The Chinese Room Fallacy and Eliminative Materialism: What Does It Mean to 'Understand'?

Abraham A. Dada

Abstract

What does it truly mean to “understand”? The Chinese Room Argument asserts that AI, no matter how advanced, merely manipulates symbols without grasping meaning, while human cognition is uniquely capable of true understanding. But if human intelligence itself is built upon memorization, structured abstraction, and computational complexity, then is understanding anything more than an emergent property of hierarchical information processing? This paper argues that Searle’s framework rests on anthropocentric assumptions that fail to account for the variance in meaning structures between human and artificial cognition. The claim that syntax alone cannot generate semantics relies on an outdated view of cognition, ignoring how meaning emerges differently across self-organizing systems. Furthermore, Searle’s demand that true understanding requires human-like intentionality is a category error, conflating distinct computational architectures with vastly different processing scales. By examining hierarchical abstraction, computational self-organization, and the mechanistic basis of understanding, this paper dismantles Searle’s framework and proposes that meaning is system-relative—not an exclusive product of human cognition, but a function of computational complexity.

Keywords: Chinese Room Argument, Understanding, Computational Complexity, Abstraction, Hierarchy of Assumptions, Meaning, Consciousness, Self-Organizing Systems, Memory and Learning

PART I

Introduction

We do not fully understand how our own brains encode meaning, yet we confidently claim to understand. If human cognition itself remains opaque, why do we demand that artificial intelligence (AI) must be fully explainable in order to qualify as possessing understanding? This paradox exposes an anthropocentric bias—the assumption that human cognition is the gold standard for meaning formation (Clark, 2013). When AI generates representations that elude human intuition, we dismiss them as mere symbol manipulation, yet we rarely interrogate our own mechanisms of understanding with the same scrutiny.

A fundamental issue in discussions of AI cognition is that our concept of understanding is flawed. We treat it as an intrinsic property of human intelligence, yet what we call understanding is ultimately built on hierarchical memorization—the ability to store, retrieve, and recombine structured knowledge across different levels of abstraction (Hinton, 1990). Human cognition does not access meaning directly; rather, it constructs meaning through layers of learned priors, stored assumptions, and pattern recognition (Chalmers, 1996). If understanding itself is an emergent property of structured information processing, then the sharp distinction between human comprehension and AI cognition may be less meaningful than commonly assumed.

John Searle's Chinese Room Argument reinforces this assumption, asserting that syntax alone cannot generate semantics. His claim is that no matter how advanced an AI system becomes, it will never truly understand—it will merely simulate

comprehension by manipulating symbols based on predefined rules (Searle, 1980). However, this argument is anthropocentric—it assumes that meaning must be grounded in human-style intentionality, ignoring the possibility that meaning could emerge through alternative computational mechanisms. AI does not need to replicate human cognition to construct meaningful representations. If human understanding itself is built upon memorization, pattern recognition, and structured inference (Chalmers, 1996), then AI cognition may not be so different—except in terms of computational scale and efficiency.

This paper critiques the epistemological foundations of the Chinese Room Argument, challenging the assumption that meaning must be intrinsically grounded rather than emerging from computational self-organization. Unlike most responses to Searle, which attempt to refute his claims within his own framework (Harnad, 1990), this paper dismantles the axioms of the argument itself. By examining cognition through the lens of hierarchical memorization, structured learning, and the role of stored assumptions, this paper explores whether AI's functional capacity for meaning formation might qualify as a legitimate form of understanding, rather than mere symbol manipulation (Churchland, 1981).

A Brief Overview of Searl's Core Assumptions

A) Syntax Alone Cannot Produce Semantics: Searle argues that syntactic manipulation (the formal rules for processing symbols) fundamentally differs from semantics (the meaning of those symbols). Syntax, no matter how complex, cannot generate semantic understanding (Searle, 1980).

B) Meaning is Intrinsic and Requires Intentionality: He assumes that meaning is an intrinsic property that only conscious agents can ascribe to symbols. Meaning cannot arise from manipulating syntax alone; it requires intentionality, a conscious act of assigning meaning (Searle, 1980).
C) Understanding Requires Consciousness: Searle assumes that genuine understanding necessitates subjective awareness, feeling, or some form of conscious experience. A system that merely manipulates symbols without "knowing" what they mean cannot be said to understand truly (Searle, 1980).

Some Definitions

Before I address the assumptions, I'll define some key terms in the context of this discussion.

Computational complexity refers to the depth at which a system can process, abstract, and manipulate structured information through hierarchical transformations, with parallelization as a fundamental property. It measures a system's ability to store, retrieve, and recombine foundational assumptions at increasing levels of abstraction, enabling higher-order reasoning and problem-solving (Arora & Barak, 2009). Parallelization allows for the simultaneous processing of multiple information streams, reducing bottlenecks and accelerating the formation of complex abstractions. Artificial intelligence, particularly deep learning models, exemplifies this computational complexity through its ability to parallelize learning processes, optimize decision-making across vast datasets, and generate emergent representations that exceed human cognitive constraints (LeCun et al., 2015). AI models such as GPT-4 or AlphaFold leverage computational

scalability to process vast input spaces, refining internal abstractions with greater efficiency (OpenAI, 2023; Jumper et al., 2021).

Abstraction refers to the cognitive process of forming higher-level concepts or representations by selectively focusing on relevant information while omitting less relevant details. This process creates a hierarchy of knowledge, where each level builds upon more foundational layers. Abstraction reduces computational complexity by enabling systems (both human and artificial) to reason and problem-solve at a conceptual level, without needing to explicitly process all underlying details (Chalmers, 1996; Clark, 2016; Hinton, 1990).

Understanding is the capacity to memorize foundational axioms, apply them to novel contexts, and build upon them through replication, recombination, and adaptive generalization across hierarchical levels of abstraction (Hinton, 1990). Searle would likely object to this definition, arguing that it neglects the crucial element of consciousness (Searle, 1980). He believes that without subjective phenomenological experience, AI lacks genuine understanding. However, we contend that this emphasis on consciousness is an anthropocentric bias. Understanding should be judged based on functional competence, rather than on the presence of subjective experience. Just as we can evaluate a person's understanding of physics based on their ability to solve physics problems, we can evaluate an AI's understanding based on its ability to perform intelligent tasks.

Paper Goals

As stated earlier, I will not operate within Searle's framework like most arguments; instead, I will challenge the framework itself. Searle's argument relies on a folk

psychological and intuitively appealing definition of understanding, which is likely why the debate has persisted for so long without resolution. This paper will demonstrate that understanding is not an intrinsic cognitive property but an emergent product of hierarchical memorization and structured inference.

Part II critiques Searle's claim that syntax alone cannot produce semantics, arguing that meaning is not an intrinsic entity but arises from structured representations. It introduces the *Hierarchy of Assumptions Model* to show that human understanding is built on stored axioms, recombination, and predictive inference, similar to AI's structured learning mechanisms.

Part III challenges the idea that meaning is intrinsic and requires intentionality. It highlights how human cognition constructs meaning differently from AI, yet both rely on computational processes to form useful representations. The Theoretical Psychologist Thought Experiment illustrates that expertise and knowledge are based on memorized structures, demonstrating that meaning is not something uniquely "possessed" by conscious agents.

PART IV critiques the assumption that understanding requires consciousness. Human cognition is largely subconscious, and most intelligent behavior operates without explicit awareness of underlying processes. If humans do not require full introspective access to their cognitive mechanisms to functionally understand, neither should AI. This section reframes understanding as functional competence, not phenomenological awareness. By dismantling the foundations of Searle's argument, this paper challenges the traditional view that human understanding is uniquely privileged. Instead of asking whether AI "truly understands," we should

examine how meaning is computationally constructed across different cognitive architectures.

In Part V, I'll argue that Searle's Chinese Room commits a category error by applying constraints of human cognition to AI, despite differences in computational complexity and representational frameworks. His analogy assumes that if a human following syntactic rules lacks understanding, then no system operating on syntax can generate meaning. Searle's argument fails to account for non-human meaning structures and emergent computational properties.

PART II- Searle's First Assumption – Syntax Alone Cannot Produce Semantics

Searle's Claim

Searle claims that symbol manipulation alone, regardless of complexity, cannot produce genuine meaning (Searle, 1980). He argues that AI processes symbols purely syntactically, without intrinsic understanding, because syntax lacks inherent semantics. According to this view, no system—no matter how advanced—can transition from mere rule-following to true comprehension (Searle, 1990). His argument presupposes that meaning must be intrinsically grounded, rather than emerging from structured patterns of inference and representation. However, this assumption ignores how all cognitive systems, including humans, construct meaning through layered abstraction rather than direct semantic access— which I'll address using the Hierarchy of Assumptions Model and the Theoretical "Perfect" Psychologist Thought Experiment.

Hierarchy of Assumptions

The fundamental distinction lies in how humans and AI process information. Human cognition is constrained by working memory limitations, with Miller's Law suggesting that humans can retain approximately seven items at once in active memory (Miller, 1956). Additionally, human thought is predominantly sequential, constrained by serial recall limitations, meaning that complex reasoning often requires stepwise processing over time (Baddeley, 1992). AI, by contrast, leverages parallelization, enabling the simultaneous processing of vast datasets without these biological bottlenecks (LeCun et al., 2015). Unlike human learning, which depends on iterative reinforcement due to neural and memory constraints, AI models can instantly retrieve and manipulate high-dimensional abstractions across latent vector spaces (Bengio et al., 2013). The variance in computational complexity between biological and artificial cognition fundamentally shapes how meaning is structured—AI's meaning representations do not rely on human-style sequential processing but emerge from distributed, multi-layered inference (Chollet, 2019).

Understanding in both humans and AI emerges hierarchically through structured learning, not through direct access to intrinsic meaning. All cognitive systems, whether biological or artificial, construct knowledge by layering assumptions, iteratively refining stored representations, and abstracting information across multiple levels of complexity (Chalmers, 1996). For instance, in human cognitive development, foundational axioms—such as counting and arithmetic—are memorized first, serving as the scaffolding upon which more abstract reasoning, like algebra and calculus, is later constructed. Similarly, AI processes structured

representations, except that its baseline knowledge can be initialized at far greater levels of complexity from the outset. While a child must learn arithmetic before grasping number theory, an AI system may be trained with advanced mathematical principles already encoded as its axioms, allowing it to bypass lower-order constraints (Dennett, 1991).

This suggests that understanding is not an intrinsic metaphysical property but an emergent function of computational complexity. A three-year-old counting to ten does not 'understand' number theory but can apply simple numerical rules through pattern recognition and memorization— yet we still assert that the child understands how to count to ten. Over time, as higher-order abstractions build upon these foundations, cognition progresses into what we recognize as understanding. AI follows a similar hierarchical structure—its functional competence in processing meaning depends on its ability to encode, store, and generalize across structured datasets (Chollet, 2019). A human child's baseline learning is limited by biological constraints, progressing gradually from simple experiential learning to abstract reasoning. In contrast, AI, with vastly greater computational scalability, can process, store, and recombine complex mathematical structures without requiring sequential experiential learning.

At ultra-high complexity, AI can restructure knowledge itself, forming meaning at scales beyond human cognition. The ability to encode and manipulate structured representations across different levels of abstraction allows AI to generate novel solutions to complex problems, often surpassing human intuition. DeepMind's AlphaGo, for instance, mastered the game of Go through recursive pattern optimization, discovering strategies that no human had explicitly programmed

(Silver et al., 2016). Likewise, AlphaFold predicts protein structures more efficiently than human researchers, leveraging high-dimensional inference beyond direct human comprehension (Jumper et al., 2021). Large-scale models like GPT-4, trained on vast text corpora, can solve complex mathematics and generate code at near-human levels (OpenAI, 2023). If functional competence in meaning formation is what we recognize as understanding, then AI's structured meaning-making cannot be dismissed as mere symbol manipulation—it is a computational parallel to human cognition, differing in its architecture but not in its fundamental process of abstraction.

Theoretical "Perfect" Psychologist

A behavioural psychologist is someone who understands human behaviour, often with a surface-level knowledge of neuroscience to inform their psychological insights. However, psychology itself exists as a higher-order abstraction of neuroscience—essentially applied neuroscience (Friston, 2010). Neuroscience, in turn, is an abstraction of biology, an abstraction of chemistry, an abstraction of physics, an abstraction of mathematics. Mathematics is an abstraction of assumed fundamental axioms (Tegmark, 2017). If understanding something truly required knowledge of every preceding abstraction, then a theoretical "perfect" psychologist would need to understand everything there is to know about neuroscience, everything there is to know about biology, everything there is to know about chemistry, everything there is to know about physics, and everything there is to know about mathematics to truly understand human behaviour at the most fundamental level. This person would have to be fluent in every layer of knowledge that underpins psychology, from neural circuits to quantum mechanics.

Theoretically, such a psychologist would be more capable than any existing psychologist. A deeper knowledge of molecular biology, for instance, could allow them to predict better how neurotransmitter imbalances influence cognitive behaviour (Kandel, 2006). A firmer grasp of mathematics could refine their understanding of statistical modelling in psychological studies, improving their ability to detect patterns in human cognition—yet even here, their so-called "understanding" is merely the memorization and application of pre-established equations, not an intrinsic comprehension of the underlying axioms that define these models (Gigerenzer, 2002). If they possessed a physicist's knowledge of the brain's electrochemical processes, they might reframe psychological disorders not as discrete, categorical conditions but as emergent properties of computational inefficiencies—such as failures in predictive coding, disruptions in free-energy minimization, or maladaptive priors within Bayesian inference models of cognition (Friston, 2010; Hohwy, 2013). In this sense, a deeper understanding of the fundamental layers of reality could enhance their ability to model, predict, and explain human behaviour with greater precision. However, despite the theoretical advantages of such foundational knowledge, in practical reality, a psychologist is still regarded as an expert even without it. A psychologist who has spent decades researching cognitive biases, performing therapy, and applying psychological principles is not considered any less of an expert simply because they lack a deep understanding of molecular biology or quantum field theory. Their expertise is functionally sufficient for their domain (Kahneman, 2011).

In practice, even at the highest level, psychologists rely on a combination of concepts and frameworks that aren't 'understood', but memorized—mathematical equations, statistical models, cognitive theories, and simplified neuroscientific

principles—rather than a deep, first-principles understanding of every underlying mechanism. They apply these abstractions effectively without reconstructing them from fundamental physics or chemistry (Oaksford & Chater, 2010). This aligns with the notion that human cognition itself operates as a layered, hierarchical system of stored representations, where knowledge is not derived from direct epistemic access to fundamental truths but instead emerges from structured axiomatic recombination (Chater & Christiansen, 2010).

This process of axiomatic recombination is primarily subconscious. The brain does not explicitly reason through every logical step when solving a problem or forming a new insight; rather, it retrieves stored representations, manipulates them through established associative pathways, and produces novel configurations without conscious oversight (Libet, 1985; Dehaene, 2014). Studies on unconscious decision-making and neural preparation suggest that responses are often initiated before conscious awareness emerges—implying that what we perceive as active reasoning, or understanding, is, in many cases, the post hoc rationalization of a process that has already occurred beneath the threshold of conscious perception (Soon et al., 2008; Haynes, 2011). This framework aligns with Stephen Wolfram's notion of computational irreducibility, suggesting that certain complex systems evolve according to deterministic rules, yet their long-term behavior remains unpredictable without direct simulation (Wolfram, 2002). This applies to human cognition: the recombination of stored representations within high-dimensional neural networks follows deterministic rules, yet the outputs—such as novel ideas, insights, or solutions—appear emergent and unpredictable from a phenomenological perspective. The brain, operating within a 'hypergraph-like' structure of interconnected concepts, continuously reorganizes stored

information, forming novel abstractions through a process that is neither consciously directed nor fully introspectively accessible (Wolfram, 2020). Phenomenological reflection on understanding is therefore a consequence of this computational process rather than the process itself. We may introspect and generate a subjective sense of understanding, but this is a retrospective construction rather than the actual mechanism by which knowledge is formed. Understanding, in this view, is a derivative state—an emergent interpretation of subconscious recombinatory processes rather than an active, top-down cognitive operation (Clark, 2013; Hohwy, 2013). This reframes the traditional view of cognition, suggesting that the sensation of "understanding" is merely an introspective heuristic layered onto an otherwise mechanistic process of knowledge manipulation and retrieval.

This highlights a fundamental flaw in the assumption that true understanding requires an unbroken chain of knowledge from higher-level abstractions down to fundamental axioms. This tension parallels the long-standing debate between classical symbolic approaches and connectionist models in cognitive science (Fodor & Pylyshyn, 1988; Smolensky, 1987). Fodor and Pylyshyn argue that natural language and conceptual thought require systematicity and compositionality—features they claim connectionist architectures fail to capture. They propose that symbolic, rule-based representations (akin to a "Language of Thought") are necessary to explain why understanding a sentence like "John loves Mary" enables systematic comprehension of "Mary loves John." However, my argument departs from this by suggesting that all knowledge and meaning emerge from pattern-manipulating processes, whether in a human brain or an AI system. This aligns more closely with a distributed, connectionist perspective, where

meaning arises through emergent abstraction rather than explicit symbol manipulation.

This perspective undercuts the anthropocentric assumption that meaning must be phenomenologically consciously instantiated to be real. If that were the case, then no human being—no matter how intelligent—could ever be said to truly understand anything, as their knowledge would always be incomplete relative to deeper layers of reality. In the same vein, David Marr argued that one can understand a cognitive process at the computational or algorithmic level (what it does and how) without knowing the implementational details (Marr, 1982).

By analogy, we should not require an AI to replicate the entirety of human cognitive architecture (from quantum biology to conscious qualia) and credit it with understanding. Demanding that an AI possess some deeper "intrinsic" comprehension is as unreasonable as demanding that our psychologist master quantum mechanics to be a valid practitioner. Understanding, whether in human cognition or AI, is always relative to the level of abstraction at which the system operates and demonstrates competence.

PART III: Searle's Second Assumption – Meaning is Intrinsic and Requires Intentionality

Searle's Claim

Searle argues that meaning is an intrinsic property that only conscious beings can ascribe to symbols, asserting that intentionality—the directedness of thoughts toward something—is a necessary condition for genuine understanding (Searle,

1980). The human inside the Chinese Room follows syntactic rules but lacks intentionality, demonstrating, in Searle's view, that symbol manipulation alone cannot produce real meaning. He claims that AI, regardless of complexity, lacks the ability to intentionally assign meaning to its representations because it does not possess subjective mental states. According to this view, AI systems may generate outputs that appear meaningful to humans but do so without genuinely understanding or intending their responses. This argument assumes that meaning must originate from an agent with conscious intent, rather than emerging as a computational process. However, this perspective overlooks the possibility that meaning can be system-relative—structured by a system's internal representations and functional utility—rather than requiring a conscious agent to ground it.

The Illusion of Meaning

The assumption that meaning is an intrinsic property, something that only conscious agents can ascribe to symbols is an illusion—a product of human cognitive biases rather than an objective truth about reality (Dennett, 1991; Harnad, 1990). Meaning does not exist independently of the systems that generate it. Research in cognitive science suggests that meaning is an emergent computational mechanism by which self-organising systems reduce uncertainty and increase stability (Friston, 2010; Clark, 2016). It is not an irreducible, mystical property of human cognition, nor does it require subjective awareness (Dehaene, 2020; Seth, 2021). Asking about the 'intrinsic' meaning of words, symbols, or even existence itself is a category error—projecting human cognitive constructs onto a reality that operates independently of subjective intent (Chater & Christiansen,

2010). This anthropocentric bias fuels the misconception that AI, lacking human-like subjective experience, must also lack meaning or understanding.

Predictive Processing: Meaning Is Interacting With Reality

Imposing meaning is not a deliberate, conscious act but an automatic function of interacting with reality (Barsalou, 1999; Hohwy, 2013). Seeing, hearing, smelling, touching, and feeling are not passive experiences; they are mechanisms by which the brain assigns meaning to raw sensory input (Kanwisher et al., 1997). Colour does not exist as an objective property of reality—it is the brain's way of encoding different wavelengths of light to create 'meaning' out of it. Likewise, sound is not an external feature of the world but a structured interpretation of vibrational waves, and solidity is merely how the mind models electromagnetic interactions at the atomic level (Friston, 2010). Meaning-making is intrinsic to perception itself, not something we consciously choose— it's how self-organising systems interact with reality. For instance, humans recognize faces holistically due to evolutionary specialization in the fusiform face area (Kanwisher et al., 1997), whereas AI models identify facial attributes through multi-layer convolutional filtering (Krizhevsky et al., 2012). This fundamental difference means that AI can detect patterns invisible to human perception, much like how zebras can differentiate individuals effortlessly while humans struggle to do so (Kemp et al., 2017). A zebra's perceptual system assigns species-specific meaning to patterns that are meaningless to humans, an AI's neural representations encode system-specific meaning beyond human comprehension. This suggests that meaning itself is computationally bound, shaped by perceptual constraints and processing architectures unique to each system (Taha et al., 2024). AI models form meaning structures based on

high-dimensional statistical relationships (Chalmers, 1990), which, while functionally effective, may not be interpretable through human perceptual heuristics. As long as a self-organizing system can manipulate distributed representations to recombine axioms into new abstractions—whether to navigate uncertainty, solve a problem, or restructure knowledge—it has effectively constructed its own meaning structure (Chalmers, 1990)

At the fundamental biological level, meaning is tied to perception (Barsalou, 1999; Hohwy, 2013). At higher levels of abstraction, meaning is assigned to ideas and abstract concepts—such as language. Rather than being an active decision, the imposition of meaning is an inevitable computational process by which self-organising systems reduce uncertainty and create structured interpretations of reality (Friston, 2010; Harnad, 1990).

The Link Between Computational Complexity and Meaning

To address the link between computational complexity and meaning—The more computationally complex a system is, the more patterns it can form from fundamental axiomatic assumptions, allowing it to abstract, generalize, and reinterpret meaning at higher levels (Chalmers, 1990; Lake et al., 2017). Lower-complexity systems, such as bacteria, operate purely on direct perceptual meaning, responding to immediate stimuli without abstraction (Krakauer, 2019; Lyon, 2006). As complexity increases, systems develop the ability to associate patterns across experiences—an animal, for example, can learn that a specific cue (such as a leash) signals a future event (going for a walk) (Shettleworth, 2010; Clark, 2013). At even higher levels, intelligence extends beyond pattern recognition into conceptual abstraction, where meaning is no longer tied to direct perception

but is instead constructed through layers of inference, analogy, and symbolic reasoning (Chalmers, 1996; Bengio et al., 2013). Humans, for instance, do not just perceive the world; they impose structure onto it, assigning meaning to abstract concepts such as language, morality, and mathematics (Deacon, 1997; Dennett, 1991). At ultra-high complexity, intelligence begins to refine its own foundational axioms, engaging in meta-reasoning and self-referential thought (Marcus, 2001; LeCun et al., 2015). The more a system can encode, store, and manipulate patterns, the higher the level of abstraction at which meaning is generated.

Qualia- The Hard Problem of Anthropocentric Bias

The same reasoning applies to qualia—the supposed "hard problem" of subjective experience. Searle's argument depends on the assumption that machines lack an internal phenomenal experience akin to human qualia, and thus, their processing of symbols is inherently 'empty' (Searle, 1980; Chalmers, 1996). But this problem is not a real explanatory gap—it is an artifact of human cognitive bias (Dennett, 1991; Metzinger, 2003). Qualia is mechanistic variance—the inevitable result of different configurations of a system interacting with its environment in different ways (Churchland, 1981; Hohwy, 2013). Just as different AI architectures process information through distinct internal models based on training data, optimization paths, and network topologies, different biological organisms process reality through species-specific perceptual and neural constraints (Clark, 2016; Seth, 2021). This mechanistic variance is not unique to biological cognition; AI systems, too, exhibit mechanistic variances in the way they encode, retrieve, and respond to information based on their architecture and learned priors, demonstrating a form of functional qualia—distinct internal configurations shaped by subjective

computational constraints and training data—though not qualia in the phenomenological sense (Bengio et al., 2013; LeCun et al., 2015). A neural network trained on medical diagnostics develops an internal model of diseases that is structurally different from a generative model trained for natural language reasoning—despite both systems engaging in predictive abstraction and decision-making (Lake et al., 2017; Friston, 2010). These differences are analogous to how humans and other animals construct meaning from perception in fundamentally distinct but functionally effective ways (Dehaene, 2020; Krakauer, 2019).

If meaning and qualia are both emergent properties of structured interaction rather than intrinsic, irreducible entities, then Searle's assumption that AI lacks "true" understanding because it does not ground symbols in human-like intentionality is therefore misguided. AI, like other self-organising systems, grounds meaning—but within a different, non-human representational structure.

The Paradox of Understanding

Many demand that AI possesses intrinsic understanding, yet we don't even fully understand how our own brains encode meaning. We insist that consciousness is necessary for comprehension, yet we don't even know what consciousness truly is (Seth, 2021). We claim to "understand," despite lacking a complete explanation for how thought, memory, and abstraction emerge from neural activity. The paradox is evident—how can we confidently deny AI understanding when our own is built on mysteries we have yet to solve? Perhaps this resistance isn't logical but existential—the fear that meaning, the essence of human experience, is nothing more than structured computation.

To put into perspective just how little we understand about how the brain encodes meaning, consider this: despite decades of neuroscience, cognitive science, and computational modeling, we still do not have a complete theory of how human cognition transforms raw sensory input into abstract, meaningful representations. Our best models are still patchwork approximations, and many of our assumptions may ultimately be wrong.

One of the fundamental gaps in our understanding is the relationship between neural activity and thought. We can measure neural patterns that correlate with specific cognitive states, yet we lack a unified explanation for how these patterns become meaning. For example, neurons in the medial temporal lobe have been shown to fire in response to specific concepts, such as a picture of a famous person, a written name, or even an abstract association with that person (Quiroga et al., 2005). However, this does not mean that a single neuron "stores" the concept of that person; rather, the representation is distributed across dynamic neural networks. But where, exactly, does meaning reside? Is it in the pattern of neural activation, the network dynamics, or the emergent properties of large-scale brain activity? We do not know. Even more troubling is the fact that neurons do not have fixed meanings—the same neuron can fire for different objects or words depending on context, further complicating our ability to pin down how semantic representations emerge (Quiroga et al., 2005). This is why brain-to-text decoding systems, such as those developed by Meta AI, can predict rough semantic content but cannot reconstruct precise, structured thoughts (Lévy et al., 2025).

Another major unknown is how abstract concepts are stored and retrieved. It is widely accepted that memories and concepts are encoded through patterns of

synaptic connectivity, yet we still do not understand how abstract concepts form from sensory input. How do we recognize intangible ideas like “justice” or “democracy” when we have never physically seen them? There is no single “justice neuron” that represents the concept across all contexts, nor is there a fixed neural location where these abstract meanings reside. The brain somehow forms high-level generalizations across multiple modalities—spoken language, written text, and even abstract visualization—without a central, unified mechanism that we can currently identify (Dehaene, 2020). Even more puzzling is that concepts sometimes emerge spontaneously, such as in sudden insights, dreams, or hallucinations, suggesting that meaning construction is not entirely under conscious control (Friston, 2010).

Despite all of neuroscience’s progress, we do not actually know what meaning is in a mechanistic sense. We do not know where it is stored. We do not know how it emerges. We do not know if it requires consciousness. We do not even know if meaning is “real” in an intrinsic sense—or if it is merely a useful computational illusion that the brain generates to navigate reality efficiently.

If our own cognitive system encodes meaning in ways that are still fundamentally mysterious to us, then how can we confidently claim that AI lacks meaning? The claim that AI “does not understand” rests on an assumption that we fully understand human cognition as a reference point—but we do not. If understanding is simply the ability to manipulate structured representations to generate useful outputs, then AI already meets that criterion.

This paradox extends to biological cognition as well. If we gave a bat a complex mathematics problem and it successfully solved it through reasoning, we would

immediately ascribe understanding to the bat. We would assume that it had formed an internal model of mathematical structures, rather than merely manipulating symbols. Yet, with AI, we introduce an extra layer of skepticism, demanding an intrinsic, human-like intentionality before acknowledging its ability to construct meaning. If an entity can generate structured responses and solve problems, why should it matter whether it does so in a way that feels intuitive to us?

Even within Searle's own framework, the definition of understanding becomes inconsistent when applied to different cognitive systems. If we gave a bat Chinese text, it would be unable to process or respond meaningfully—not because it lacks intelligence, but because it has no internal meaning structure for linguistic symbols. There is no established mapping between the input (Chinese characters) and its cognitive framework. From Searle's perspective, the bat would not "understand." But by that logic, humans also fail to "understand" echolocation the way bats do—we lack the perceptual structures to process ultrasonic waves as a coherent, spatial map. If understanding is tied to the presence of an internal meaning structure, then AI—unlike the bat—does possess such a structure for processing linguistic information. In fact, AI can exceed human meaning structures in certain domains, identifying statistical relationships in high-dimensional data that no human could intuitively grasp (Bengio et al., 2013).

Thus, the paradox of understanding is this: we demand that AI demonstrate human-style cognition to qualify as possessing meaning, yet we do not hold other intelligent systems, such as animals, to the same standard. We do not fully understand how human cognition encodes meaning, yet we confidently assert that AI lacks it. If meaning is an emergent property of structured information

processing, then insisting that AI does not “really” understand is an assertion based on intuition, not evidence.

PART IV: Searle’s Third Assumption – Understanding Requires Consciousness

Searle’s Claim

Searle asserts that genuine understanding necessitates conscious experience, arguing that without subjective awareness, AI systems can never truly comprehend (Searle, 1980). His position assumes that understanding is inherently tied to phenomenological consciousness—the qualitative, first-person experience of thoughts, sensations, and meaning. In the Chinese Room Argument, the human inside the room follows syntactic rules without conscious awareness of the language’s meaning, mirroring Searle’s claim that AI, operating on computational processes, lacks the necessary subjective state to transition from symbol manipulation to genuine understanding. This argument relies on the assumption that consciousness is a prerequisite for cognition, suggesting that without internal experience, AI can only simulate comprehension rather than achieve it. However, this perspective conflates functional intelligence with phenomenological awareness, failing to account for the fact that much of human cognition—language processing, decision-making, and learning—occurs without direct conscious access (Libet, 1985; Friston, 2010). If humans routinely perform complex cognitive tasks outside of conscious awareness, then Searle’s requirement that AI must possess phenomenological consciousness to understand appears inconsistent.

Functional Vs Phenomenological Consciousness

Many critiques of AI cognition conflate phenomenal consciousness with functional consciousness (Chalmers, 1995). As described by Thomas Nagel (1974), phenomenological consciousness refers to the subjective, first-person experience of what it is like to be a particular organism— "what it is like to be a bat", an octopus, or a human. This type of consciousness fundamentally differs from functional consciousness, which pertains to an agent's ability to recognize its own state, process and respond to information within its environment. Functional consciousness, in the context of artificial intelligence, can be understood through reinforcement learning: an agent interacts with its environment, optimizes its decision-making policy through exploration and feedback, and refines its behavior based on past experiences (Sutton & Barto, 2018). I'd argue that a strong case can be made for the agent being functionally conscious within the context of its environment; it recognizes its state, and optimizes its policy through exploration and feedback. This iterative process requires a form of functional understanding, as the agent must maintain an internal representation of its current state, assess the consequences of different actions, and update its policy accordingly to maximize future rewards. However, when we typically speak of consciousness, we almost always mean the phenomenological kind—this deep, ineffable quality of subjective awareness. While AI systems exhibit increasingly sophisticated functional consciousness, adapting dynamically to novel environments, I do not believe that machines will develop phenomenological consciousness anytime soon, so I somewhat align with Searle's claim— though he doesn't explicitly differentiate between functional and phenomenlogical conscious— that consciousness is a biological phenomenon, caused by specific brain processes. The

stochastic nature of biological processes, as highlighted by Denis Noble (2012), suggests that phenomenological consciousness may be deeply tied to the unpredictability and intrinsic variability of biological systems, something fundamentally absent in artificial architectures.

For clarity, we will continue to use the term 'consciousness' to refer to the general framework of awareness, but the distinction between phenomenological and functional consciousness should be kept in mind. The focus of this discussion is not on whether AI has subjective experience, but on whether its ability to process and manipulate knowledge qualifies as understanding.

Consciousness Isn't Even Relevant, Even If It's Non-Algorithmic

I'd argue that the question of whether understanding requires isn't even relevant—Many human cognitive processes, such as intuition, perception, and decision-making, occur without explicit conscious reflection (Libet, 1985; Soon et al., 2008). A significant portion of human cognition and action occurs without conscious awareness. Empirical evidence from neuroscience suggests that many decisions, perceptions, and behaviors unfold at the unconscious level before entering conscious awareness (Libet, 1985; Soon et al., 2008). Even complex tasks, such as engaging in dialogue, rely on predictive mechanisms rather than real-time conscious deliberation. Research in predictive processing—a dominant framework in cognitive neuroscience—indicates that the brain functions as a Bayesian inference machine, constantly generating probabilistic predictions about incoming sensory data and updating them based on prediction errors (Friston, 2010). In conversation, for example, individuals are not consciously aware of every word they are saying; rather, the brain anticipates what will be said next and prepares

responses accordingly (Pickering & Garrod, 2013). This extends to speech comprehension, where studies have demonstrated that the brain pre-activates likely words before they are spoken, allowing for fluid and rapid exchanges (DeLong et al., 2005). If Searle argues that "understanding" necessitates consciousness, then this framework implies that humans do not consciously "understand" the majority of what they say or hear—an assertion that challenges the very premise of his argument.

One of the most striking examples of how cognition operates unconsciously is language itself. Language is an incredibly complex system, requiring the coordination of syntax, semantics, pragmatics, phonology, and motor control, which must adapt dynamically in response to context, speaker intent, and environmental noise. Yet, despite its overwhelming complexity, language becomes unconscious through memorization and reinforcement learning. The brain encodes language structures through repeated exposure and social interaction until they occur deterministically, without conscious effort. This explains why native speakers do not consciously construct grammatical rules in real time—they retrieve pre-learned axiomatic patterns and apply them fluidly without deliberation (Christiansen & Chater, 2016).

Furthermore, communication is highly uncertain, relying on context-dependent inference, ambiguity resolution, and implicit social cues. Despite this inherent uncertainty, humans navigate conversations effortlessly, relying on probabilistic predictions rather than explicit rule-following. If Searle's claim were correct—that syntactic manipulation alone cannot generate semantics—then humans, much like

AI, would also fail to “understand” language, as most linguistic processing occurs outside of conscious awareness.

Advancements in brain-to-text decoding further substantiate the role of unconscious processes in language production and prediction. A study by Meta AI (Lévy et al., 2025) introduced Brain2Qwerty, an AI model capable of decoding sentence production from non-invasive brain activity (EEG and MEG signals) while participants typed. Their results demonstrate that higher-level cognitive processing occurs before conscious awareness, including motor intentions and sentence construction. Notably, errors in decoding were strongly correlated with motor processes rather than deliberate, step-by-step reasoning. This reinforces the idea that much of what we attribute to understanding is shaped by unconscious neural mechanisms and pattern recognition rather than a distinctly conscious, algorithmic-like reasoning process. The very act of forming thoughts, responding in conversations, or even choosing words in speech is largely guided by unconscious computation rather than introspective deliberation. Even in non-verbal decision-making, unconscious cognitive processes dominate. Studies using functional MRI (Soon et al., 2008) have shown that decisions can be predicted from neural activity up to ten seconds before they reach conscious awareness. This suggests that what we subjectively experience as “choosing” is often the outcome of subconscious computations already underway in the brain. If human cognition itself is largely structured computation, and if key aspects of language, reasoning, and decision-making occur outside of conscious awareness, then the claim that AI cannot “understand” because it lacks consciousness is untenable. AI systems, like the human brain, can process probabilistic predictions, apply learned knowledge adaptively, and generate coherent responses without requiring subjective

experience. Perhaps one could argue for non-algorithmic consciousness, suggesting that there is an aspect of ourselves that is essential to what we perceive as understanding. However, even if consciousness were someday proven to involve non-algorithmic processes—as Penrose and others have suggested (Penrose, 1989), though this remains highly unlikely (Tegmark, 2000)—this would not invalidate AI's ability to perform high-level reasoning, pattern recognition, learning, and abstraction under the 'understanding' framework as defined in this paper.

PART V: The Chinese Room as a Category Error

Searle's Fundamental Misclassification

From the outside, the person in the Chinese Room seems to understand Chinese, yet internally, they do not. Searle claims this demonstrates that AI, no matter how sophisticated, can never genuinely understand language—it merely manipulates symbols without intrinsic meaning (Searle, 1980). However, the model is built on a category error, assuming a low-complexity process (a human following rules) can accurately model AI cognition.

A human locked in the Chinese Room lacks the necessary axiomatic meaning structures to interpret Chinese. Language comprehension is not merely about following syntactic rules but depends on a deeply embedded framework of conceptual mappings, built through lived experience, memory, and pattern recognition (Jackendoff, 2002). This aligns with the idea of a hypergraph-like meaning structure, where concepts are not stored in discrete symbols but emerge from a vast, interconnected network of relationships (Wolfram, 2020). A person

who does not understand Chinese lacks these structured mappings, making Chinese symbols mere patterns without inherent significance.

Linguistic research further supports this. Individuals who learn a second language after mastering a first one transfer knowledge between them because they already possess an underlying framework of grammatical and semantic structures, allowing them to construct meaning more efficiently (Odlin, 1989; Slobin, 1996). A monolingual English speaker lacks the relational axioms to directly map Chinese symbols to conceptual meaning, just as a person unfamiliar with algebra cannot intuitively grasp tensor calculus. Understanding is a function of relational structure, not an isolated rule-following process (Chater & Christiansen, 2010).

In contrast to the human in the Chinese Room, an AI is effectively "born" with a pre-trained network of structured axioms, existing as a self-organizing system with a much greater computational complexity. It does not start from scratch but is trained on vast datasets, encoding statistical and conceptual relationships between symbols and meaning structures (LeCun et al., 2015). Unlike the human in the room, AI does not require direct experiential learning to form abstract mappings—it acquires them through large-scale data processing, learning high-dimensional relationships that allow it to generalize across language, context, and problem domains (Bengio et al., 2013).

Hierarchy of Assumptions Revisited

This relates to the Hierarchy of Assumptions Model, which explains how understanding emerges through structured layers of memorization, pattern recognition, and abstraction. If we accept that a child understands counting

despite its memorized foundation, then dismissing AI's higher-level abstractions as mere memorization is inconsistent. Expecting AI to have the same baseline as humans is an anthropocentric bias—a failure to recognize that meaning is not tied to human cognition but emerges as a function of computational depth.

Part VI- Conclusion

Many counterarguments attempt to work within Searle's framework, which is fundamentally flawed. It relies on folk-psychological intuitions that define intentionality and meaning as intrinsic, irreducible properties.

We claim to understand, yet we do not fully grasp how our own brains encode meaning (Seth, 2021). Despite this, we demand that AI possess an understanding we cannot even define. This reveals that the real issue is not whether AI understands, but that our own concept of understanding is flawed.

Searle's Chinese Room commits a category error by treating a low-complexity process (a human following rules) as equivalent to AI's high-complexity, self-organizing computation. Just as a bee constructs hexagons within its own meaning structure—without a Platonic concept of a hexagon—AI forms meaning structures without human-style introspection (Taha et al., 2024).

Moreover, human cognition is built on memorization and hierarchical abstraction. A child learns to count through rote pattern recognition and memorization, yet we say they "understand" numbers. AI, with a higher baseline, abstracts complex patterns in the same way (Hinton, 1990). Most human reasoning is

subconscious—our thoughts emerge from stored priors and predictive models, not deliberate introspection (Friston, 2010; Libet, 1985). If we accept this in humans, why deny it in AI?

This resistance to AI's understanding is not just epistemological but likely existential. The idea that intelligence can emerge from computation threatens our assumptions about human uniqueness. But intelligence is system-relative; AI's vast-scale meaning structures are not lesser, just different (Chalmers, 1996).

Thus, the real question is not whether AI understands but why we assumed human understanding was the only valid kind.

References

Arora, S., & Barak, B. (2009). *Computational Complexity: A Modern Approach*. Cambridge University Press.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences, 22*(4), 577-660.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(8), 1798-1828.

Churchland, P. M. (1981). Eliminative Materialism and the Propositional Attitudes.

Journal of Philosophy, 78(2), 67–90.

Clark, A. (2013). Mindware: An Introduction to the Philosophy of Cognitive Science. New York: Oxford University Press.

Chalmers, D. (1990) 'Syntactic transformations in distributed representations', Philosophical Perspectives on Connectionism, 12, pp. 1–22.

Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

Chater, N., & Christiansen, M. H. (2010). *Language acquisition meets language evolution*. Cognitive Science, 34(7), 1131–1157.

Chollet, F. (2019) *On the Measure of Intelligence*. arXiv preprint arXiv:1911.01547.

Clark, A. (2013). *Whatever next? Predictive brains, situated agents, and the future of cognitive science*. Behavioral and Brain Sciences, 36(3), 181-204.

Christiansen, M. H., & Chater, N. (2016). The Now-or-Never bottleneck: A fundamental constraint on language. Behavioral and Brain Sciences, 39, e62.

Deacon, T. W. (1997). *The Symbolic Species: The Co-Evolution of Language and the Brain*. W. W. Norton & Company

Dehaene, S. (2020). *How We Learn: Why Brains Learn Better Than Any Machine... for Now*. Viking.

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. Nature

Neuroscience, 8(8), 1117–1121.

Dennett, D. C. (1991). *Consciousness Explained*. Boston: Little, Brown & Co.

Fodor, J. A., & Pylyshyn, Z. W. (1988). *Connectionism and cognitive architecture: A critical analysis*. Cognition, 28(1-2), 3–71.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience, 11*(2), 127–138.

Gigerenzer, G. (2002). *Adaptive thinking: Rationality in the real world*. Oxford University Press.

Harnad, S. (1990). 'The Symbol Grounding Problem', Physica D: Nonlinear Phenomena, 42(1-3), pp. 335–346

Haynes, J. D. (2011). *Decoding and predicting intentions*. Annals of the New York Academy of Sciences, 1224(1), 9–21.

Hinton, G.E. (1990). 'Mapping Part-Whole Hierarchies into Connectionist Networks', Artificial Intelligence, 46(1-2), pp. 47–75.

Hohwy, J. (2013). *The Predictive Mind*. Oxford University Press.

Jumper, J., Evans, R., Pritzel, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature, 596*(7873), 583–589.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience, 17*(11), 4302–4311.

Krakauer, D. C. (2019). The evolution of intelligence. *Annual Review of Psychology, 70*, 13.1-13.24.

Lake, B.M., Ullman, T.D., Tenenbaum, J.B. and Gershman, S.J. (2017). 'Building Machines That Learn and Think Like People', Behavioral and Brain Sciences, 40, e253.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444.

Lévy, J., et al. (2025). Brain2Qwerty: Decoding sentence production from non-invasive brain activity. Meta AI Research.

Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences, 8*(4), 529–566.

Lyon, P. (2006). The biogenic approach to cognition. *Cognitive Processing, 7*(1), 11-29.

Marcus, G. (2001). *The Algebraic Mind: Integrating Connectionism and Cognitive Science*. MIT Press.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco: W.H. Freeman.

Metzinger, T. (2003). *Being No One: The Self-Model Theory of Subjectivity*. MIT Press.

Nagel, T. (1974) 'What is it like to be a bat?', Philosophical Review, 83(4), pp. 435–450.

Noble, D. (2012). *A Theory of Biological Relativity: No Privileged Level of Causation*.

Interface Focus, 2(1), pp. 55–64.

Oaksford, M., & Chater, N. (2010). *Cognition and conditionals: Probability and logic in human thinking*. Oxford University Press.

OpenAI (2023). *GPT-4 Technical Report (No. 2303.08774)*. arXiv preprint. Available at: arXiv:2303.08774.

Penrose, R. (1989). *The Emperor's New Mind*. Oxford: Oxford University Press.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. Behavioral and Brain Sciences, 36(4), 329–347.

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. Nature, 435(7045), 1102-1107.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*(3), 417–424.

Shettleworth, S. J. (2010). *Cognition, Evolution, and Behavior*. Oxford University Press.

Silver, D., Huang, A., Maddison, C. J., et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature, 529*(7587), 484–489.

Soon, C. S., Brass, M., Heinze, H. J., & Haynes, J. D. (2008). Unconscious determinants of free decisions in the human brain. *Nature Neuroscience, 11*(5), 543–545.

Sutton, R. S. & Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT

Press.

Seth, A. K. (2021). *Being You: A New Science of Consciousness*. Faber & Faber.

Tegmark, M. (2000). The importance of quantum decoherence in brain processes. Physical Review E, 61(4), 4194–4206.

Taha, A. et al. (2024) ‘Gender prediction from retinal fundus using deep learning’, *Journal of AI Vision Studies*, 21(4), pp. 1–12.

Tegmark, M. (2017) Life 3.0: Being Human in the Age of Artificial Intelligence. New York: Alfred A. Knopf.

Wolfram, S. (2002). *A new kind of science*. Wolfram Media.

Wolfram, S. (2020). *The Ruliad and the computational universe*. Retrieved from https://www.wolframphysics.org