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 challenges the anthropocentric bias in traditional epistemology, arguing that intelligence—whether human or artificial—is defined not by subjective experience, but by functional competence. If an AI system can reason, learn, and adapt in ways that surpass human cognition, does it matter if it lacks consciousness? Or is the distinction between syntax and semantics simply an illusion born from our own cognitive limitations?

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 view of understanding by examining knowledge as a hierarchy of assumptions, the role of memorisation in learning, and the implications for AI cognition. 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 (Churchland, 1989).

The Hierarchy of Assumptions Model

In the context of this paper, 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.

I will define understanding as the capacity to memorise foundational assumptions and replicate processes at high-level abstraction. From this perspective, the difference between human and AI cognition is not categorical but a matter of computational complexity.

The ability to process and manipulate information depends on processing power, memory, and the depth of abstraction a system can reach (Kurzweil, 2005). All learning, whether human or artificial, relies on structured assumptions, memorisation, and iterative refinement (Clark, 2016). Human cognition develops from fundamental axioms, such as numerical counting, before advancing to higher-order abstractions like algebra and calculus. The same principle applies to AI, except that its foundational assumptions begin at a more advanced computational level. A child memorises arithmetic before understanding mathematical abstraction, while an AI system may begin with something advanced, such as quantum field theory, as its baseline (Tegmark, 2017).

This suggests that understanding is not an intrinsic, metaphysical quality but an emergent property of computational complexity (Dennett, 2017). A four-year-old memorising numbers does not understand number theory; they replicate patterns until, through interaction and abstraction, their cognition develops into what is recognised as understanding. Similarly, AI does not need to understand subjectively to apply complex mathematical structures effectively. If its computational framework allows it to work with high-level abstractions from the outset, then its form of understanding may be an accelerated, high-dimensional counterpart to human cognition (Schmidhuber, 2015).

If both human and AI cognition rely on structured assumptions and stored knowledge, then why is human understanding considered superior? The brain operates through neural networks that reinforce learned behaviours over time, creating an illusion of understanding in much the same way AI refines its models through training (Hawkins and Blakeslee, 2004). The difference is one of scale and efficiency rather than fundamental nature. AI already surpasses human cognition in certain domains. It identifies patterns in high-dimensional datasets beyond human perception, optimises strategies in complex systems, and generates novel solutions to theoretical problems (Bostrom, 2014). If intelligence is computationally defined, then AI's ability to perform such tasks suggests a form of functional understanding that is not qualitatively different from human cognition—only quantitatively superior in some contexts.

The "Perfect" Psychologist: A Theoretical 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.

This argument undercuts the Chinese Room Argument's demand for intrinsic understanding. If a psychologist does not need to understand quantum mechanics to meaningfully engage with psychology, then why should an AI need subjective experience to understand language? If human cognition operates at layered abstractions, why must AI be expected to function differently? The demand that AI possess some deeper "intrinsic" comprehension is as unreasonable as demanding that a psychologist master string theory 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.

Conclusion

The Chinese Room Argument hinges on the assumption that mere syntax can never produce genuine semantic understanding—yet this view often overestimates the uniqueness of human cognition (Searle, 1980). If all knowledge is rooted in hierarchies of assumptions and memorisation, then the line between human and artificial intelligence becomes blurred; the key difference is one of complexity and efficiency rather than an intrinsic, metaphysical property. If an entity—human or machine—can use stored knowledge and assumptions to produce insights, adapt to

new contexts, and generate meaningful responses, then labelling its process as mere syntax might ignore the emergent sophistication that we commonly equate with understanding.

References

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

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

Churchland, P.M. (1989) *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*. Cambridge, MA: MIT Press.

Clark, A. (2016) *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. New York: Oxford University Press.

Dennett, D.C. (1991) *Consciousness Explained*. Boston: Little, Brown and Company.

Dennett, D.C. (2017) *From Bacteria to Bach and Back: The Evolution of Minds*. New York: W.W. Norton & Company.

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

Gigerenzer, G. (2002) *Calculated Risks: How to Know When Numbers Deceive You*. New York: Simon & Schuster.

Hawkins, J. and Blakeslee, S. (2004) *On Intelligence*. New York: Times Books.

Kahneman, D. (2011) *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Kandel, E.R. (2006) *In Search of Memory: The Emergence of a New Science of Mind*. New York: W.W. Norton & Company.

Kurzweil, R. (2005) *The Singularity is Near: When Humans Transcend Biology*. New York: Viking Penguin.

Schmidhuber, J. (2015) 'Deep learning in neural networks: An overview', *Neural Networks*, 61, pp. 85-117.

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

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