most of these information measures are closely related to Shannon information, and none of them quantify semantic content.

Unlike a mathematical or physical theory of communication, semantic information concerns the meaning of a message. The idea that meaning cannot be reduced to pattern recognition alone and that a physical representation of something is fundamentally different from the thing being represented becomes intuitively evident with the famous Gestalt figures. For example, the popular Rubin's vase-face figure, where our mind switches between the visual interpretation of a vase and two faces, shows how patterns *in-formed* into material structures (the figure on a piece of paper) have no meaning in and of themselves. A meaning-making mind must convert physical information into a coherent semantic whole and eventually even make a choice between mutually exclusive ones. It is a cognitive process that highlights how the significance of things isn't inherent in the things themselves "out there" but rather emerges as a perceived content in us.

For about three decades, these questions were largely ignored as irrelevant philosophical subtleties. Unfortunately, physical information was often conflated with semantic information, and both were simply referred to as "information" without further distinctions. However, the progress of IT, the ever-increasing (physical) information processing power of computers, the advent of AI, and new findings in neuroscience have forced us to reconsider these simplistic assumptions. Questions about the nature of consciousness and the mind have resurfaced, and the relationship between computational and mental states has become a matter of debate.

However, what are consciousness and semantics?

Consciousness in biological systems is widely believed to emerge from complex, integrated activity within the brain. Rather than residing in a single structure, it is thought to arise from the dynamic coordination of sensory processing, memory, attention, and self-representation across distributed neural circuits. Theories such as Integrated Information Theory[2] and Global Workspace Theory[3] propose that consciousness depends on both the quantity and organization of information processing—where integration and accessibility of information are key. This idea relies on the assumption that in biological organisms, the neural architecture enables a unified experience of perception and thought, giving rise to the subjective awareness we call consciousness. However, the true origin and nature of consciousness remain largely debated. For comprehensive reviews on the subject, see: [4] or [5].

Thus, there is no universally accepted definition of consciousness (for a review, see [6]). In this context, the term 'phenomenal consciousness' suffices to illustrate our point. Phenomenal consciousness refers to the subjective, qualitative aspects of experience—often described as 'qualia'—which include raw sensory perception like the redness of blood, the pain of a toothache, and the sweetness of sugar. Phenomenal consciousness encompasses more than just sensory qualia; it includes all conceptual and nonconceptual experiences related to time, space, causality, the body, the self, and the world. From now on, we will simply refer to phenomenal consciousness as consciousness.

It is important to note that consciousness is exclusively a first-person subjective experience that cannot be fully captured through third-person empirical investigation. While I cannot demonstrate that others are conscious, I am aware that I am conscious, and therefore, I have no reason to doubt that other human beings like me are also conscious.What is semantics? Here, we adopt a general understanding. Even though we will focus on the linguistic dimension of semantics due to the disruptive impact of LLMs, we refer not only to formal linguistic competence, which encompasses lexical semantics (the retrieval of individual word meanings) and compositional semantics (the construction of meaning from multi-word utterances), but also to a type of 'non-verbal semantics' grounded in world knowledge and functional linguistic competence, applicable in non-verbal contexts (such as extracting meaning from a picture). This can be described as 'general conceptual knowledge,' which does not necessarily rely on linguistic inputs but is crucial for fluent language use. While LLMs excel in formal competence, they often struggle with the functional aspects of language[7]. In the following, we will focus on the complex reasoning that relies on the non-linguistic or 'pre-linguistic' world knowledge and that determines the functional linguistic skills. subsuming it under the label of 'true semantic understanding' or just 'semantics.'

The Explanatory Gap Between Symbols and Semantics

In 1980, one of the most debated arguments against computationalism came from Searle, who formulated his celebrated "Chinese Room Argument." This thought experiment was designed to challenge the notion that a Turing machine–and, by extension, any (non-quantum) computer or AI system running an algorithm–can truly understand something simply by processing symbols[8]. Searle, who does not understand Chinese, imagined himself locked in a room with a set of rules (in English) for manipulating Chinese symbols. By receiving Chinese characters as input and following the rules to produce the correct output in Chinese, he could mimic a native Chinese speaker, though without understanding what the Chinese symbols actually referred to–that is, what they meant. No matter how perfect a simulation of understanding might be, it does not imply genuine comprehension of the language. One is merely following the syntactic rules of a program without any grasp of meaning (semantics).

By highlighting the distinction between syntax (symbol manipulation) and semantics (meaning), this thought experiment demonstrates the possibility of correctly processing symbols without understanding what they mean. Thus, it raises the question of whether machines, while capable of the former, are also capable of the latter. There is only the signifier without the signified, in the sense that a Turing machine manipulates symbols according to syntactic rules (an algorithm) without having either an abstract conceptual designation or any real-world referents.[1]

Based on these arguments, Searle introduced the distinction between "strong AI"–nowadays referred to as "Artificial General Intelligence" (AGI)–and "weak AI"[9]. There is no consensus on what exactly "general intelligence" means. Here, as will become clear throughout this essay, general intelligence refers to a system possessing what I like to call "semantic awareness"–that is, a form of cognition capable of reasoning and generalization abilities based on a genuine semantic comprehension (in a sense that will be clarified later) of language, data, and, eventually, sensory inputs. Such a system would go beyond the mere emulation currently achieved by symbol manipulation, best-fit algorithms, or next-token predictions.

Nobel laureate Roger Penrose also expressed his doubts in his seminal book *"The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics"* where he laid out his argument that the mind can't be equated with a Turing machine and emphasizes the role of non-algorithmic processes. His ideas were influenced by Gödel's incompleteness theorems, which demonstrate the existence of true mathematical statements that cannot be proven algorithmically within any formal system. While he does not provide a

singular, definitive explanation, Penrose contends that conscious understanding isn't computational. The quality of understanding requires awareness and arises from the conscious mind's ability to perceive and interpret reality in ways that transcend purely mechanistic or computational systems, suggesting that consciousness involves non-algorithmic processes that are tied in particularly to quantum mechanics[10].

In the 1990s, Searle's thought experiment and Penrose's argument were reinforced by the so-called "symbol grounding problem," discussed by Harnad[11]. This issue highlights the difficulty of explaining how symbols in computational systems–whether words, numbers, streams of bits, signals, or more complex representations–can carry meaning without being grounded in sensory experiences or real-world interactions. A gap remains between functional symbols and number-crunching processes on the one hand and meaningful mental states on the other. The divide between syntax and semantics persists.

The question becomes even more complex when we consider that symbols do not always refer to concrete objects or real-world phenomena but can also represent abstract and intangible concepts like "beauty," "justice," or even the concept of "abstraction" and "meaning" itself.

Harnad raises the issue that we do not really know what "meaning" itself is. We cannot simply assume that semantics can be reduced to a form of computation, as computation follows specific syntactic, logical, and mathematical rules based on symbol manipulation–not their meanings. He referred to this as the "symbol/symbol merry-go-round"–the idea that symbols can refer only to other meaningless symbols. Avoiding this infinite regress requires a bridge between those representations and the things or concepts to which they refer.

Even for a Turing machine that convincingly passes the Turing test–that is, its cognitive processes are indistinguishable from those of a human, the question would remain: Does it meaningfully connect its internal representations to their referents and its causal relationships, or is it a mere Searle's Chineese room mimicry, without any real semantic awareness of what they mean? Does it truly understand? What does it mean to "truly understand"? Harnad realized that *"there is a difference between inert words on a page and consciously meaningful words in our heads"*[12].

Thus, following mechanical rules in symbol manipulation is one thing, and what humans mean by "understanding," "knowing," or "comprehension" is another. Semantics remains outside the formal description of any computational model.

The question, then, is: If symbols are meaningless in and of themselves, what generates meaning? While Shannon's notion of physical information is conceptually and mathematically well-defined, it is unclear

how to define semantic information and even less clear how minds convert the former into the latter. How do neural activity patterns supposedly constitute and implement meaning? How do symbols and signs encoded in language elicit meaning in the brain through their reciprocal exchange? It seems that a mind is required as an interface between Shannon information (which refers to physical patterns or states in a medium) and semantic information (which meaningfully grounds the symbol to its referent in a much less tangible mental domain).

The most promising approach to answering these questions and aiming at "naturalizing meaning" seemed to be representationalism–an attempt to explain mental states and their contents in terms of representations and intentionality. However, decades of research in cognitive science and philosophical debates aimed at connecting mental processes that manipulate representations of external reality with cognition, perception, and consciousness have led to little substantial progress.[2] Currently, non-representational theories emphasizing cognition as deeply rooted in bodily experiences and actively engaged with the environment are receiving more attention. (We will address these in the concluding section.) The advent of Large Language Models (LLMs) and their limitations suggests that this must be the case. However, I will argue that while embodied cognition–that is, cognitive processes rooted in interactions with the environment–might provide valuable insights into the relationship between computation, symbols, representations, and sense-making semantics, it can, at best, be only a necessary but not sufficient condition. Phenomenal consciousness and its whole psychological dimension cannot be sidestepped, as it is inextricably intertwined with meaning-making.

Meaning in a Time of Large Language Models

With the advent of transformer-based Large Language Models (LLMs), particularly in their popular form as ChatGPT, these questions elicited renewed interest. LLMs are designed to understand and generate human-like text based on vast amounts of data. Using deep learning neural networks, they predict and generate word sequences by learning patterns, grammar, context, and semantics from large datasets. Trained on diverse sources, including books, websites, and conversations, LLMs can answer questions, complete texts, translate languages, summarize information, and engage in natural dialogue. We have all been impressed by their abilities, which seem to suggest that they can reason effectively and possess a degree of semantic understanding.[3] The deep learning architecture underpinning these models is the transformer, whose key feature is its ability to effectively process and understand long-range dependencies in input data. Key components of transformer-based LLMs include: the self-attention

mechanism (which weights the importance of different tokens in a sequence relative to one another); positional encoding (accounting for the order and relative position of words); multi-head attention (using multiple attention "heads" to capture different types of relationships between words in parallel, thereby enhancing the model's flexibility in understanding context); feedforward neural networks (applying neural networks to refine each token's representation based on the context derived from the attention mechanism); and layer normalization and residual connections (stabilizing training and improving information flow through the network, often reinforced by learning from human feedback).

In simpler terms, transformers in LLMs process and generate text by learning the probability distribution of long-range sequences of words and predicting the next word based on those previously generated. The transformer architecture has proven to be exceptionally powerful, especially in tasks that seem to require some level of semantic understanding. This fact challenges the notion that symbol manipulation alone is insufficient for cognitive states involving deep understanding.

Do LLMs truly understand? Do they possess a semantic understanding beyond imitation? Do they ground symbols as humans do?

While LLMs perform surprisingly well, nobody knows exactly why. Their internal complexity is such that they must be regarded as a black box. The issue is further complicated by the possibility of distinguishing different forms of "grounding." Modern LLMs, as artificial neural networks, compute over high-dimensional vectors rather than discrete symbols, making it more accurate to speak of a "vector grounding problem"[13]. Millière and Mollo distinguish five notions of grounding: referential (how linguistic items refer to real-world objects), sensory-motor (such as linking textual and visual representations), relational (how words relate to other words through definitions), communicative (the establishment of an intersubjective, rule-based language to communicate meaning and ensure mutual understanding), and epistemic (the relationship between linguistic expressions and factual knowledge-based data). Using an analogy with Chalmers' distinction between the easy and hard problems of consciousness[14], the hard problem of symbol grounding pertains to referential grounding. The other forms of grounding are, so to speak, "easy" in the sense that they can be addressed through a more or less sophisticated representational theory. Harnad's "symbol/symbol merry-go-round" relates to the referential grounding.

However, Millière and Mollo argue that LLMs can partially overcome the referential grounding problem, thereby acquiring more than interlinguistic functions through human fine-tuning. This fine-tuning endows LLMs with normative, world-involving functions that allow them to develop intrinsically

meaningful representations of the world–going beyond mere word associations by adding an epistemic layer that connects language to world references. Through human feedback, LLMs gain knowledge about how the world operates, which enables them to form world-model representations rather than purely linguistic ones, even without embodied cognition with sensorimotor skills. In this way, LLMs at least partially address the referential grounding problem.

Yet, these world-involving functions are ultimately human-induced; the human factor is the bridge between language and the world. Human feedback via supervised learning provides modern LLMs with a form of referential grounding that aligns vector representations of words with real-world referents, endowing them with intrinsic meaning that, once fine-tuned, may no longer rely directly on human input. The only way for LLMs to achieve any form of referential grounding is through the intervention of a semantically grounded human agent who sets the normative framework, thereby guiding the non-grounded agent in establishing a world model–that is, a world-system of referents. Thus, even if this approach succeeds, it merely shifts the symbol grounding problem from the machine to the human agent: If an artificial neural network cannot achieve symbol (or vector) referential grounding, how are those biological neural networks inside our skulls able to do so?

Moreover, empirical evidence suggests that there is not anything or "anyone" in an LLM that genuinely "understands" beyond best-guessing and mimicry. Recent research has shown that LLMs can deviate in their responses when presented with irrelevant information, with performance declining when an unrelated sentence is added to a problem. This does not alter the problem's semantic content but appears to "distract" the model[15]. The illusion of "reasoning" displayed by LLMs relies largely on recognizing patterns with a strong token bias rather than true comprehension; this also accounts for their limited generalization ability [16]. Further studies reveal that LLMs lack genuine mathematical reasoning; instead, they replicate logical steps seen in training data without understanding. For instance, rephrasing the same question can yield different answers, and adding redundant information irrelevant to the solution significantly lowers their mathematical reasoning performance. A slight increase in task difficulty can similarly lead to notable performance drops[17]. LLMs tend to translate statements into operations (e.g., interpreting "discount" as "multiplication") without understanding the underlying semantics. This behavior aligns more closely with probabilistic pattern matching than with actual logical reasoning. Such limitations are evident even in basic grade-school math problems and are expected to become more apparent in complex mathematical assessments. That LLMs do not go much beyond sophisticated forms of "semantic-free" imitation is also suggested by the fact that they are inherently

limited by their ability to solve problems that significantly differ from those they saw during their training session.[4] Additionally, the idea that LLMs exhibit "emergent abilities"–skills that are absent in smaller models and that appear only in larger ones–is also under scrutiny. These abilities may be fictitious and not an inherent result of scaling models up in complexity[18].

Other investigations have highlighted AI's limitations in reasoning, as demonstrated by LLMs' failures in seemingly simple tasks such as counting words or reversing a list[19]. This raises the question of why LLMs are generally effective in multi-step reasoning yet struggle with surprisingly trivial problems. Chain-of-thought prompting, which involves generating intermediate reasoning steps before arriving at a final output, exhibits characteristics of both memorization reasoning (in which the model mimics patterns learned from training data) and probabilistic reasoning (in which the model selects the most probable output based on the input)[20]. Further studies have examined how transformer LLMs tackle compositional tasks that require breaking problems into sub-steps and synthesizing those steps into precise answers. It turns out that they often simplify these tasks by matching linearized subgraphs–training examples that mirror the computations needed to solve test examples–without developing systematic problem-solving skills. Moreover, abstract multi-step reasoning problems based on autoregressive generation (the statistical prediction of the next sequence value based on previous values) tend to deteriorate rapidly as task complexity increases[21].

A study conducted by OpenAI[22] with Large Reasoning Models (LRMs)—a new type of AI designed to excel at complex problem-solving by breaking tasks into smaller, logical steps—revealed that while these models perform better on reasoning benchmarks, they experience a complete collapse in accuracy beyond certain complexities. Their reasoning effort increases with problem complexity up to a point, after which it declines. In high-complexity tasks, they fail to develop generalizable reasoning capabilities, experiencing total collapse. Additionally, their limitations in performing exact computations became apparent; they might succeed in one task but repeatedly fail to provide correct answers in others. Apple's take is that these models create an illusion of thinking but lack true reasoning. They are essentially large pattern matchers that falter when they encounter data outside their training set, leading to breakdowns in generalization. Before we can consider them capable of genuine thinking processes with true reasoning abilities, something crucial is still missing.

This suggests that LLMs' abilities scale asymptotically: Beyond a certain threshold, adding more neural layers, faster processing, increasing training data, enhancing supervised reinforcement learning, and

further fine-tuning or prompt engineering do not significantly improve their performance.[5]

Additionally, unlike humans, LLMs cannot correct their erroneous guesses[23]. This aspect is crucial, as one might argue that humans are prone to similar flaws. However, humans can recognize their logical, inferential, and deductive mistakes, which allows them to restart the problem-solving process with different premises or methodological approaches. This capacity to overcome biases and reach the desired result is possible only because of a semantic awareness that extends beyond mere symbol manipulation, enabling individuals to identify incorrect or nonsensical outcomes even if they do not know the correct answer.

After all, it does not require complex scientific research to recognize the shortcomings of LLMs in reasoning, semantic competence, and the understanding of inherent meaning. Sustained use and interaction often reveal interesting insights.[6] For example, if one were to ask how many U.S. states have names beginning with the letter K, one might be told that there are four: Kansas, Kentucky, Kansas, and Kentucky. These issues similarly affect generative AI. If instructed to create a picture with no elephant in the room, it might repeatedly generate images featuring an elephant in a room, thereby demonstrating a failure to grasp negation (the "not-questions"). When asked to depict someone writing with their left hand, the model may refuse to comply, always producing an image of a right-handed person instead. This makes sense if we keep in mind that the model makes probabilistic guesses rather than semantic ones, as it has been trained predominantly on a large dataset where the proverbial sentence 'elephant in the room,' and images of right-handed individuals are quite common.

Similarly, self-driving cars' ability to classify objects and people is impressive, yet countless instances reveal their lack of comprehension regarding what they are "seeing" (e.g., distinguishing between real people on the street and fictional figures on a billboard). Classifying and categorizing are not the same thing as understanding. The critical questions are whether a car can effectively navigate without a semantic understanding of its environment–recognizing the street, cyclists, traffic lights, and so on–and whether a level V self-driving car could achieve comprehension that even humans might struggle to acquire without conscious experience.

It could be argued that further fine-tuning and additional training sessions might resolve these issues and reduce the error rate. After all, humans also make mistakes. However, this is not the decisive point. The crux of the matter is that when the model fails, it can do so spectacularly, sometimes producing absurd answers that clearly illustrate a lack of genuine understanding.

On the other hand, it is also true that humans are not entirely immune to similar senseless perceptual hallucinations or ludicrous cognitive failures, particularly in individuals with neurological disorders. Oliver Sacks's best-selling book "*The Man Who Mistook His Wife for a Hat"* compiles case studies of such rare but real phenomena. However, the question remains whether such pathologies can be addressed without consciousness. Neural networks fit data and classify information, whereas humans integrate sensory data into a unified experience—what we might call a "perception of meaning"—that represents a semantic whole, no matter how senseless it may appear. Humans can overcome these hallucinations precisely because of conscious experience, while machines are confined to sense-making that cannot go beyond relational systems and statistical inferences based on abstract tokens[24].

Overall, we are beginning to realize that despite billions being invested in research and development, LLMs continue to hallucinate, fabricate information, and lack genuine understanding of the world. They excel in deep learning but struggle with generalized abstract reasoning, and they have difficulty understanding wholes in terms of their parts.[7] Though we are dealing with a black box, we know it is ultimately a Turing machine working with symbols according to logical, syntactic, context-based, or probabilistic rules in someone else's natural language. These symbols stand for external referents grounded in the subjective experiences of someone else–experiences that the symbols themselves cannot convey to others. While it would be unfair to label LLMs as mere "stochastic parrots," they still fall short of human capabilities in natural language inference, their dialog is limited in information and can be unreliable, frequently failing in differentiating between fact and fiction, generating inaccurate claims (for a good summary of the strengths and weaknesses of LLMs see [25].

While LLMs' cognitive skills are undoubtedly impressive and surprising, it is becoming increasingly evident that there is still no true reasoning or common-sense knowledge that extends beyond sophisticated Chinese room machinery. There is no "ghost in the machine" with semantic awareness; fundamentally, the machine does not understand a thing.

Does the machine 'cognize'? It certainly doesn't align with the human understanding of 'cognition.' When we look at a green apple or a yellow pear, we don't perceive vectors or matrices. However, one might argue that nonetheless, the machine does possess a form of cognition as well. We shouldn't assume that human-like cognition is the only possible one. In fact, representing reality as a complex relationship of multidimensional vector space enables AI to engage in meaningful conversations with us.

Yet, without succumbing to anthropocentrism and placing our cognition at the center of the universe, we need to acknowledge the differences between human and machine meaning making. If there is none, the

question arises how the internal workings ultimately represented as digits in a memory chip, translate into the rich experience of perceiving the 'greenness' of an apple, feeling its shape and size, and even tasting it. A memory cell labeled "apple" or "pear" is merely a symbol encoded in the physical state of a circuit. There exists a subjective and experiential dimension that cannot be encapsulated by a symbol alone, regardless of how complex the computations leading to its identification may be. This issue is not exclusive to machines; it also applies to humans. If you have never tasted an apple, no amount of description—be it through language, science, mathematics, computation, or neurological explanation—can truly convey what that experience is like.

If even the most advanced AI can pass a molecular biology exam yet remains far from achieving human intelligence–or even the intelligence of a cat or dog–something might be fundamentally flawed in our understanding of what intelligence truly is. Alternatively, we might be overlooking a crucial aspect of our nature as living beings.

Meaning and Qualia

Following Wigner's famous question about *"The Unreasonable Effectiveness of Mathematics in the Natural Sciences"*[26], for reasons we are still struggling to fully comprehend, mathematics appears to mirror aspects of the world. At least in physics, chemistry, and, to some extent, biology, it serves as a descriptive framework for physical reality. Perhaps the question is not so much why mathematics is effective but why it mirrors some aspects of the world at all.

One might ask a similar question in the domain of AI: If LLMs lack a true understanding of what they are doing, why are they so unreasonably effective?

It is challenging to discern precisely how modern LLM implementations–essentially black boxes–operate, given that they employ millions of neurons and billions of weighting parameters and manipulate trillions of numbers, vectors, matrices, and tensors. As of the time of this writing, there is no definitive answer. For now, we can only speculate.

For example, Wittgenstein might not have been entirely off: Language, like mathematics, isn't merely a human cognitive creation unrelated to the organization of the world but, rather, might reflect its structure. We might conjecture that the statistical distribution of words is not solely tied to human cognition but also offers a window into the conditional structure of the world as mediated through human language. The patterns embedded in language might reflect the patterns inherent in the world.

Linguistic patterns track real-world patterns, and which representational structure may be captured within LLMs[27]. If so, given that LLMs are pattern matchers trained to "autocomplete" based on complex relationships between vast numbers of tokens (representing properties of the world and their interrelations), their effectiveness might be less surprising.

Nonetheless, while any mathematical or formal proposition acquires meaning only when it is communicated from grounded semantic agents to other grounded semantic agents, one could go a step further and suggest that all of what we perceive and conceptualize about the world is ultimately a symbol, sign, token, or image within our cognitive awareness. This might be even less surprising to the philosophical idealist, who posits that the world itself is an "idea" or symbolic expression. Such a view aligns with certain Eastern philosophical and mystical theories in which the universe represents the expression of a creative, transcendental "real-idea," forming the foundation for all meanings, signs, words, and human language[28].

Be that as it may, it is not necessary to venture into metaphysical speculations to capture some essential aspects of how our sense-making cognitive processes emerge. The primary challenge here is not conceptual or theoretical but, rather, the shift from a third- to first-person perspective.

In this respect, it might also be interesting to consider Harnad's recent follow-up, which he formulated in the form of a dialogue with ChatGPT-4[29]. According to Harnad, what ChatGPT lacks is *"a direct sensorimotor grounding to connect its words to their referents and its propositions to their meanings."* Machines operate within the realm of symbol manipulation without any grounding in real-world physical interactions and experiences; they lack the subjective and conscious dimensions of understanding. He further suggests that the only solution is to embody AI models–that is, to ground them–with sensory and motor outputs that allow them to interact with and learn from their environment–thereby building a world model.

However, he carefully distinguishes the symbol grounding problem from any allusion to the hard problem of consciousness (or the "problem of qualia" or the "explanatory gap"[14]). The symbol grounding problem concerns how symbols acquire meaning by linking them to their referents in the real world or to concepts that enable understanding. In contrast, the philosophical issue of qualia involves how neural processes give rise to the subjective experiences of "what it is like" to be in a qualitative state of consciousness. The symbol grounding problem is about representation and understanding in artificial

systems, not about why the complex machinery in our brain supposedly gives rise to subjective experiences.

Qualia–the introspectively accessible, subjective, phenomenal aspects of our conscious lives, such as the bitterness and warmth of coffee, the redness of an apple, the smell of freshly cut grass, or the sharp pain from a paper cut–are not reducible to mere information processing. Our private experiences of pleasure, pain, feelings, and emotions, along with sensory perceptions like seeing, hearing, touching, smelling, and tasting, convey qualities of the world such as colors, sizes, and shapes. These experiences are not merely concepts or abstract symbols in a quantitative database; they are grounded in a qualitative experiential dimension.

Harnad's insights are inspiring; however, in my view, he fails to close the circle. He begins by associating meaning with "the subjective, felt experience of understanding," the "phenomenological aspect of what it feels like to mean or understand something"–elements that AI lacks, as it has no "direct experiential grounding" or "direct sensorimotor experiences." Yet, somewhat surprisingly, he stops short of explicitly concluding that, for a machine to achieve what he terms "direct symbol grounding" (and, ultimately, a state of AGI), it must be conscious.

Disentangling the symbol grounding problem from the hard problem of consciousness and then reiterating that natural language must be grounded in real-world experiences is a misstep. This is because the former is an aspect of the latter. The symbol grounding problem is itself "grounded" in the problem of qualia and the explanatory gap. Asking how a symbol can be meaningfully connected to what it represents without allowing the connection to be linked to phenomenal consciousness is a form of philosophical self-censorship.

I submit that the gap between syntax and semantics is qualitatively much deeper than previously assumed. Scaling alone will not bridge this gap. While embodied architectures could extend text-based information to include sensory-based data with feature-detecting and abstracting capacities from real-world engagement, this would not transform Shannon-type sensory information into semantic information, thereby grounding the ungrounded in any miraculous way.

Sensorimotor embodiment without qualia–that is, without phenomenal consciousness and its sentient, subjective experiences–could, nonetheless, lead to more nuanced forms of non-human intelligence. However, there is no reason to believe that replacing or augmenting the linguistic symbols in an LLM super-dictionary database with a sensorimotor super-dictionary of pixels and mathematical functions representing environmental sensory signals could provide genuine grounding beyond syntactic

computational understanding within an abstract categorical space. Names and verbal descriptions require not only a sensorimotor embodiment but also a conscious experience of the features they represent to be grounded in a semantic space. Vast data sets (textual, numerical, and/or sensory), immense computing power, and any form of embodiment may be necessary to build meaningful world models, but it is hard to believe they would be automatically sufficient to allow an AI system to gain a true understanding of the properties of the real world with which it is engaging without having a subjective experience of the very same properties.

It will remain a "reasoning" based on complex classification schemes that formally represent relationships between things and phenomena but without actual understanding. The result is still akin to a Turing machine or a "Chinese room mimicry" that requires far more data for learning than humans do and that will ultimately reveal its limitations. After all, one can teach children how to roll a ball without giving them millions of examples, because a child doesn't see the world through numerical or symbolic representations but through a sense-making mind that relies on lived experiences. Biological organisms' meaning-making has an essential component that machines fundamentally lack–a gap that information processing alone cannot bridge. To put it in Vallor's words: *"We are more than efficient mathematical optimizers and probable next-token generators"*[30].

Meanwhile, taking a first-person perspective reveals that our semantics-based cognition is inherently tied to conscious experience–always. We cannot truly understand colors, sounds, tastes, smells, temperature, touch, or sight without having directly experienced seeing a color, hearing a sound, tasting chocolate, smelling an odor, or touching an object. Is there a difference in "meaning" between, say, listening to a symphony of Tchaikovsky and "knowing" it only in the form of a digital transcription? An LLM might "know" everything textually about apples or the Fourier transform of a sound signal, but there is no semantics to extract, as all this remains an abstract, multidimensional vector representation defined by weighted numerical parameters. Without experiencing "what it is like" to taste an apple or feel its texture or to be immersed in the chills and thrills of musical rapture, it cannot achieve any form of symbol–or vector–grounding. Similarly, someone might explain all the chemistry and physics of an H2O molecule, but you cannot grasp the essence of wetness until you have felt the sensation of water yourself.

This leads us to Jackson's famous knowledge argument[31]: Neurophysiologist Mary may have all the physical information about color vision, but she still lacks knowledge if she has not experienced the qualia of colors.

The question is: Can any physical information be linked to semantic information without the presence of subjective experience?

One could reverse the argument and ask, "What kind of 'knowledge' could a purely phenomenological experience without physical information or conceptualization convey?" It is unlikely that this would lead to semantic comprehension. I might experience the blueness of the ocean, the chill of an environment, the sourness of a lemon, or its shape in my hands without having any concept or understanding of what the ocean, the environment, or a lemon truly represents.

This does not mean, however, that congenitally blind individuals cannot understand others discussing spatial properties, geometric forms, light, colors, or other visual sensations. They have an indirect understanding of these concepts, achieving referential grounding much like an LLM–not through sentience but via information mediated by other sentient agents. And it is known that blind subjects are not defective in their learning of language, of space, and objects. Blind learners not only learn the forms of language but their meanings as well[32]. The key difference between even the most advanced AI and a person with sensory impairments is that the absence of sight in humans is compensated for by other subjective experiences, like sound, touch, taste, and smell. Moreover, blind human beings retain the ability to perceive the world through other subjective experiences beyond the senses as well. These experiences enable the indirect semantic comprehension of visual concepts even if direct visual perception is lacking, as the representation is supported by other forms of conscious experience–something a machine cannot replicate.[8]

What kind of comprehension do blind individuals have of the visible world, and to what extent can they eventually recover an ordinary understanding of it? This is not a new question. For instance, in the 17th century, William Molyneux proposed a thought experiment involving a congenitally blind person who has learned to recognize objects solely through touch. Imagine this person can distinguish a sphere from a cube by touch but has never seen them. Now, suppose this person could suddenly see. Molyneux questioned whether, if the sphere and cube were placed on a table, this person would be able to identify which was which *without* touching them.

The question was debated by philosophers like Locke and Berkeley, who essentially agreed that if our mind cannot establish a relationship between the sensations of the tactile and visual worlds, the answer to Molyneux's question must be negative. The connection between these two realms–specifically, grounding visual experience in meaningful semantic content known only through tactile experience–cannot be established through mere intellectual inference. To create this connection, one must link real-

world experiences with previously grounded concepts derived from other kinds of lived experiences into some world-model.

Interestingly, the Molyneux problem can now receive a scientific answer. It is possible to gain some insights into the state of "meaningless awareness" from individuals affected by congenital cataracts that are later treated. A congenital cataract is an organic anomaly present at birth that clouds the eye's natural lens and can result in amblyopia–a disorder in which the brain fails to process visual stimuli. In the 1930s, Marius von Senden became the first to describe the perception of space and shape in congenitally blind individuals before and after surgery[33]. When the sight of previously blind patients was restored and their bandages removed, they did not see the world as one might assume. Instead, they experienced a blotch of chaotic colored patches that meant nothing to them. They required time and practice to make sense of what they saw.

In 2011, neuroscience shed further light on this topic. An Indian research project, "Project Prakash," aimed at treating blind children while also seeking answers to scientific questions about how the brain develops and learns to see, provided evidence supporting Locke and Berkeley's views[34]. In the treatment of congenitally blind children aged 8 to 17, researchers found that, upon gaining sight, these children struggled to visually match objects they had previously known only through somatosensory information. However, this capacity developed quite rapidly; their skills in relating visual perception to somatosensory sensation improved within a few days of sight restoration and were nearly fully present within a few months. Physical information could finally be grounded in semantic information through sensory tactile experiences, as the latter was already rooted in a meaningful conceptual unity.

This indicates that linking tactile knowledge to visual knowledge is not an innate ability. Furthermore, it demonstrates that meaning emerges not only from the relationships among words, symbols, or vectors but also through the contextualization of different forms of qualitative experiences. Only after this experiential stage can the mind connect a symbol (the signifier) to its referent in the real world (the signified).[9]

One thing that must be noted is that semantic content does not require an experiential link to the physical world–that is, direct sensorimotor engagement. Most notably, abstract and intangible concepts like "freedom," "courage," "truth," "beauty," "justice," and "wisdom" lack objective referents in the physical world. Nonetheless, they are deeply meaningful to us and, in a sense, real and concrete. Their concreteness arises from subjective feelings and a "perception of meaning" rooted in the experiential dimension of the mind. These concepts are thoughts, often accompanied by an inner emotional state,

that also have a qualitative meaningful aspect. Even purely abstract entities, such as numbers, lack physical reality outside our minds (or, as Platonists would argue, exist only in a world of perfect Forms), and the symbols for numbers hold no inherent significance unless they are associated with a "number sense"–an experiential link between symbol and quantity mediated by a subjective mental experience of extension or quantity.[10]

Qualia are not limited to sensory experiences; thoughts are "mental qualia," and emotions are "emotional qualia." The "perception of meaning" is, above all, a subjective "mental quale" that emerges from and is intertwined with other qualitative experiences, whether somatosensory, affective, or otherwise. In this way, there exists a phenomenon of what "it is like to be" in each of these internal states–an aspect that any symbol-to-symbol or representation-to-representation relationship cannot fully capture.

Moreover, semantics has a wholistic character. The perception of meaning is always associated with an integration of information that neuroscience still struggles to understand [35]. How the brain integrates information from various sensory inputs (sight, sound, touch, etc.[11]) and distinct neural processes into a unified, coherent experience is unclear. For example, when we look at an apple, we don't perceive separate visual features like color, shape, and size individually; rather, we see a single, whole apple. It is what I referred to as a semantic whole that suddenly appears in our awareness by looking at a figure made of pixels, parts, structures, boundaries, colors, etc. The question of how different parts of the brain "bind" features together into one experience, despite being processed in different brain regions, is commonly known as the "binding problem" and is intimately related to the unity of consciousness.

The implications for AI are that without a conscious subject (an entity or individual capable of conscious qualitative sensations of the properties of the world), a machine cannot grasp anything beyond the purely abstract "understanding" of its representations, symbols, letters, words, sentences, signs, and numbers. A self-driving car, no matter how advanced or sophisticated its neural network and information processing system might be, cannot comprehend what an image of a street, a cyclist, or a traffic light represents if it lacks the complementary subjective experience and an integrative process that binds them into a meaningful unity. True knowledge cannot be achieved through functional processes alone; understanding requires a specific type of qualitative phenomenal dimension. An image can be converted into vector spaces, matrices, relationships, curve fittings, and probability laws, beyond their complex functional relationships. However, unless the light signals hitting the pixels of the image on the CCD camera elicit a subjective lived experience in someone or something, all these mathematical abstractions, multidimensional vectors, or neural representations will remain meaningless physical

information without semantic grounding. While physical information can be measured by (negative) Shannon entropy, without sentience, nothing can convert it into semantic information. This is because "understanding" is an aspect of phenomenal consciousness itself and can't be abstracted from the hard problem of consciousness.

Semantics, Life, and the Unconscious

Thus, consciousness plays a fundamental role in the transition from symbols to semantics. An information processing system may interact with the environment through sensory-motor embodiment, but that alone does not make it a true semantic agent. An embodiment allowing for the representation of a world model should not be confused with a system experiencing the world. For it to achieve this, it must possess a comprehension of the world based on experiences, not just representations, abstract symbolic descriptions, or indirect linguistic grounding–that is, "qualia-less" physical information. Transitioning from the vectorization of language to the vectorization of the environment may represent a significant advancement in AI, but this alone is unlikely to overcome its semantic deficiencies, as these will still be ungrounded quantifying numbers or symbols all the way down. Embodiment may be a necessary condition, but it is not sufficient to transform unconscious comprehension into conscious comprehension. To achieve a deeper understanding of the world, including its contents, properties, and related concepts, I must grasp it qualitatively through conscious experience–that is, with qualia in the form of sensations, feelings, and lived sensory perceptions. A tight and inextricable relationship exists between feeling and knowing, sensing and understanding, perceiving and comprehending. We cannot disconnect these aspects of cognition and treat them separately. While the hard problem of consciousness addresses how consciousness emerges from material and/or functional processes, such as neural activity in the brain, the symbol grounding problem explores how meaning arises from these same processes. Symbol grounding is impossible without experience because meaning is an inherent aspect of consciousness. To understand something implies being conscious of it. Meaning emerges not only due to a relationship between abstract tokens, but also as a pre-linguistic association of experienced qualities into a unified construct apprehended and comprehended as a singular aware sensation. Meaning *is* subjective perception; the "perception of meaning" *is* a quale in itself.[12]

Embracing this 'semantics needs consciousness' standpoint might have significant implications for theories in biology and cognitive science that explore the relationship between cognition and the defining characteristics of life. Examples include autopoiesis and enactivism, complex dynamical

systems theories based on predictive processing, the Free Energy Principle theory, the Extended Mind hypothesis, Integrated Information Theory, the Cellular Basis of Consciousness model, Rosennean complexity, and biosemiotics, to name just a few.

Due to space limitations, this essay cannot summarize these theories, not even superficially. However, it can be said that one common theme they emphasize is the close intertwining of cognition and life processes. This suggests that cognition is not merely a high-level brain function but, rather, is foundational to life itself. The mind is not a passive container or a separate system that simply models the world; instead, cognition emerges from the direct interaction between organisms and their environments (and, eventually, among multiple agents, objects, or biological networks). The embodied mind encompasses more than the brain; it represents a synergistic relationship between the brain and the body, in which cognition is a process of minimizing predictive error and energy expenditure to maintain self-identity and is deeply rooted in bodily experiences actively engaged with the environment. Cognition, problem-solving skills, perception, and goal-directedness are not limited to brain functions in complex organisms but, rather, are essential traits of life, present even in simple bacteria, slime molds, and plants.

The relevant common trait in the present context is that, according to these perspectives, meaning-making is a fundamental characteristic of all living systems as well. It emerges from the continuous interaction between an organism and its environment. Meaning is enacted through the adaptive coupling between the internal state of the organism–whether unicellular or multicellular–and its surroundings (e.g., what supports survival is meaningful, while what threatens existence becomes meaningless). In biosemiotics, life is viewed as a network of semiotic (sign-based) processes, with organisms constantly interpreting signals and symbols from their environment. Here, meaning emerges as an intrinsic feature of life through signs and codes operating at cellular, genetic, and ecological levels[36].

In other words, these approaches aim to reduce meaning-making starting from non-reductionist and wholistic perspectives described by different forms of material causation. They seek to naturalize meaning by beginning with basic physical information processes and bottom-up or top-down complex dynamics accounting for the emergence of semantic information in cognitive biological systems. For instance, this is visible in the “experience-blind” naturalism pursued by enactive theory, which attempts to reduce experience to a “naturalized phenomenology.” While enactivism recognizes the limitations of traditional naturalism and advocates for a “new naturalism,” it still struggles to account for the

emergence of experience within the natural world. There are good reasons to believe this will end similarly to the pragmatist philosophy of nature from the previous century[37].

Consequently, I argue that for these same reasons, attempts to explain the emergence of “semantic agency” in nature from an experience-blind perspective are bound to fail. As this paper has argued, semantics is inherently tied to conscious experience and, therefore, cannot be separated or abstracted from it. To put it metaphorically, trying to account for the emergence of meaning in living systems without conscious experience is like trying to explain water waves without water.

I would like to address one of the potential objections to this viewpoint.

Many of our cognitive processes occur without our conscious awareness. Experimental psychology demonstrates that sophisticated cognitive abilities seem to occur without consciousness. For instance, individuals can process semantic content and focus on objects even when masking techniques prevent the information from reaching their reported access consciousness. In cases of blindsight, individuals who are cortically blind can respond to visual stimuli they do not report to perceive, due to lesions in the primary visual cortex. Remarkably, they can often correctly identify objects that they believe they do not see.

Does this show that understanding is possible in the absence of consciousness?

I believe this conclusion is premature. Our understanding of consciousness is largely based on a superficial subjective experience. What we refer to as 'unconscious' or 'lack of attention' may actually represent another form of conscious awareness and attention—not unconsciousness, but a subliminal, non-metacognitive––that is, non-reportable––type of awareness. We often confuse different states of consciousness with being 'unconscious' because we cannot relate them to our ordinary state of awareness. In these states, we may not be genuinely unconscious; instead, we may simply lack mnemonic access to those experiences.

What is it like to be in a conscious state without memory? Are we truly 'unconscious' during dreamless sleep? Are we unconscious during anesthesia? Are we unaware while in a hypnotic trance or sleepwalking?

I argue that we do not have definitive answers. Phenomenal consciousness—that is, some form of sentience—can be present but quickly fade within seconds. Consciousness-mediated semantic cognition might still function in these altered states. Therefore, we should avoid making hasty conclusions, as this remains an open question.

Is the AI-AGI Gap a Conscious Semantics Gap?

As for any AGI narrative, since the success of ChatGPT and the impressive capabilities of LLMs, there has been a marked increase in discussions about the imminent arrival of AGI–human-like intelligence. However, many of these discussions overlook the intrinsic connection between general intelligence and consciousness.

Enhancing AI performance by adding another trillion neurons, increasing the number of parameters, providing sensory-motor environmental interaction, integrating it with neuro-symbolic AI, or developing more advanced deep learning algorithms, internal feedback loops between sensory acquisition and a central processing system to construct more accurate world-models, etc., will no doubt advance current AI capabilities. However, it will not accomplish the "semantic trick." The elephant in the room must be addressed: the possibility that biological intelligence, with its semantic awareness, might not be replicable in machines unless we learn how to manufacture consciousness itself (if it is even possible to do so). Without consciousness, true AGI cannot exist. The current widespread discourse on the impending AGI revolution, which is predicted to shape our future, warrants a more critical examination that considers the first-person perspective.

Though I am an "AGI skeptic," the rationale presented here does not necessarily imply that AGI is impossible or that consciousness in machines is impossible. The claim is that if we want the machine to become semantically aware—where, by 'awareness,' we mean a subjective dimension capable of experiencing—it must become conscious as well. Because awareness and consciousness are inextricably intertwined. Otherwise, it remains semantics without awareness, such as Searle's Chinese room.

On the other hand, recognizing the relationship between meaning and consciousness might offer new approaches to longstanding questions. The distinction between conscious intelligence and machine intelligence (or even the distinction between machine and living organism itself?) becomes apparent when, sooner or later, the machine makes a glaringly irrational mistake–one that no conscious, semantically aware agent would make. This reveals its lack of genuine semantic understanding and could be taken as evidence of its absence of subjective experience. In principle, this could suggest the basis for a new test of intelligence, potentially replacing the famous Turing test. The goal would be to develop a test designed to assess whether a machine possesses a deep, semantic understanding comparable to that of a conscious being building its meanings beyond symbol manipulation in a semantic space determined by

qualitative experiences. If it does, this would imply a form of intelligence that points to some level of experiential awareness, even if not necessarily human.

This line of thought could also serve as an argument for or against the hypothetical existence of "philosophical zombies" (or "p-zombies"). A p-zombie is imagined as a being physically identical to a human and behaving exactly like a conscious person but lacking any inner life or conscious experiences. However, because a p-zombie would not have the sentience necessary for meaning-making, it would eventually reveal itself by making naïve errors, much like non-conscious AI systems do. This would betray its lack of consciousness and semantic understanding. It is reasonable, then, to conclude that beings that act like humans yet lack subjective experiences do not exist, as otherwise, we would have already distinguished them from conscious humans. A test centered on meaning-making provides a stronger and potentially more convincing proof of the presence or absence of consciousness and true intelligence in a machine than the Turing test does.

One possibility could be to draw inspiration from modern tests that evaluate forms of consciousness in animals and apply them to future machine intelligence that could potentially exhibit signs of conscious awareness. At this stage, I can't provide concrete, implementable testing frameworks, but considering the possibility that meaning-making and consciousness are interconnected offers, in and of itself, a new perspective that could potentially lead to innovative practical approaches and methodological tests as well.

It is important to note that no contemporary AI exhibits any signs of agency and personhood [38]. Without input, these systems do not act autonomously. The scientific concept of agency—encompassing intentionality, autonomy, self-determination, purposeful decision-making, and its connection to cognition—is itself a complex subject[39]. However, it is clear that regardless of the definition we adopt, nothing resembling agency exists in current AI systems. The question of whether agency and consciousness might be intrinsically related deserves more attention, particularly as we speculate about the potential emergence of AGI.

Conclusions

I have focused primarily on the relationship between human-like experience and semantics because it represents the most familiar form of cognition for us. Nonetheless, I believe similar arguments can apply to non-human animals, with sentience also serving as the basis for meaningfully navigating the world.

Conversely, one could argue that machine learning without any conscious experience could be viewed as a form of "understanding" as well. Avoiding an anthropocentric perspective means acknowledging that our concept of understanding need not be confined to human ways of making sense of the world; indeed, computers could "understand" things in their own manner.

In fact, recent findings in mechanistic interpretability[40], the field dedicated to understanding the internal reasoning processes of trained neural networks and LLMs, reveal that these systems do not rely solely on statistical inference. LLMs also develop internal structures that are functionally analogous to key aspects of human understanding. They perceive connections and form learned vector features within a multidimensional latent space, which humans typically bind and coalesce under a single concept. However, the understanding exhibited by LLMs is fundamentally different from human understanding. We need new frameworks for understanding that can embrace the emerging forms of intelligence we are creating, without assuming that human comprehension is necessarily unique, exclusive, or superior. That said, it is important to question whether machine understanding differs from human understanding precisely because of its lack of subjective experience.

On the other hand, animal forms of consciousness may result in entirely different types of semantic awareness of the world. We should not assume that human subjective experience is the only standard either. Perhaps there even exists an extraterrestrial civilization with a radically different understanding of reality. What do we know?

This is plausible, as human semantics might be only one of infinitely many varieties of semantics. However, this paper's main claim does not focus on any specific form of human cognition. Rather, the paper posits that semantics is inherently tied to conscious experience, whether human, animal, extraterrestrial or else. If semantics is not grounded in some form of consciousness and sentience, without any perception of the world's properties through qualitative subjective experiences, then it would, by definition, be a p-zombi semantics. In principle, we would eventually be able to distinguish zombies from other conscious beings. Their "understanding," exhibited by a cognition devoid of qualia–no matter how complex, powerful, or sophisticated–would remain at the level of a Turing machine head processing physical information on a strip of tape. Nothing would convert this physical information into genuine semantic understanding.

The realization of AGI, if it ever happens, will first require the creation of conscious machines. There is a foundational, principled argument against the imminent arrival of AGI, as we remain far from developing systems that transcend memorization, pattern matching, or probabilistic assessments to achieve

genuine, complex abstract reasoning based on conscious experience. Therefore, a first-person perspective, not only in the philosophy of mind but also with regard to addressing fundamental questions about AI, is essential. Approaching AI from this perspective might also help us gain a deeper understanding of ourselves.

Whatever the case, framing certain philosophical problems can serve as the initial step toward developing new methodologies and practical applications. Our actions are influenced not only by our knowledge but also by our underlying philosophical worldviews. If programmers, developers, and software engineers explored deeper questions about the origin and nature of consciousness, intelligence, the mind, and life, they could potentially gain insights that a purely technical perspective cannot provide.

Footnotes

[1] One might argue that the meaning of symbols is rooted in their "use"—specifically, in the function they serve within a given context. However, from a more computational and reductive perspective, one could add that there is no inherent 'use' or function in the cell state that a Turing machine's head reads and reacts to, other than the meaning assigned by an external semantic agent.

[2] For a good summary of these theories and their inability to solve the problem of meaning in AI shortly before LLMs took center stage, see[41].

[3] Here, the term "reasoning" encompasses not only problem-solving abilities that rely on logical processes like deduction, induction, abduction, multi-step inference, or mathematical skills, but also those requiring abstraction, generalization, semantic discrimination rather than plausible best guesses, and the capacity to organize thoughts and conceptual structures into coherent, meaningful wholes–that is, what is commonly referred to as "common sense," "rationality," or "thinking" in human cognitive skills.

[4] Anyone with teaching experience recognizes this behavior in students who cheat. When they do not know the answer, they abandon meaningful reasoning and instead try to arrive at the correct answer by guessing–mimicking the methods and rules they have seen applied but without truly understanding the problem-solving method.

[5] This is also reminiscent of the fact that, contrary to popular belief, humans do not possess the largest brain. Other species have larger brains in terms of size, number of neurons, or weight. What, then, determines human cognitive dominance?

[6] For many more examples, see Dr. W. Hsu's collection of LLM failures: https://lnkd.in/eUW6TYCY.

[7] For further review on how close we are to AGI, see also[42], and references therein.

[8] One of the most dramatic and well-known examples of this is the story of deafblind Helen Keller.

[9] One might argue that in human infants, certain cognitive developments occur prior to full conscious awareness. This suggests that some forms of understanding can arise without phenomenal consciousness. However, this conclusion relies on the questionable assumption that newborns have no subjective experiences, which contradicts all external evidence. The burden of proof lies with those who claim that babies lack sentience.

[10] Children who, for whatever reason, do not learn to relate the number symbols, or a perception of numerosity of objects, to a "number sense" are most likely to develop mathematical learning disabilities like dyscalculia[43].

[11] A careful first-person investigation reveals that this isn't the case only with sensory inputs, but with thoughts and emotions as well.

[12] I like to reframe this as the "hard problem of semantic awareness."

References

1. ^*Shannon CE (1948). "A Mathematical Theory of Communication." Bell Syst Tech J. **27**(3):379–423. doi:10.1002/j.1538-7305.1948.tb01338.x.*
2. ^*Oizumi M, Albantakis L, Tononi G (2014). "From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0." PLoS Comput Biol. **10**(5):e1003588. doi:10.1371/journal.pcbi.1003588.*
3. ^*Baars BJ (1988). A Cognitive Theory of Consciousness. New York: Cambridge University Press.*
4. ^*Seth AK, Bayne T (2022). "Theories of Consciousness." Nat Rev Neurosci. **23**:439–452.*
5. ^*Kuhn RL (2024). "A Landscape of Consciousness: Toward a Taxonomy of Explanations and Implications." Prog Biophys Mol Biol. **190**:28-169. doi:10.1016/j.pbiomolbio.2023.12.003.*
6. ^*Gulick RV (2014). "Consciousness." In: The Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/consciousness/.*
7. ^*Mahowald K, et al. (2024). "Dissociating Language and Thought in Large Language Models." Trends Cog n Sci. **28**(6):517–40. doi:10.1016/j.tics.2024.01.011.*
8. ^*Searle J (1980). "Minds, Brains and Programs." Behav Brain Sci. **3**:417–57.*

9. ^ Searle J (1990). "Is the Brain a Digital Computer?" Am Philos Assoc. **64**(3):21–37.
10. ^ Penrose R (1989). The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics. Oxford: Oxford University Press.
11. ^ Harnad S (1990). "The Symbol Grounding Problem." Physica D. **42**(1-3):335–346. doi:10.1016/0167-2789(90)90087-6.
12. ^ Harnad S (2007). "Symbol Grounding Problem." Scholarpedia. **2**(7):2373. doi:10.4249/scholarpedia.2373.
13. ^ Mollo DC, Millière R (2023). "The Vector Grounding Problem." doi:10.48550/arXiv.2304.01481.
14. a, b Chalmers D (1995). "Facing up to the Problem of Consciousness." J Conscious Stud. **2**(3):200–19.
15. ^ Shi F, Chen X, Misra K, et al. (2023). "Large Language Models Can Be Easily Distracted by Irrelevant Context." International Conference on Machine Learning, ICML 2023. PMLR. pp. 31210–31227. https://proceedings.mlr.press/v202/shi23a.html.
16. ^ Jiang B, Xie Y, Hao Z, et al. (2024). "A Peek into Token Bias: Large Language Models Are Not Yet Genuine Reasoners." https://arxiv.org/abs/2406.11050.
17. ^ Mirzadeh I, Alizadeh K, Shahrokhi H, et al. (2024). "GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models." https://doi.org/10.48550/arXiv.2410.05229.
18. ^ Scheffer R, Miranda B, Koyejo S (2023). "Are Emergent Abilities of Large Language Models a Mirage?" 37th Conference on Neural Information Processing Systems (NeurIPS 2023).
19. ^ McCoy RF, Yao S, Friedman D (2024). "Embers of Autoregression Show How Large Language Models Are Shaped by the Problem They Are Trained to Solve." PNAS. **121**(41):e2322420121. doi:10.1073/pnas.2322420121.
20. ^ Prabhakar A, Griffiths TL, McCoy T (2024). "Deciphering the Factors Influencing the Efficacy of Chain-of-Thought: Probability, Memorization, and Noisy Reasoning." doi:10.48550/arXiv.2407.01687.
21. ^ Dziri N, Ximing L, Sclar M, et al. (2023). "Faith and Fate: Limits of Transformers on Compositionality." Adv Neural Inf Process Syst. **36**:70293–70332.
22. ^ Shojaee P, et al. (2025). "The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models Via the Lens of Problem Complexity." https://ml-site.cdn-apple.com/papers/the-illusion-of-thinking.pdf.
23. ^ Kambhampati S (2024). "Can Large Language Models Reason and Plan?" Ann N Y Acad Sci. **1534**:15–18. doi:10.1111/nyas.15125.
24. ^ Sacks O (1985). The Man Who Mistook His Wife for a Hat. New York: Summit Books.

25. ^Lappin S (2024). "Assessing the Strengths and Weaknesses of Large Language Models." J Log Lang Inf. **33**:9–20. doi:10.1007/s10849-023-09409-x.

26. ^Wigner EP (1960). "The Unreasonable Effectiveness of Mathematics in the Natural Sciences." Commun Pure Appl Math. **13**(1):1–14. doi:10.1002/cpa.3160130102.

27. ^Queloz M (2024). "Can Word Models Be World Models? Language as a Window Onto the Conditional Structure of the World." https://philpapers.org/rec/QUECWM.

28. ^Masi M (2024). "The Nature and Origin of Language in Abhinavagupta and Sri Aurobindo." Upcoming.

29. ^Harnad S (2024). "Language Writ Large: LLMs, ChatGPT, Grounding, Meaning and Understanding." Front Artif Intell. **7**. doi:10.3389/frai.2024.1490698.

30. ^Vallor S (2024). "The Danger Of Superhuman AI Is Not What You Think." Noema, Technology & the Human. https://www.noemamag.com/the-danger-of-superhuman-ai-is-not-what-you-think/.

31. ^Jackson F (1982). "Epiphenomenal Qualia." Philos Q. **32**:127–136. doi:10.2307/2960077.

32. ^Landau B, Gleitman LR (1985). Language and Experience: Evidence From the Blind Child. Cambridge (MA): Harvard University Press.

33. ^Senden M v (1960). "Space and Sight." Methuen, London.

34. ^Held R, Ostrovsky Y, de Gelder B, et al. (2011). "The Newly Sighted Fail to Match Seen With Felt." Nat Neurosci. **14**:551–553. doi:10.1038/nn.2795.

35. ^Herzog M (2008). "Binding Problem." In: Binder MD, Hirokawa N, Windhorst U (eds) Encyclopedia of Neuroscience. Berlin, Heidelberg: Springer. doi:10.1007/978-3-540-29678-2_626.

36. ^Else L (2010). "A Meadowful of Meaning." New Sci. **207**(2774):28–31.

37. ^Barrett FN (2024). "Experience and Nature in Pragmatism and Enactive Theory." Phenom Cogn Sci. doi:10.1007/s11097-024-10012-z.

38. ^Browning J (2024). "Personhood and AI: Why Large Language Models Don't Understand Us." AI Soc. **39**:2499–506. doi:10.1007/s00146-023-01724-y.

39. ^Virenque L, Mossio M (2024). "What Is Agency? A View From Autonomy Theory." Biol Theory. **19**:11–15. doi:10.1007/s13752-023-00441-5.

40. ^Beckmann P, Queloz M (2025). "Mechanistic Indicators of Understanding in Large Language Models." doi:10.48550/arXiv.2507.08017.

41. ^Froese T, Taguchi S (2019). "The Problem of Meaning in AI and Robotics: Still With Us After All These Years." Philosophies. **4**(2):14. doi:10.3390/philosophies4020014.

42. ^ *Ananthaswamy A (2024). "How Close Is AI to Human-Level Intelligence?" Nature. **636**:22–25. doi:10.1038/d41586-024-03905-1.*

43. ^ *Decarli G, Sella F, Lanfranchi S, et al. (2023). "Severe Developmental Dyscalculia Is Characterized by Core Deficits in Both Symbolic and Nonsymbolic Number Sense." Psychol Sci. **34**(1):8–21. doi:10.1177/09567976221097947.*

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.