Given that modern systems seem to pass the Imitation Game, are we rationally entitled to treat them as thinking beings?

By Faraz Fookeer

In this essay, I argue that modern systems — namely, contemporary artificial intelligence (AI) systems — cannot be rationally treated as thinking beings. The question of whether we are rationally entitled to treat AI applications as thinking beings forces us to ask whether behaviour that reasonably and convincingly resembles human intelligence justifies the attribution of thought, or can genuine thought be perceived from something more than linguistic success? In this essay, I define what a “thinking being” is, and I outline the scope of AI systems and applications that form the subject of the main question. I then present the three experiments that I use to argue my point: Turing’s seminal imitation game experiment (Turing 1950), and two thought experiments: the Chinese room (Searle 1980) and blockhead (Block 1981). Finally, I use the latter thought experiments to explain why AI systems cannot be rationally treated as thinking beings.

To evaluate whether an AI system can be thought of as a “thinking being” it is necessary to define what such a being is, in the context of this argument. From a philosophy perspective a “thinking being” can form and modify internal inner pictures or symbols (representations) that are about something in the world (Searle 1980). For example humans can hold an idea of an apple that represents a physical object called an “apple’ in the wider world. We give this object meaning. Daniel Dennett (1987) refers to thinking beings as intentional systems, agents where reasoning is driven by their internal beliefs and desires. Turing (1950) flattens the concept of a thinking being to an implicit definition of something capable of meaningful contextual linguistic interactions. For the purpose of my argument, I align myself with Searle’s definition of a thinking being where he goes further and defines representation as a state with semantic content i.e. meaning, differentiating it from mere syntactic symbol manipulation (Searle 1980). A “thinking being” then is one that does not merely output correct information but one that does so via comprehension, intentionality and an awareness of the world and meaning.

The AI systems that form the scope of this argument are specifically part of the large language model (LLM) family where commercial examples include OpenAI’s market-leading ChatGPT, Google’s Gemini and Anthropic’s Claude. These are deep-learning systems built on so-called transformer neural network architectures and trained on vast datasets through self-supervised and reinforcement learning. Linguistic input is tokenised into high-dimensional vector representations that are then decoded to predict the next token or sequence (Elastic 2024). An LLM generates statistically likely continuations of language based on patterns in its training data. The outputs gleaned from such models appear intelligent, coherent and almost thoughtful in their approach to human interaction, and in the following paragraphs I argue that this is a

fallacy: these systems ought not to be treated as thinking beings. The scope of my argument is confined to LLMs that output text since versions of these models have passed the Turing test (Jones and Bergen 2025).

In our interactions with people we cannot determine what goes on behind their eyes. We are not privy to their internal states: mental, spiritual or otherwise. We infer the intentions of others via their behaviour patterns. Turing's seminal Imitation Game (Turing 1950) extends this logic to machines and poses the question "Can machines think?" This test hinges on behavioural equivalence to prove that machines possess intelligence and disregards the necessity of creating a mental reality of the world - an internal representation. Turing turned a complex philosophical question into a testable criterion for measuring "intelligence". An intelligence that is purely based on behavioural cues without the consideration of any inner mental states.

Thirty years later John Searle challenged Turing's behavioural criterion directly. In his Chinese Room thought experiment Searle imagines himself locked in a room where he is given Chinese inputs, an English instruction manual used to manipulate the Chinese symbols and a means to output these symbols to the world. From an external observer, the room appears to understand Chinese, however John - following English instructions - understands nothing of the symbols that he is passing across his desk (Searle 1980). Searle reaches the conclusion that computation is purely syntactic. Symbols are manipulated via a set of rules, absent of any meaning. He goes on to argue that thought is semantic - it involves intentionality and the "aboutness" of mental states. Searle goes on to define "weak AI' and "strong AI", accepting the former and rejecting the latter. Weak AI is simplistic and aligns with Turing's model of a thinking machine - it acts as a tool to aid the study of cognition. Whereas Turing's model provides an optimistic, strong AI view of the world, Searle rejects this notion of a strong artificial intelligence arguing that such a machine is a mind that would need to understand concepts and derive thought through ascribing meaning to the world.

I agree with Searle's view of the world due to the nature of LLMs themselves. These machines stack probabilistic numerical methods atop one another and bind them with linear algebra, operating through multi-layered vector transformations and token predictions ("Overview of Large Language Models" 2024). They mimic human speech patterns and behaviour to a point yet frequently fail to provide accurate or context-appropriate responses - a problem known as hallucination (Huang et al. 2023; "Hallucinations in LLMs" 2024). Although they are trained on human-derived data, they cannot understand what this data means through word use or association alone. Their outputs remain syntactic rather than semantic, supporting Searle's argument that symbol manipulation is not understood. They resemble the Chinese Room writ large: systems that reproduce patterns of language without genuine comprehension. Even their personalities can be artificially tuned - recent reports note that GPT-5 was perceived as less friendly or personable than GPT-4o, reflecting the fact that such "traits" are design features rather than signs of genuine personality (Ars Technica 2025; Business Insider 2025). Hence, while impressive simulators of human language, these systems cannot currently be rationally considered thinking beings.

My argument is furthered by the work of Ned Block and his Blockhead thought experiment (Block 1981). This consisted of a black box containing a lookup table that held every imaginable conversation. Given an input the machine will retrieve the output from the lookup table and, to the external observer, it would seem that the machine is intelligent enough in its replies. Such a device would pass the Turing test with flying colours yet it is merely relying on pattern recognition to generate the correct answers. In a sense, Block's machine is closer to modern day LLMs as there is a greater degree of pattern matching that its internal mechanism is required to do compared to Searle's Chinese room "machine". Block goes on to argue that the behaviour of a system alone cannot be a marker of intelligence as two identical outputs may result from differing internal states. One could argue that the act of looking up data in a table could constitute mechanical "thought" but there is no understanding or meaning ascribed to this action. This machine is devoid of understanding. I am in agreement with Block's argument that although behaviour is a condition for determining intelligence, it is not a sufficient one. To be justified in ascribing thinking to LLMs we must understand their internal processes and, as described above, these are too mechanistic to generate thought.

I concede that there may be grounds for a rebuttal of my argument. Gnawing at the end of my mind are two questions that, although nascent and not yet thought through, deserve to be highlighted as their answers could form a contrarian stance. The first is the idea of frame of reference. Could it be that, from the machine's statistical computational frame of reference, the act of using probability to associate tokens is a form of thought? The second question is embedded in the first and concerns the flow of time: given that the entirety of humanity's lexicon is finite then repeating the same token associations within a stream of time can give rise to meaning?

In the literature both Dennett (1987) and Clark (2016) provide a functionalist and predictive-processing perspective which could serve as a rebuttal to my position. This view counters the position by basing thought on a system's internal organisation mirroring the causal roles of human cognition. Extending this idea to LLMs arrives at an answer to the first question of my rebuttal, that is, large language models may implement a rudimentary form of predictive reasoning as they anticipate linguistic outcomes in a manner that is structurally analogous to how the human brain predicts sensory input. However, my argument stands as one cannot underestimate the difference between prediction and understanding, where human cognition is embedded in perception, emotion and intentionality.

To conclude, I argue that modern systems - LLM-based artificial intelligence systems - cannot be rationally thought of as thinking beings. We have defined a "thinking being" as one with an internal, malleable, representation of the world and an intentionality that provides the being with meaning about itself and its environment (Searle 1980). The Imitation Game is purely behaviouristic and the natures of the players are not known (Turing 1950). Both Searle and Block have shown us that behaviour is insufficient for determining whether a system thinks. The system's internal state is unknown and whether outputs are generated via instructions (Searle 1980) or via pattern recognition (Block 1981) these are merely simulations of human cognition - semantic understanding is missing (Searle 1980). With contemporary LLMs, we know that they generate human-like linguistic outputs based on statistical numerical methods. We know that

they do not think and that their architectures are incapable of generating thought. Until these machines can ground their outputs in meaning, they cannot be rationally thought of as thinking.

References

Ars Technica (2025) ‘ChatGPT users hate GPT-5’s “overworked secretary” … miss their GPT-4o buddy.’ 8 Aug. Available at: https://arstechnica.com/ai/2025/08/chatgpt-users-outraged-as-gpt-5-replaces-the-models-they-love/ (Accessed 25 October 2025).

Block, N. (1981) ‘Psychologism and Behaviourism.’ *Philosophical Review,* 90 (1), pp. 5–43.

Business Insider (2025) ‘Sam Altman says GPT-5’s “personality” will get a revamp — but it won’t be as “annoying” as GPT-4o.’ 13 Aug. Available at: https://www.businessinsider.com/sam-altman-openai-gpt5-personality-update-gpt4o-return-backlash-2025-8 (Accessed 25 October 2025).

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

Dennett, D. C. (1987) *The Intentional Stance.* Cambridge, MA: MIT Press.

Elastic (2024) *What Are Large Language Models (LLMs)?* Available at: https://www.elastic.co/what-is/large-language-models (Accessed 22 October 2025).

Huang, Q. et al. (2023) *A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges and Open Questions.* arXiv pre-print. Available at: https://arxiv.org/abs/2311.05232 (Accessed 25 October 2025).

‘Hallucinations in Large Language Models: Types, Causes, and Approaches for Enhanced Reliability’ (2024) *ResearchGate* [online]. Available at: https://www.researchgate.net/publication/385085962_Hallucinations_in_LLMs_Types_Causes_and_Approaches_for_Enhanced_Reliability (Accessed 25 October 2025).

‘Overview of Large Language Models’ (2024) arXiv pre-print. Available at: https://arxiv.org/html/2307.06435v9 (Accessed 25 October 2025).

Jones, C. R. and Bergen, B. K. (2025) *Large Language Models Pass the Turing Test.* arXiv pre-print. Available at: https://arxiv.org/abs/2503.23674 (Accessed 25 October 2025).

Searle, J. R. (1980) ‘Minds, Brains and Programs.’ *Behavioral and Brain Sciences,* 3 (3), pp. 417–457.

Turing, A. M. (1950) ‘Computing Machinery and Intelligence.’ *Mind,* 59 (236), pp. 433–460.