# A Critical Examination of Searle's Chinese Room Thought Experiment

Zhuoya Tian *

Faculty of Arts, The University of Hong Kong, Hong Kong, China

* Corresponding Author Email: u3643936@connect.hku.hk

**Abstract.** Searle's Chinese Room thought experiment argues that computers cannot understand language, but this view has been challenged from multiple perspectives. According to critics, comprehension entails a relationship to consciousness and the outside world in addition to manipulating symbols. There is little evidence to support Searle's denial of machine intentionality, and functionalism, which emphasizes the functional aspects of mental experiences, provides an alternative viewpoint. Furthermore, developments in contemporary artificial intelligence suggest that, under certain situations, machines are capable of processing information efficiently. Our knowledge of machine comprehension is changing as a result of the quick advancements in deep learning and natural language processing.. Although Searle's experiment has its limitations, it has had a lasting impact on discussions in philosophy and artificial intelligence, prompting us to reflect on the relationship between humans and machines and driving further research in this area.

**Keywords:** Chinese Room Thought Experiment; Functionalism.

## 1. Introduction

The Chinese Room experiment was proposed by John Searle in the 1980s to explore whether computers can truly understand language. The experiment asks whether computer symbol processing is comparable to human language comprehension by simulating a person who does not comprehend Chinese operating symbols based on rules inside a locked room.

In recent years, research on the Chinese Room experiment has deepened. Scholars have not only criticized Searle's arguments but also offered new perspectives. For instance, researchers are focusing on advancements in natural language processing (NLP), examining how machine learning and deep learning change our definitions of understanding. Many recent studies indicate that computer systems can exhibit near-human understanding capabilities in specific contexts through contextual analysis and semantic reasoning.

This paper will clearly discuss which aspects of Searle's Chinese Room experiment still hold philosophical value and which aspects have been challenged by modern technological developments. By incorporating the most recent developments in natural language processing and artificial intelligence, this paper will reexamine the theoretical underpinnings and limitations of Searle's Chinese Room experiment, exploring its philosophical significance in the context of modern technology. Prior to discussing the difficulties posed by functionalism to the experiment, the article will analyze the main counterarguments against Searle's Chinese Room experiment, first analyzing the difference between symbol manipulation and true understanding, and then examining the crucial role of consciousness in the understanding process. Through this analysis, it is hoped to provide a more comprehensive perspective that reveals the limitations of Searle's Chinese Room experiment and offers insights for future research in artificial intelligence.

## 2. Experiment Overview

In his article "Minds, Brains, and Programs," John Searle first introduced the Chinese Room thought experiment to critique the theory of strong artificial intelligence.[1] He based his argument on the "story understanding" programs conceptualized by Roger Schank and others, imagining a

"Chinese Room" that could fully exemplify such a program but lacked semantic understanding, thereby challenging the core idea of strong AI — those programs are equivalent to minds

Specifically, the room contains a set of Chinese symbols and a corresponding English rulebook. Searle, who can only read English, manipulates the incoming Chinese characters by following the rulebook's instructions to generate an output in a different order. He is unaware, nevertheless, that the arriving Chinese symbols constitute a Chinese inquiry and that the rearranged Chinese symbols he produces are the proper responses to that question. It would seem to an outsider that Searle comprehends the Chinese inquiry and gives the right response. However, in practice, Searle never comprehends the Chinese question; instead, he is simply manipulating symbols in accordance with the rulebook, mimicking the functioning of a computer..Thus, Searle concludes that "a program by itself cannot constitute a mind, and the formal syntax of a program does not guarantee the emergence of mental content."

## 3. Counterarguments

### 3.1. Counterarguments One: The Distinction Between Symbol Manipulation and Understanding

The core argument of Searle's Chinese Room experiment is that computers can simulate language understanding through symbol manipulation, but this does not mean they truly understand language.Despite the experiment's apparent support for Searle's theory, skeptics like Daniel Dennett and Harry Frankfurt contest this dichotomy, contending that the lines separating comprehension from symbol manipulation are not as sharp as they may appear..

#### 3.1.1 The Surface Phenomenon of Symbol Manipulation

In Searle's experiment, the person inside the room manipulates Chinese symbols according to the rules without understanding their meaning. Critics counter that in practice, this symbol manipulation might nonetheless promote effective communication even though it does not amount to "understanding" in the conventional sense. The ability of contemporary natural language processing systems, such GPT models, to produce coherent text, for instance, suggests that computers are capable of partially capturing the intricate linguistic patterns.This aligns with philosopher Joseph Weizenbaum's view of the ELIZA program; despite ELIZA's limited understanding, its interactive effects demonstrate the potential of symbol manipulation.[2]

#### 3.1.2 The Multi-layered Nature of Understanding

Understanding language is not merely a simple response to symbols; it involves multiple layers, including semantics, pragmatics, and context. Linguist Noam Chomsky emphasizes the generative capacity of language, noting that humans can create an infinite number of new sentences, a capability that does not entirely rely on a specific understanding of each symbol.[3] Critics argue that computers, by learning from vast amounts of data, can develop similar abilities, thus achieving a certain degree of "understanding."

Harry Frankfurt's research posits that understanding is a multi-layered process that can manifest in different forms across various systems.[4] At the same time, comprehension and aim are strongly linked. He contends that genuine comprehension depends on a person's goals and intents; people need to be able to discern symbols' meanings in certain settings in addition to manipulating them. This point of view contradicts the notion in Searle's Chinese Room experiment that reduces comprehension to simple manipulation of symbols. According to Frankfurt, a system is said to have some level of understanding if it can react in accordance with context and aim. Additionally, according to this viewpoint, Searle's experiment ignores the fundamental aspects of comprehension and does not sufficiently take into consideration the critical roles that context and purpose play in the cognitive process.

Philosopher Daniel Dennett believes that understanding is not limited to uniquely human subjective experiences but can manifest in different ways across various systems. [5] He points out that while computers lack consciousness, they can exhibit a certain degree of "understanding" through complex algorithms and large-scale data processing. This contrasts with Searle's view, which restricts understanding to the realm of subjective experience.

### 3.1.3 The Practicality of Language Use

The meaning of language is often derived from its actual use, as Ludwig Wittgenstein stated, "The meaning of a word is its use in the language."[6] Critics contend that computers are capable of contributing to this dynamic language use by giving symbols new meanings. Computers can comprehend the situational context of language use to some degree, as demonstrated by the ability of machine learning models to modify their replies in response to context. In stark contrast to the static symbol manipulation in Searle's experiment, this realism highlights the dynamic nature of language. This practicality emphasizes the dynamic nature of language, sharply contrasting with the static symbol manipulation in Searle's experiment.

Harry Frankfurt also highlights the practicality and functionality of understanding. He argues that understanding should not be viewed solely as an internal subjective experience; rather, it should focus on its actual effects in communication and interaction.

## 3.2. Counterarguments Two: The Absence of Intentionality

Intentionality is an important concept in the philosophy of mind. Searle argues that intentionality is a characteristic of certain mental states, through which these states refer to or involve objects or states of affairs in the world.[7] As Daniel Dennett puts it, "Intentionality is the language of the mind; it is the way it interacts with the world." Dennett is talking about intention and consciousness. According to him, comprehending a mental state's intentionality is essential to comprehending how it interacts with the outside world.[8] Intentionality, to put it simply, is about the mind's capacity to engage with the outside environment.

Searle posits that the mind consists of two components: syntax and semantics, with semantic content, or the meaning of language, being the most crucial. Computer programs rely solely on syntactic operations, making them purely symbolic. Consequently, the operations and processing of language by computers are considered to lack external semantic reference. Searle concludes that the fundamental difference between humans and computers lies in intentionality, which he claims is unique to the human biological brain, whereas computers do not possess intentionality.

Dennett, however, opposes this view. He illustrates his point: "A shopping list written on a piece of paper only has metaphorical intentionality from the agent who wrote it. Similarly, a shopping list retained in the memory of the same agent has the same kind of metaphorical intentionality, for the same reasons." This implies that even though a computer program is created by a person, computers can still be thought of as having "intentionality." As Dennett points out, even though the shopping list itself lacks actual intentionality, we might nevertheless consider it to have "metaphorical intentionality" as it aids in achieving a particular objective.[8] Stated differently, there is now insufficient evidence to support the claim that computers lack intentionality and only humans do.

## 3.3. Counterarguments Three: The Challenge of Functionalism

Functionalism argues that the definition of mental states should be based on their functions or roles, rather than their material basis. According to functionalists like Hilary Putnam, mental states are defined by their connections to other states and their functions within a system.[9] Stated differently, the core of mental states is not their particular forms of implementation, but rather their interactions with other mental states and the external environment. According to philosopher Julian Baggini, functionalism holds that "different materials can realize the same mental functions."[10]

Searle's Chinese Room thought experiment aims to demonstrate that machines cannot understand language due to their lack of genuine understanding. However, from a functionalist perspective, this

experiment has significant shortcomings. First, the Chinese Room experiment envisions a person who does not understand Chinese processing Chinese symbols through rules; although he can generate correct responses, he does not understand their meaning. Searle emphasizes the subjective experience of understanding, arguing that a machine's symbol manipulation cannot be equated with understanding.[1] However, functionalists contend that the definition of understanding should focus on its functionality, specifically whether the ability to process symbols can produce appropriate responses in specific contexts.

### 3.3.1 Understanding from a Functionalist Perspective

Within the functionalist framework, understanding is not merely an expression of internal consciousness but rather how a system effectively processes information and interacts with its environment. This point of view is expressed by Douglas Hofstadter in his work *Gödel, Escher, Bach*, where he contends that understanding can appear in different implementations as long as these systems are able to carry out the same tasks. Put another way, even while machines might not be as sentient as people, they might be thought of as having some understanding if they are able to process information efficiently and react correctly in particular situations.[11]

Functionalists, such as Hilary Putnam, also support this perspective, noting that complex systems can achieve "understanding" through appropriate functions. Even if a machine's operations are based on algorithms, they can still "understand" inputs and generate responses to a certain degree.[12]In this sense, Searle's Chinese Room experiment fails to adequately consider the importance of function and effect, simplistically equating understanding with the subjective experience of consciousness and overlooking the functionalist emphasis on system performance.

### 3.3.2 Refuting Searle's Assumptions

Searle's experimental hypothesis views the philosopher inside the Chinese Room as an entity lacking the ability to understand. But functionalism places more emphasis on the connection between function and effect than it does on the reality of consciousness. The philosopher in the room is capable of doing symbol transformations and answers even if he does not comprehend Chinese. Functionalists would contend that Searle's experiment does not provide a thorough knowledge of how mental states function.

Moreover, the functionalist perspective also points out that the understanding capabilities of machines and humans are not necessarily mutually exclusive. Putnam argues that as long as machines can process information effectively in a given context, they can be considered to possess some form of understanding. [9] As David Chalmers notes, the relationship between consciousness and understanding is not a simple causal one, but rather a complex functional interaction.[13]

A functionalist critique of Searle's Chinese Room thought experiment highlights the variety of knowledge and how closely it relates to functionality. Searle ignores the functionalists' more comprehensive concept of mental states by rigidly linking understanding to consciousness. Functionalists contend that even while robots might not be conscious, they can be thought of as having the ability to understand if they are able to process language and interact with their surroundings.

### 3.3.3 Multiple Realizability

Multiple realizability posits that the same mental states or cognitive functions can be instantiated in various physical systems. This concept challenges the notion that mental processes are strictly tied to biological substrates, suggesting that both biological brains and artificial systems can exhibit similar cognitive capabilities.

The Chinese Room Argument implies that understanding is inherently tied to human-like cognition. Multiple realizability, on the other hand, permits the possibility that mental states, like comprehension or belief, may manifest in several ways. For instance, cognitive processes can arise from both human brains and sophisticated AI systems, demonstrating that cognition is not limited to biological things, just as a computer program can execute on different hardware.[14]

### 3.4. Counterarguments Four: The Impact of Emerging Technologies

With the rapid advancements in deep learning and natural language processing technologies, our definition of machine understanding is evolving. The rigid division between symbol manipulation and understanding observed in Searle's experiment is challenged by contemporary artificial intelligence systems, which not only produce grammatically accurate text but also engage in semantic reasoning depending on context. Modern natural language processing systems, such GPT-4 and GPT-o1, are able to provide responses that are both fluent and appropriate for the context, showcasing language processing abilities that extend beyond simple symbol matching..For instance, deep learning models trained on vast amounts of data can capture semantic relationships, leading to a deeper understanding of language. This suggests that comprehension requires on efficient information processing and contextual analysis rather than being only the product of subjective experience.Contextual awareness is becoming more and more possible in modern AI systems, especially those that use Artificial General Intelligence (AGI). Modern AI can communicate with its surroundings through sensors and feedback systems, unlike the solitary person in Searle's thought experiment. Future technological developments may force us to reconsider what machine understanding is and admit that, in certain situations, machines are capable of comprehending. At that point, Searle's views may once again be challenged.

## 4. Conclusion

In conclusion, although Searle's Chinese Room experiment has sparked extensive discussion in the field of philosophy, its core arguments have faced multiple challenges. Understanding is more than just manipulating symbols; it also entails making the link between awareness and the outside world. Furthermore, there is insufficient evidence to support Searle's claim that robots can never be deliberate from the standpoint of intentionality. However, functionalism stresses that mental states should be defined according to their functions rather than their physical foundation. The advancement of contemporary artificial intelligence challenges the conventionally limited conceptions of understanding by showing that robots are capable of efficient information processing and responses in particular settings. Finally, with advancements in deep learning and natural language processing technologies, our views on machine understanding continue to evolve.

In the future, scientists will further explore the nature of machine understanding, particularly how to assess and define the standards of understanding in different contexts. Additionally, there will be a focus on the performance of artificial intelligence in practical applications to better understand its capabilities in cognition and language processing.

Despite its drawbacks, Searle's Chinese Room experiment has had a significant influence on conversations about consciousness, understanding, and artificial intelligence. It has made us consider how humans and machines interact and has grown to be a crucial area of study in both philosophy and technology.

## References

[1] Searle J. Minds, Brains, and Programs[J]. 1980.

[2] Weizenbaum J. ELIZA—a computer program for the study of natural language communication between man and machine[J]. Communications of the ACM, 1966, 9(1): 36-45.

[3] Chomsky N. Syntactic structures[M]. Mouton de Gruyter, 2002.

[4] Frankfurt H. The importance of what we care about[J]. Synthese, 1982: 257-272.

[5] Dennett D C, Dennett D C. Consciousness explained[M]. Penguin uk, 1993.

[6] Wittgenstein L. Philosophical investigations[M]. John Wiley & Sons, 2009.

[7] Searle J. Intentionality: an essay in the philosophy of mind[M]. Cambridge University Press, 1983.

[8] Dennett D C. Précis of the intentional stance[J]. Behavioral and brain sciences, 1988, 11(3): 495-505.

[9] Putnam H. Mind, language and reality[M]. Cambridge University Press, 1979.

[10] Fosl P S, Baggini J. The philosopher's toolkit: a compendium of philosophical concepts and methods[M]. John Wiley & Sons, 2020.

[11] Hofstadter D R. Gödel, Escher, Bach: an eternal golden braid[M]. Basic books, 1999.

[12] Prigogine I, Stengers I. The end of certainty[M]. Simon and Schuster, 1997.

[13] Chalmers D J. The conscious mind: In search of a fundamental theory[M]. Oxford Paperbacks, 1997.

[14] McGrath, S. W., & Russin, J. (2024). Multiple Realizability and the Rise of Deep Learning. arXiv preprint arXiv:2405.13231.