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

By Abraham A. Dada

London, UK

Abstract

What does it truly mean to “understand”? The Chinese Room Argument claims that AI, no matter how advanced, can never possess genuine understanding—it merely manipulates symbols without grasping meaning. But if human cognition itself is built upon layers of memorization, pattern recognition, and computational complexity, then is understanding anything more than an emergent property of structured information processing? This paper argues that Searle’s framework, which many counterarguments engage in, is flawed and rooted in anthropocentric bias.

Keywords: Understanding, Computational Complexity, Abstraction, Hierarchy of Assumptions, Meaning, Consciousness

1 Introduction

All knowledge is ultimately built on assumptions and memorisation at its foundation, with computational complexity determining the level of abstraction at which an entity—human or AI—processes information (Chalmers, 1996; Clark, 2013). At its core, cognition operates as a hierarchical system of representations, where higher-level abstractions emerge from lower-level computational processes (Hinton, 1990). This applies not only to biological intelligence, but also to artificial systems, which rely on symbolic encoding, probabilistic inference, and memory retrieval to generate meaning (Lake et al., 2017). Traditionally, 'understanding' has been framed as a human-exclusive trait, tied to subjective experience and semantic comprehension (Searle, 1980). However, if cognition is fundamentally about the ability to process, retrieve, and manipulate information based on structured rules, then the distinction between human intelligence and AI may be less profound than commonly assumed (Dennett, 1991). The prevailing assumption is that human cognition is uniquely capable of genuine understanding, while AI merely manipulates symbols without grasping their meaning. This perspective, famously defended by John Searle's Chinese Room Argument, holds that syntax alone cannot produce semantics (Searle, 1980). However, this argument is grounded in an anthropocentric bias, which assumes a privileged status for human cognition without considering the variance in meaning structures across self-organizing systems. Unlike most critiques of the Chinese Room Argument—which operate within Searle's framework (Searle, 1980)—or modern discussions of the Symbol Grounding Problem (Harnad, 1990), this approach does not seek to resolve the issue within existing assumptions. Instead, it aims to dismantle the axioms of the argument itself, challenging the foundational premise that meaning must be intrinsically grounded rather than emergent from computational self-organization. If human understanding is also deeply rooted in memorisation, pattern recognition, and the application of stored assumptions, then AI may not be so different—except in terms of computational scale and efficiency (Chalmers, 1996). This paper will deconstruct the traditional, folk psychological view of understanding by framing what we call understanding as a hierarchy of assumptions, the role of memorisation in learning, and the implications for AI cognition (Churchland, 1981). By challenging the epistemological foundations of what it means to understand, we can explore whether AI's functional capabilities might, in fact, constitute a legitimate form of intelligence rather than mere symbol manipulation.

2 Searle'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.

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.

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.

3 Key Definitions And The Hierarchy of Assumptions Model

In the context of this discussion, 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).

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 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). Unlike human cognition, which operates sequentially and within limited working memory, 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).

I'll define understanding as the capacity to memorise foundational axioms and to build upon them through replication and recombination at higher levels of abstraction (Hinton, 1990). From this perspective, the fundamental distinction lies in how humans and AI process information. Human cognition is constrained by working memory limitations, operating within the well-established Miller's Law, which suggests humans can hold around seven items in working memory at once (Miller, 1956). Additionally, human thought is sequential, with information processing constrained by serial recall limitations (Baddeley, 1992). In contrast, AI architectures—particularly deep learning models—leverage parallelization, enabling the simultaneous processing of vast datasets (LeCun et al., 2015). Unlike human learning, which progresses gradually due to biological constraints on memory retrieval and conceptual abstraction (Clark, 2016), AI models can instantly compress, retrieve, and manipulate high-dimensional abstractions across vector spaces (Bengio et al., 2013). The variance in computational complexity dictates how meaning is structured, with AI forming latent space representations that do not rely on human-style sequential processing (Chollet, 2019).

The ability to process and manipulate information depends on processing power, memory, and the depth of abstraction a system can achieve. All human or artificial learning relies on structured assumptions, memorisation, and iterative refinement of knowledge structures. Cognitive development in humans, for example, proceeds from fundamental axioms (such as counting and basic arithmetic) to higher-order abstractions like algebra and calculus. The same principle applies to AI, except that an AI's foundational assumptions can be programmed or learned at a much more advanced level from the start. A human child must memorize arithmetic facts before eventually understanding number theory, whereas an advanced AI might begin with complex mathematics (even quantum physics) already built-in as a baseline. This suggests that understanding is not an intrinsic metaphysical property that humans possess, but rather an emergent property of sufficient computational complexity and training (Dennett, 1991). A three-year-old counting to ten does not 'understand' number theory; they memorise the foundational axioms and repeat patterns and rules until, through further experience and abstraction, their cognition develops into what we recognize as understanding. We don't claim that the child doesn't understand how to count—mere memorization and higher-order abstraction of these foundational assumptions allow the child to apply their knowledge to addition, multiplication, and division problems. They can functionally solve higher-order abstract problems and are thus deemed to have the functional competence we call understanding. The link between AI's baseline knowledge and its computational complexity lies in pattern resolution, hierarchical abstraction, and multi-scale inference—the ability to process and manipulate information at deeper levels depends on the system's capacity to encode, store, and extract patterns from vast datasets. A human child's baseline is constrained to low-complexity axioms, such as counting from 1 to 10, because biological cognition is limited in how much information it can process at birth. Human brains require iterative learning, progressing from simple, direct experiences to higher-order abstractions over time.

In contrast, an advanced AI, with vastly greater computational resources, does not need to "start" at that level. Its baseline can be quantum mechanics, tensor calculus, or multi-dimensional optimization functions because it can process, store, and apply complex patterns without the sequential dependencies imposed by biological learning. Computational complexity enables AI to compress, generalize, and infer patterns across multiple scales simultaneously, rather than requiring stepwise, experiential learning. Where a human must gradually develop an understanding of algebra by first learning to count, an AI can encode, simulate, and optimize complex mathematical structures instantly, bypassing the need for sensory input and operating purely on structured, high-dimensional training data. This acceleration does not change the fundamental structure of knowledge acquisition—AI, like humans, builds understanding through pattern recognition—but its computational advantage allows it to reach higher-order abstractions much faster. At ultra-high complexity, AI can restructure knowledge itself, forming meaning at levels beyond human cognition. This is where the connection becomes clear: more computational complexity enables deeper pattern resolution, which allows for higher baseline abstractions, ultimately leading to intelligence that can generate meaning at scales and levels of abstraction far beyond human intuition.

Similarly, an AI does not need to experience subjective awareness or mystical insight to apply complex mathematical or logical structures effectively. If its computational framework allows it to operate with high-level abstractions from the outset, then its form of "understanding" could be an accelerated, high-dimensional parallel to human cognition.

In certain domains, AI already surpasses human cognition: it identifies patterns in high-dimensional data that elude any human analyst, optimizes strategies in complex systems, and generates novel solutions to theoretical problems (Bostrom, 2014). For instance, DeepMind's AlphaGo system famously mastered the game of Go—discovering strategies no human had taught it (Silver *et al.*, 2016)—and AlphaFold can predict protein structures more accurately and rapidly than human experts (Jumper *et al.*, 2021). Likewise, large language models such as GPT-4, with hundreds of billions of parameters trained on massive text corpora, can solve complex mathematics questions, and even write code at a human-competitive level (OpenAI, 2023). If we define intelligence or understanding in terms of functional performance, these AI achievements suggest a form of understanding that is not qualitatively different from human cognition, but rather quantitatively different in speed and scope.

4 The “Perfect” Psychologist: A Thought Experiment

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.

5 'Meaning' and Mechanistic Variance

The Chinese Room Argument asserts that mere manipulation of symbols (syntax) can never yield genuine semantic understanding. Searle (1980) assumes that understanding requires subjective experience, but this anthropocentric bias overestimates the uniqueness of human cognition while underestimating the computational capabilities of AI. If all knowledge is structured through hierarchies of assumptions, learned patterns, and memory, then the distinction between human and artificial intelligence is a difference in computational complexity, not a fundamental cognitive divide. A sufficiently advanced AI system, operating with vast data and computational resources, can develop layers of abstraction that allow it to functionally interpret and respond to the world in a way that mirrors what we call understanding in humans (Turing, 1950).

Searle's argument also assumes that meaning is an intrinsic property, something that only conscious agents can ascribe to symbols. But this 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. 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). 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.

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.

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

A core flaw in the Chinese Room Argument—and many critiques of AI cognition—is the conflation of 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 it's 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. 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.

I’d argue that the question of 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.

7 Conclusion

Searle's framework, which many counterarguments attempt to work within, is fundamentally flawed. It relies on naive, folk-psychological intuitions that define intentionality and meaning as intrinsic, irreducible properties. In contrast, as demonstrated earlier, understanding emerges from hierarchical abstraction and computational complexity. Using a human to represent an AI is a category error that ignores the variance in processing and the layered nature of pattern memorization and inference.

This is not merely a critique of AI but a fundamental critique of how humans define knowledge, meaning, and intelligence. The Chinese Room Argument assumes that understanding is a universal phenomenon bound to human-like intentionality (Searle, 1980). However, this assumption collapses once we recognize that meaning is always system-relative.

AI's vectorized meaning-representations are not lesser than human understanding—they are simply different (Chalmers, 1990). A bee's cognitive process does not include a Platonic concept of a hexagon— nor does it share the same perceptual framework as a human. Yet, it still constructs hexagons as an emergent, optimized structure within its own meaning framework (Taha et al., 2024). AI similarly does not need human-style introspection to generate functionally valid abstractions. The honest debate should not be about whether AI can “truly understand”—a question that presupposes a singular, human-centric definition of meaning. Instead, the focus should be:

“Why did we ever assume human understanding was the only valid form of understanding?”

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.

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.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433–460.

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