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

Abraham Dada

London, UK

abraham@stoado.com

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.

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, starts processing information. This epistemological reduction challenges the notion that human understanding is intrinsically different from artificial intelligence. 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 rests on an anthropocentric bias that presumes a privileged status for human cognition. 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.

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.

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. 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. From this perspective, the difference between human and AI cognition is not categorical but instead a matter of computational complexity and degree of abstraction. The ability to process and manipulate information depends on processing power, memory, and the depth of abstraction a system can achieve. All learning—whether human or artificial—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 four-year-old reciting numbers does not understand number theory; she repeats patterns and rules until, through further

experience and abstraction, her cognition develops into what we recognize as understanding. Similarly, an AI need not experience subjective awareness or mystical insight to effectively apply complex mathematical or logical structures. 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.

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, which is an abstraction of chemistry, which is an abstraction of physics, which is an abstraction of mathematics, and mathematics itself 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. 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 better predict how neurotransmitter imbalances influence cognitive behaviour (Kandel, 2006). A stronger grasp of mathematics could refine their understanding of statistical modelling in psychological studies, improving their ability to detect patterns in human cognition (Gigerenzer, 2002). If they possessed a physicist's knowledge of the brain's electrochemical processes, they might reframe certain psychological disorders as computational inefficiencies rather than traditional diagnoses. 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

clinical 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 the domain they operate within (Kahneman, 2011). 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. 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).

This thought experiment undercuts the Chinese Room argument's demand for some additional *intrinsic* understanding beyond functional performance. If a psychologist does not need to perceive quantum processes in neurons to meaningfully engage with psychology, then why should an AI need subjective experience to understand language? Humans operate with layered abstractions and use information appropriate to the level at hand; that is sufficient for practical understanding. We don't require a person to be a physicist to say they understand a car engine, nor do we require them to *feel* what the engine "feels." By analogy, we should not require an AI to replicate the *entirety* of human cognitive architecture (from quantum biology to conscious qualia) in order to 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.

The Chinese Room 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. 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 (Dada, 2025c). It

is not an irreducible, mystical property of human cognition, nor does it require subjective awareness. 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. 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. Seeing, hearing, smelling, touching, and feeling are not passive experiences; they are mechanisms by which the brain assigns meaning to raw sensory input (Dada, 2025c). 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. This extends beyond sensory input into abstract cognition, where higher-order meaning—such as assigning purpose to ideas, events, or deities—emerges as a natural extension of the brain’s predictive mechanisms. 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 (Dada, 2025c).

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'. But this problem is not a real explanatory gap—it is an artifact of human cognitive bias (Dada, 2025b). Qualia is just mechanistic variance—the inevitable result of different configurations of a system interacting with its environment in different ways. 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 (Dada, 2025d). 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. 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. These differences are analogous to how humans and other animals construct meaning from perception in fundamentally distinct but functionally effective ways.

If meaning and qualia are both emergent properties of structured interaction rather than intrinsic, irreducible entities, then Searle's assumption that AI lacks understanding due to the absence of "true" meaning or subjective experience collapses. The only difference is that humans experience these processes introspectively, while AI does not need to. But introspection itself is an illusion of cognitive architecture, not a prerequisite for intelligence. The argument that "machines lack understanding" ultimately falls apart once we remove the flawed

premise that meaning and qualia must exist as metaphysical properties rather than computational ones.

Even if one-day consciousness is proven to involve non-algorithmic—as Penrose and others entertain (Penrose, 1989), although 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. Human intelligence itself correlates with structured computation, even if it happens to be instantiated in biological neural networks rather than artificial architectures. Many human cognitive processes, such as intuition, perception, and decision-making, occur without explicit conscious reflection (Libet, 1985; Soon et al., 2008). This suggests that understanding, as a functional property, does not require subjective awareness—it simply requires an adaptive system that organizes and applies stored knowledge at different levels of complexity.

Conclusion

The real question is not "Can an AI ever have human-like subjective consciousness such that it truly understands?" but rather "Is subjective consciousness even necessary for intelligence and understanding?". The evidence presented here suggests that it is not. Insisting that AI can't achieve "real understanding" simply because it lacks a specific human-like phenomenology may itself be a fallacy.

References

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

Churchland, P. M. (1981). Eliminative Materialism and the Propositional Attitudes. *Journal of Philosophy, 78*(2), 67–90.

Dada, A. (2025b). *A Very Short Essay: Qualia Is Just Mechanistic Variance, The Hard Problem of Anthropocentric Bias*. AbrahamDada.com Essays.

Dada, A. (2025c). *A Short Essay: How Self-Organising Systems Abstract "Meaning"*. AbrahamDada.com Essays.

Dada, A. (2025d). *The Mind Beyond the Senses: Axiomatic Thought and Its Independence from Perception*. AbrahamDada.com Essays.

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

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

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

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

Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in

voluntary action. *Behavioral and Brain Sciences, 8*(4), 529–566.

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

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.

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

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.

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

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.