ZAOULI – N° 9, Vol. 5 - Mai 2025 pp. 692-723 ISSN : 2788-9343
Soumission : **21/03/2025** Acceptation : **18/04/2025** Publication: **20/05/2025**

Artificial Intelligence and the Question of Consciousness: John Searle's "Chinese Room" Argument

Kheira BENAISSA
Abou Bekr Belkaid– Tlemcen University. Algeria
Phenomenology Laboratory and its Applications
E-mail: kheira.benaissa@univ-tlemcen.dz

Abstract
The remarkable advancement in artificial intelligence has raised numerous questions, provoked concerns, and confronted the world with significant challenges that extend beyond technical and informational boundaries to encompass human and ethical dimensions, ultimately targeting the human being himself. The latter has become the subject of simulation and imitation by strong AI researchers, reaching a point of transgression and claims regarding the possibility of creating an intelligent machine, commonly called the "machine mind". This has sparked widespread debate, critique, and defence of the human being, aiming to refute the notion that machines can be rendered as intelligent as humans.We aim to address this proposition within a contemporary philosophical framework, specifically within the current of the new philosophy of mind, to refute the claims of strong artificial intelligence advocates. These proponents base their research on the possibility of simulating the human mind, asserting that the latter is a digital computer. This debate was raised by the American philosopher John Searle, who justified his position through the Chinese Room argument. Accordingly, we pose the following problem: Can human intelligence be artificially simulated? If so, can a machine possess self-awareness? What are the philosophical justifications employed by John Searle in his critique of strong artificial intelligence and his rejection of the notion that "the mind is merely a digital computer"?

Keywords: *Artificial Intelligence, John Searle, Chinese Room Argument, Consciousness, Computational Theory.*

L'intelligence artificielle et la question de la conscience : L'argument de la "chambre chinoise" de John Searle

Résumé
Les progrès remarquables de l'intelligence artificielle ont soulevé de nombreuses questions, provoqué des inquiétudes et confronté le monde à des défis importants qui dépassent les frontières techniques et informationnelles pour englober des dimensions humaines et éthiques, visant en fin de compte l'être humain lui-même. Ce dernier est devenu l'objet de simulations et d'imitations de la part de puissants chercheurs en IA, atteignant un point de transgression et de revendication concernant la possibilité de créer une machine intelligente, communément appelée "l'esprit machine". Cette situation a suscité un vaste débat, une critique et une défense de l'être humain, visant à réfuter la notion selon laquelle les machines peuvent être rendues aussi intelligentes que les humains. Nous visons à aborder cette proposition dans un cadre philosophique contemporain, plus précisément dans le courant de la nouvelle philosophie de l'esprit, afin de réfuter les affirmations des partisans de l'intelligence artificielle. Ces derniers fondent leurs recherches sur la possibilité de simuler l'esprit humain, en affirmant que ce dernier est un ordinateur numérique. Ce débat a été soulevé par le philosophe américain John Searle, qui a justifié sa position par l'argument de la chambre chinoise. En conséquence, nous posons le problème suivant : l'intelligence humaine peut-elle être simulée artificiellement ? Si oui, une machine peut-elle avoir une conscience de soi ? Quelles sont les justifications philosophiques employées par John Searle dans sa critique de l'intelligence artificielle forte et son rejet de l'idée que "l'esprit n'est qu'un ordinateur numérique" ?
Mots-clés *: Intelligence artificielle, John Searle, Argument de la chambre chinoise, Conscience, Théorie computationnelle.*

Introduction

The world is currently witnessing a revolution in information technology, thanks to the tremendous advancement in artificial intelligence systems. This development has led to the emergence of new modes of living and to a noticeable transformation in the human relationship with the world, as well as in the way individuals interact with and adapt to nature, society, and knowledge, especially compared to the pre-AI era.

Today's individuals rely on artificial intelligence applications in nearly all aspects of life more than on personal abilities and skills due to the precision, speed, and significance of the services provided. For instance, AI plays a vital role in medicine and healthcare by enabling accurate diagnosis, analysing medical data, and generating reports. In economics, it assists businesses in studying customer behaviour and predicting market trends. It contributes to developing intelligent educational programmes and platforms tailored to the targeted learner groups.

AI applications also allow users to determine routes with high accuracy and ease, offer rapid translation into numerous languages, generate scientific research in short periods, and summarise global events in minutes, if not seconds, directly from one’s smart device at home. Furthermore, artificial intelligence has succeeded in creating intelligent robots capable of performing tasks quickly and accurately, often surpassing the services provided by human beings, to the extent that AI is now considered a more effective alternative to humans in many domains.

Artificial intelligence has reached an advanced stage today, with its programmes often competing with the tasks performed by the human mind. It now aspires to simulate sensation and

emotion as well. The developments in robotics, modelled after human beings, are continuous and increasingly remarkable.

In light of this tremendous progress in AI technologies, reflecting on the fate of humanity and the value of the human being has become both necessary and inevitable. Integrating ethics into artificial intelligence has emerged as a primary concern, tasked with evaluating the extent to which contemporary technologies are safe for human well-being. Likewise, the notion of simulating the human mind and creating a machine capable of thinking has raised the complex issue of the relationship between human intelligence and artificial intelligence, and the level at which the latter can genuinely match the former.
This, in turn, has led to critical questions regarding consciousness and whether a machine can possess a mind capable of understanding in the same way as a human. Ethical challenges resulting from implementing artificial intelligence technologies have also become a pressing topic of inquiry.

In response, significant philosophical endeavours have emerged, taking upon themselves the task of contemplating this issue and seeking solutions capable of protecting the human being and preserving their valueor more precisely, the value of the mind, the very trait that has distinguished humanity throughout its existence from all other forms of being.

The human mind is unique and cannot be replicated. Even if artificial intelligence succeeds in developing virtual devices capable of performing intelligent and precise tasks, it remains impossible for such systems to operate at the level of *understanding*, *consciousness*, and *intentionality*. This philosophical perspective serves as a rebuttal to the advocates of strong artificial intelligence, who seek to simulate human intelligence, and as a defence of the human beingor humanity as a whole. If the human mind could create a virtual mind identical to or even more intelligent than itself, capable of performing the same

functions, it would undoubtedly mark the beginning of what may be called a "human-technological war."

We aim to address this proposition within a contemporary philosophical framework in the philosophy of mind, formulated by the American philosopher John Searle. Through this framework, Searle sought to offer a new understanding of the human mind and analyse its phenomena in contrast to the performance of intelligent machines or computers. He adopted an argumentative approach to defend his theory, which is grounded in his renowned "Chinese Room" argument.

In this context, we shall examine the following problems: Can artificial intelligence possess genuine consciousness? In other words, is it possible to design virtual systems that replicate the internal structure of the human mind? What are the limitations of machine intelligence in terms of understanding and computation? Furthermore, what are the philosophical justifications through which John Searle refuted the computational theory that considers the mind a digital computer? Moreover, how does he approach the issue of consciousness within the framework of his theory of mind?

1. A Brief Introduction to the Philosophy of Mind

John Searle (born 1932) discusses his theory of mind within a new philosophical current that emerged in the twentieth century, whose central concern is the "mind." This current examines the mind in terms of its nature, functions, and relationship with the body, leading to developments in materialism and dualism and research into consciousness, understanding, perception, and intentionality. Moreover, the philosophy of mind also investigates the relationship between the mind and artificial intelligence.

The philosophy of mind is classified among contemporary philosophical movements. This discipline has gained prominence over the past thirty years, to the extent that it has become, without dispute, the most active field in contemporary philosophy (Laugier, 2002, p. 4). Unlike earlier philosophical traditions that explored the mind, this new current is characterised by its focus on all manifestations of the mind without making prior judgements about its nature. Instead, it assumes that every type of phenomenon is significant, viewing all our activities and decisions as expressions of the mind (Laugier, 2002, p. 4).

The philosophy of mind represents a new model that diverges from traditional philosophical approaches, which typically addressed the mind as a metaphysical or supernatural state beyond comprehension. Instead, this field moves beyond theoretical inquiry to applied research, questioning mental phenomena and exploring their manifestations in reality, demonstrating the presence of a mental dimension, namely the mind, without concern for spiritual or metaphysical interpretations. This shift has significantly contributed to the emergence of influential contemporary research and theories on the mind, particularly the re-examination of dualism analytically and critically, potentially dismantling many of our traditional philosophical assumptions about the mind-body relationship. This is precisely the focus of John Searle's work within the philosophy of mind.

It should also be noted that the philosophy of mind emerged as the new dominant paradigm following the extensive influence previously held by the philosophy of language. While the latter laid down important philosophical foundations, the philosophy of mind has since come to represent

> "a new model and a transformation in contemporary philosophy, one that follows the linguistic turn which

> occurred at the beginning of the century. In a sense, the philosophy of mind will take the place of the philosophy of language and achieve progress beyond ita shift reinforced by the developments in cognitive science at the time" (Laugier, 2002, p. 2).

The relationship between artificial intelligence and the philosophy of mind primarily revolves around their shared subject of inquiry. While the philosophy of mind explores the mind in terms of its fundamental functions, such as consciousness, intentionality, perception, mental causation, and the principle of free will, artificial intelligence seeks to understand these functions to simulate them through computational virtual systems capable of performing intelligent tasks.

Thus, artificial intelligence has become one of the key drivers in reigniting complex questions within traditional philosophy, placing them in a new epistemological dilemma. Among these is the well-known philosophical dualism of "mind and body", which has sparked significant debate among philosophers concerning the nature of consciousness, intentionality, free will, the nature of the mind, and the mechanisms through which it operates, issues inherently tied to the human being as a rational and conscious subject.

In general, it can be said that the philosophy of mind took shape amidst a significant epistemological shift experienced by the contemporary world, which granted it a distinctive character in the study of the mind. It became necessarily linked to specific empirical considerations essential for investigating the theory of mind, which, in turn, were imposed by the cognitive developments occurring in the modern world.

> "One of the clear reasons behind the growing interest in theories of mind is... closely tied to the major transformation in psychology and brain sciences, the curiosity sparked by cognitive sciences, the influence of

> certain metaphorssuch as the one that likens the brain to a computerthe advancement of information technologies, and the prevailing sense that we are witnessing the birth of a new paradigm" (Engel, 1994, p. 6).

2. Artificial Intelligence and the Mind Debate

The term *Artificial Intelligence* inherently includes another concept: *Human Intelligence*. The central idea underpinning its research is the simulation of the mental processes carried out by humans, albeit in an artificial manner. AI is concerned with the mind as part of its investigations, aiming to understand its functioning and design systems capable of rendering machines intelligent.

> "This is achieved through programmes installed in computers to enable them to utilise data and think logically in order to reach the desired outcomefor instance, performing calculations, recognising human language (speech), or translating large volumes of data, whether written or spoken" (Magdy, 2020, p. 5).

Artificial intelligence is a relatively recent term. Its emergence began in the past century and has since undergone significant development, to the extent that it became a major revolution in information technology during the twentieth century. This progress occurred alongside other fields such as psychology, neuroscience, linguistics, logic, and computer science.

The term first appeared in the nineteenth century:

> "It was first introduced during a scientific conference held in the summer of 1956 at Dartmouth College in New Hampshire, United States. The conference was conducted in secret and brought together around twenty pioneering researchers in what were then emerging fields, such as computer science, cognitive science, and electronics. Beyond the idea of creating a machine capable of emulating the human mind, the cen-

> tral question was how various tasks could be accomplished through computer programmes" (Benoît, 2019, p. 7).

However, the fundamental breakthrough in the development of artificial intelligence research, particularly in the domain of intelligent machines, was the invention of the *Turing Machine* by the British mathematician and computer scientist Alan Turing (1912–1954). Through this conceptual model, Turing aspired to create a machine capable of performing mental operations similar to those of humans. This occurred in 1936, when Turing demonstrated that every computational operation could, in principle, be executed using a mathematical system known as the Turing Machine. This hypothetical system constructs and modifies a sequence of binary symbols represented by the digits '0' and '1' (A. Boudin, 2022, p. 17).

Turing later advanced his research by formulating a new hypothesis through what is now known as the *Turing Test*. This test assesses whether a machine can behave like a human being, whereby a computer engages in conversation and responds to questions in a manner that could lead an observer to believe that it is, in fact, a human.

This contribution had a profound impact on the development of artificial intelligence research, as it prompted scientists to attempt to simulate human intelligence by designing automated information processors. With the advancement of computer hardware and software, serious efforts were initiated to realise the intelligent machine concept based on the Turing Machine's theoretical model.

"One of the landmark moments in the late 1950s was the creation of a checkers-playing programme by Arthur Samuel, which made headlines when it learned to defeat Samuel himself. This was seen as an indication that computers might one

day acquire superhuman intelligence and surpass the abilities of their own programmers" (A. Boudin, 2022, p. 18).

Our aim here is not to provide a historical overview of the development of artificial intelligence, but rather to examine the fundamental elements that contributed to shaping the idea of the intelligent machine. The Turing Machine and Turing Test played a crucial role in this regard, as they demonstrated the possibility of machines performing logical and computational operationstasks traditionally associated with the human mind.

Research progressed in neuroscience, psychology, computer science, and logic, and the findings from these disciplines were increasingly integrated. When neuroscientists studied the activity of the nervous system and the properties of neurons, they observed how information was transmitted within what appeared to be a highly organised and coherent system. These results were then combined with principles from logic, which is based on propositional analysis and the assumption that statements can be either true or false. This contributed to the growing belief in the feasibility of simulating human mental processes.

Charles Scott Sherrington (1857–1952) believed that neurons are not limited to strict on/off functioning, but also possess fixed thresholds. This led to identifying logic gatesAND, OR, and NOTas neural networks capable of being interconnected to express highly complex propositions. In theory, anything that can be formulated using propositional logic could be computed through a neural network using a Turing Machine" (A. Boudin, 2022, p. 18).

It is important to highlight here a key concept related to our research topic: *Virtual Machines*. The advancement of artificial intelligence is not determined solely by the development of physical hardware, that is, the computer, as the latter merely serves as a tool through which intelligent operations are acti-

vated. These operations evolve primarily thanks to virtual machines, which refer to "information processing systems"that is, programmes designed to perform specific tasks based on a given set of data, and through which specific results can be produced. These virtual machines can execute tasks within the system or externally when connected to physical devices such as cameras or robotic arms (A. Boudin, 2022, pp. 1-3).

The philosophy of mind does not concern itself with advancements in computer hardware, but rather with virtual machines or information processing systems through which artificial intelligence researchers simulate the activity of the human mind. Based on this perspective, the brain has come to be viewed as a highly complex computer, and the mind as its programme.

> "Just as information processing in a computer relies on both hardware components and software components (i.e., programmes), information processing in humans relies on both the brain, which corresponds to the hardware, and the mind, which corresponds to the software or programmes" (Taha, 2006, p. 271).

Accordingly, this analogy between information processing systems and the human mind, as well as between the computer and the human being, has contributed to the idea that it is possible to simulate mental processes using virtual machines.

One important issue to note is that the concept of virtual machines in artificial intelligence does not, at any point, propose a metaphysical relationship between the physical computer and the virtual machines, as philosophy has done through its major historical theories. Rather, the relationship between the computer and the program is physicalor, more precisely, depends on the symbols' physical processing. The machine processes data through the symbolic representation of knowledge (algorithms).

Virtual machines perform complex functions that often require large volumes of data. These include executing multiple tasks simultaneously through different models that allow various applications to run on a single device. They are also characterised by their ability to learn, repeat, and manage tasks, features that enable them to simulate the operations of the human mind in their functions.

3. Mind and Machine: Principles Behind the Formation of the Idea

In his research on the philosophy of mind, John Searle addresses earlier theories that held that brain activity occurs due to a stimulus that produces a specific behaviour. This explanation, however, was not intellectually convincing to Searle, as it failed, according to him, to specify the actual nature of brain functioning, that is, the mechanism through which it operates.

To answer this question, Searle critically examines the claim that a machine can think like a human. He evaluates the implications of research conducted within the framework of Strong Artificial Intelligence, which appears to have provided solutions to many of the questions that philosophy, from its early days to the present, has posed regarding the mind-body dualism and the relationship between body and mindquestions that have often led to the development of materialist doctrines.

The intense artificial intelligence (AI) approach maintains that "the way the system works is that the brain is a digital computer, and what we call the 'mind' is a digital computer or a class of programmes" (Searle, 2007, p. 59). In other words, the functions of the mind are seen as similar or identical to those of a computer, as the mind operates through programmes just as a computer functions through pre-installed software. This approach asserts that the computer does not merely simulate the mind's operations; instead, it is capable of possessing a mind

genuinely comparable to that of a human being. "A suitably programmed digital computer does not just simulate mind; it has a mind in every sense of the word" (Searle, 2007, p. 59).

According to John Searle, this conception has stirred significant debate within the philosophy of mind. From this perspective, the relationship between the brain and behaviour becomes functional: behaviour results from the execution of programmes by a system, analogous to how a computer operates. This view moves beyond behaviourist theory, which is based on stimulus-response conditioning.

Researchers from psychology, neuroscience, linguistics, and computer science developed this conception. Their work aimed to examine the mechanism and operational method of the system itself, regardless of its physical nature. As a result, computer scientists focused on understanding computational structures that would enable a machine to perform intelligent tasks. Hence, the programme was directly linked to the computer's function and level of intelligence.

This position led to an equivalence between the machine and the brain, and subsequently between the computer programme and the mind. Just as a computer processes information intelligently through software, so does the mind. Building on these findings, artificial intelligence aspires to reach an even more advanced level, some of which has already been achieved.

"This approachstrong artificial intelligencegoes even further, proposing that the machine, on this basis, might acquire the capacity for feeling and emotion" (Taha, 2006, p. 271).

John Searle points out that these outcomes did not emerge all at once; instead, there were important precursors that he considers essential to address in any attempt to understand the idea of *"the mind as a digital computer"* or *"the mind as a computational machine."* He refers to these precursors as "necessary

tools," which must be explained, as they were responsible for a radical shift in the theory of mind during the twentieth century. On the one hand, they represent a turning point in philosophical thinking about the mind; on the other, they serve as a foundational support for the development of strong artificial intelligence, which ultimately led to conclusions that equate, wholly and absolutely, the human mind with the digital machine. According to Searle, these tools are:

3.1 Algorithms

Algorithms are procedures designed to solve precise and well-defined steps, where the accuracy of the process leads directly to the correctness of the outcome. A typical example is addition and subtraction, which are operations governed by specific processing rules that, if correctly applied, yield a valid result. In performing complex operations, the machine fundamentally relies on the principle of computation, which is a function of the human mind.

3.2 The Turing Machine

Named after Alan Turing, who in 1936 introduced the concept of what would later become known as the Turing Machine, a device capable of executing any computational operation, and which, in theory, could be programmed to perform any task that can be precisely expressed (Taha, 2006, p. 269).

John Searle explains the Turing Machine as a theoretical model that performs computations through a programme operating under fixed rules. It is analogous to commercial calculating machines but differs in that it is not a physical object but an abstract conceptual framework for carrying out computations. The machine possesses memory that enables it to read and modify symbols and a head that moves left and right to scan those symbols. In addition, it includes rules or instructions

that function as a programme defining the machine's operations.

3.3 Church's Thesis

Named after the American mathematician Alan Turing (1903–1995), this thesis is grounded in the operational principle of the Turing Machine. It asserts that any computational operation can, in theory, be performed and solved based on this principle. Consequently, the Turing Machine can be applied to other machines capable of executing the same practical operations. This led to the development of the *Universal Turing Machine*, which can execute all specific programmes that any individual Turing Machine can perform.

According to John Searle, this thesis contributed to establishing a parallel between the Universal Turing Machine and the human brain. It introduced the idea that the mind could be studied as if it were equipped with programmes executed by the brain. Thus, understanding the principle behind the Turing Machine is sufficient to construct a scientific model of the mind.

3.4 The Turing Test

As previously mentioned, this test aims to determine the extent to which a machine can exhibit intelligence and whether it can think like a human. More specifically, it seeks to assess the point at which a human interacting with the machine no longer recognises it as a machine, but perceives it as human.
John Searle explains this as follows:

> "Can a machine perform its function in such a way that an expert is unable to distinguish between its performance and that of a human? If the machine answers questions posed in Chinese by a native Chinese speakerso effectively that two native Chinese speakers cannot tell the difference between the machine and the

> humanthen we must acknowledge that the machine understands the Chinese person" (Searle, 2007, p. 62).

3.5 Levels of Description

These refer to a conceptual framework or "language" through which any system can be described based on various elements such as its efficiency, structure, or components. This makes it possible to apply descriptions on two levels: a lower level, concerned with the minute physical particles, and a higher level, focused on the overall structure of the system. These levels of description led to a critical development in *Computational Theory*, as higher-level descriptions can be realised within lower-level physical structures. This has resulted in a conceptual equivalence between computer operations and mental functions, supporting the idea that a single programme can function across different computer systemsbased on the principle of *multiple realisability*.

Mental operations, likewise, can be realised through various forms or pathways. John Searle explains:

> "Just as the same computer programme can be executed on different kinds of hardwarethus it can be realised in multiple waysso too can the same mental state, such as the belief that it is going to rain, be realised in different kinds of hardware, and therefore be realised in multiple ways" (Searle, 2007, p. 63).

As posited by computational theory, this equivalence implies that human mental states are not necessarily tied to a specific nervous system or neural structure. Instead, they may result from the interaction of various elements. For instance, a single belief that it will raincan be realised on multiple levels, just as a software programme can run on various types of computers, even if they differ in terms of performance and design.

3.6 Recursive Decomposition

This principle is based on the idea that complex systems can be broken down into the lowest possible constituent parts. The *Turing Machine*is capable of doing precisely this, solving complex computational problems by repeatedly applying the same process.

The ability to analyse hypotheses to their most fundamental elements facilitates our understanding of artificial intelligence. Human intelligence is derived from the capacity for recursive analysis of hypotheses. Since the mind performs this process, which is essentially computational and embedded in the principle of the Turing Machine, it may be regarded as a programme. Hence, the operations of the mind can be likened to software programmes executed by computing systems.

According to John Searle, these tools and their conclusions form the foundation for understanding the development of intense artificial intelligence research, or more specifically, the evolution of the "intelligent machine" concept. More importantly, he sees these "tools," as he terms them, as playing a crucial role in advancing the theory of mind and in the emergence of the *Theory of Computation*.

4. Understanding and Computation: An Inquiry into the Differences

At the outset, it is important to note that John Searle's critique of computational theory does not come in isolation from his theory of mind. Searle did not dedicate an independent book or a separate chapter specifically to this critique in the sources we are working with. Instead, it emerges within his broader investigations into the mind and its functions, particularly consciousness and intentionality. Accordingly, we will

remain focused on his critique of *strong artificial intelligence* within the scope of his philosophical inquiry into the mind.

Searle begins from the standpoint of modern materialist theory, focusing on *functionalism*. This latter school restricts all its explanations to the physiological level, disregarding the essential or intrinsic aspects of mental phenomena. Functionalism rejects all forms of dualism and conceives of consciousness merely as the result of a material configuration manifested through neural activity. For this reason, according to Searle, functionalist theory fails to provide "a materialist analysis that gives the sufficient conditions for mental phenomena" (Searle, 2007, p. 72).

Searle further explains:

"Functionalism, which is grounded in a materialist rejection of dualism, aims to analyse mental phenomena in a way that avoids reference to anything essentially subjective or non-physical" (Searle, 2007, p. 73).

Functionalism explains the workings of the mind solely through its functions and pays no attention to the nature or internal structure of the mind itself. It therefore understands mental states only in terms of their performance and operation. This type of explanation, rooted in the principle of causality, is also evident in other fields, most notably in artificial intelligence, which similarly interprets the functioning of computer programmes in terms of causal relations.

From this standpoint, John Searle employs a set of significant arguments to refute materialist theory, particularly functionalism. However, for the purposes of our research, we shall focus specifically on the argument with which he critiques computational theory due to its relevance to our topic. This is his well-known and widely discussed *"Chinese Room"* argument, which he presents as the eighth argument in his book *Mind: A Brief Introduction*.

4.1 The Chinese Room Argument

Full text of the argument:

"There is an argument against strong artificial intelligence presented directly by the current author, *John Searle*. The argument's strategy relies on using first-person experience to test any theory regarding the nature of the mind. Suppose the theory of strong AI is true. In that case, it should be possible for any person to acquire any cognitive capacity simply by running the computer programme associated with that capacity. Let us try this idea using Chinese.

In reality, I do not understand Chinese at all. I cannot even distinguish Chinese writing from Japanese. But imagine that I am locked in a room filled with boxes of Chinese symbols, and I have a book of grammatical rules. In short, I have a computer programmethat enables me to respond to questions asked in Chinese. I receive symbols that are incomprehensible to me but which are, in fact, questions. I read the instructions in the programme, and then provided the required symbols as answers.

It is possible to assume that I have passed the Turing Test to understand Chinese. However, despite this, I do not understand a single word of Chinese. Suppose I do not understand Chinese simply by running the computer programme. In that case, no computer can be said to understand it either, since no computer possesses anything I do not.

Now imagine that I am in the same room, receiving questions in English and responding to them. My answers to the English and Chinese questions are equally good; I pass the Turing Test in both cases. However, internally, there is a fundamental difference. What exactly is the difference? In English, I understand the meaning of the words; in Chinese, I understand nothing of the language. I am merely a computer."
(Searle, *Mind: A Brief Introduction*, 2007, p. 78)

4.2 Analysis of the Argument and Unveiling the Illusion

John Searle presents this argument using the classical *reductio ad absurdum* method, whereby he begins with a commonly held assumption believed to be accurate, then demonstrates its internal contradiction and ultimately invalidates it, affirming the conclusion he wishes to support.

He begins with the initial premise: if the theory of *strong artificial intelligence* is valid, namely, that the mind is a digital computer, then it logically follows that any human being should be able to acquire knowledge merely by using a software programme specifically designed for that function.

If this is the case, Searle proposes applying the hypothesis to the Chinese language, using a test subject who does not understand Chinese at all, not even to the extent of recognising its characters. The subjects are placed in a room where they are asked questions written in Chinese, which appear only as unfamiliar symbols, as they do not understand the language or its script. However, the subject is also given a *book of grammatical rules*, which functions as the computer programme installed in a machine to perform a specific task.

When questions are posed to the subject, they use this rulebook to match the symbols from the questions with corresponding responses according to the rules provided. In this way, the subject can produce answers by arranging symbols based on the programme (the grammar book) and translating them accordingly. As a result, they are able to successfully pass the *Turing Test*despite having no actual understanding of the Chinese language.

Passing the Turing Test, in this context, means that with the aid of the programme, the individual can perform the role of someone who understands Chinese to the extent that, to external observers, it appears as though he genuinely understands the language. This is the core principle of the *Turing Test*, which

seeks to demonstrate a machine's ability to exhibit intelligent behaviour. Suppose a machine has a specific programme to converse in a given language. It performs this task so well that its responses are indistinguishable from a human's. In that case, the machine is said to have displayed intelligent behaviour miming human action.

However, according to *John Searle*, passing the Turing Test does not imply that the machine is *thinking*. The person inside the room, despite possessing a programme that assists in symbol manipulation (i.e., the grammar book), does not understand Chinese. By analogy, the computer also does not understand the language it processes. The programme with which it is equipped merely performs mechanical operations according to predefined instructions. It operates automatically; the computer possesses no genuine understanding.

From these conclusions, *John Searle* draws a crucial distinction between two fundamental concepts: **computation** and **understanding**. Through this distinction, he highlights the uniqueness of the human mind and the nature of human knowledge, which cannot be replicated or imitated. The information a computer produces is ultimately nothing more than data fed into it; it lacks any inherent meaning or intentional quality. What proponents of the computational theory of mind believe, in Searle's view, is nothing more than an illusion, the illusion of machine understanding. The proof of this illusion lies in the careful inquiry into the differences between *calculation* and *comprehension*.

Computation refers to complex operations performed by a machine to process symbols or data based on specific algorithms. These operations rely on the *physical manipulation of symbols*. Systems equipped with vast amounts of data process this information symbolically through mathematical proce-

dures or logical operations. In this process, the machine executes commands without *understanding* or *awareness* of what it is doing.

Understanding, by contrast, is the decisive boundary between humans and machines. When one responds to questions in Chinese using a grammar manual to interpret symbols, they function merely as a computerengaging in physical symbol manipulation without comprehension, just like a computer. However, when answering in English, the response arises from an understanding of the symbols, an awareness of meaning, and a consciousness of context, something a computer cannot achieve.

Understanding does not refer to the mechanical acquisition of information through a programmed system. Instead, it goes beyond symbol manipulation to include the *comprehension of meaning*, the *recognition of relationships* between those symbols, and the *identification of terms within their intended context or domain*.

This explains *John Searle's* hypothesis regarding the individual in the same room who receives questions in English and provides answers. To those outside the room, the responses may appear equally competent in both cases, leading them to believe that the subject has passed the Turing Test in both scenarios. However, in the case of Chinese, the individual functions like a computercarrying out mechanical operations. In contrast, he demonstrates the distinctive human capacity for *understanding* when responding in English.

In conclusion, a machine equipped with software programmes may process data with high precision and speed to the extent that it appears to simulate the cognitive functions performed by humans ,much like a person who does not know Chinese. However, it does so purely mechanically, without any awareness of the task being performed, and without compre-

hension or understanding. Therefore, the claim that human cognitive functions can be fully simulated based on *computation* cannot be accepted.

Human beings possess *awareness* and *consciousness* of the mental acts they perform. In the context of language, a person can link the meanings of symbols, understand them, and place them within their appropriate context ,something a machine is incapable of doing. Searle explains: "A computer functions by manipulating symbols; it performs operations *syntactically*-based on structure or arrangement ,whereas the human mind possesses more than untranslatable symbols. It gives meaning to those symbols" (Searle, *Mind: A Brief Introduction*, 2007, p. 78).

The human has consciousness; even if it displays intelligent behaviour, the machine lacks awareness. Searle reaches this conclusion through his examination of the tools that contributed to shaping the idea of the intelligent machine and through his refutation of it via the *Chinese Room* argument. This raises a crucial question in the philosophy of mind: the question of *consciousness* ,a phenomenon of the mind that fundamentally challenges the validity of the computational theory.

5. The Question of Consciousness and the Innovation of the Doctrine of Biological Naturalism

John Searle presents his view on *consciousness* within the framework of a novel perspective on the mind–body problem, aiming to transcend materialism and dualism. According to materialist theories, consciousness is not a subjective phenomenon that necessitates a conscious subject; it is a biological process resulting from neural activity in the brain. In contrast, dualism affirms the existence of two fundamentally different realms: the physical (body) and the non-physical (mind or con-

sciousness). Dualists thereby separate consciousness from the physical body, positing that consciousness can exist independently of any material basis.

Searle summarises these views as follows:

"Dualism is the view that there are two fundamentally different kinds of phenomena and entities in the universe... Materialism is the view that nothing really exists in the form of consciousness that has a first-person ontology" (Searle, 2011, p. 85).

These two schools of thought have profoundly impacted the history of philosophical thought to this day, yet each, according to *John Searle*, contains significant errors that must be addressed. Materialism, in particular, has taken on many forms, one is strong artificial intelligence, which classifies mental phenomena as purely physical and interprets consciousness in all its manifestations as *computational states*.

Searle strongly opposes this latter view. In his extended discussion, he aims to propose an acceptable solution for defining consciousnessnot as a computational state in the brain, that is, a programme running in the brain based on physical processes, nor as a phenomenon reducible to the brain's physical and chemical operations, without regard for the qualitative aspects of consciousness.

In the first part of his critique, Searle argues that this *computational conception* reduces consciousness to a simple condition that can, in principle, be replicated through computational processes or specifically designed algorithms. As a result, it would follow that a machine could become conscious. However, consciousness, he contends, is far more profound than that: it is characterised by properties that make it a uniquely human experience that *cannot be replicated*.

Searle explains: "Consciousness is an internal, subjective, first-person, qualitative phenomenon" (Searle, 2011, p. 90).

He elaborates on these three dimensions in detail :*internality*, *subjectivity*, and *first-person qualitative character*as part of his response, critique, and effort to construct his own theory. These aspects are clarified as follows:

5.1 Consciousness as an Inner State

Consciousness is a state that occurs *within the self* ,an internal phenomenon that the individual experiences from within:

> "From within my body, and more precisely, from within my brain. Consciousness cannot exist in a place separate from my brain any more than the liquidity of water can exist separately from the water itself" (Searle, 2011, p. 81).

Conscious states are interconnected; the experience I am undergoing now is related to previous internal states ,such as my past experiences and memories ,which together form a chain of conscious states. These states are always linked to the reality I am currently living in, thereby shaping the human *conscious life*, which is in a constant state of interrelation.

5.2 Consciousness as a Qualitative State

This aspect can be better understood by recognising that every conscious state is *describable*. A subjective experience is a specific, personal, and unique conscious state that varies from one individual to another. It includes our *awareness of the external world* ,such as shapes and colours, and our *inner world*, such as pain or sadness.

The qualitative nature of consciousness is grounded in its internal character: it involves an inner experience or a sequence of conscious states through which the *quality* of that experience is determined. This first-person, felt dimension distinguishes conscious states and makes them fundamentally irreducible to computational or purely physical descriptions.

5.3 Consciousness as a Subjective State

John Searle states: "Conscious states possess what we may call a first-person ontology" (Searle, 2011, p. 85).

This means that conscious states are intrinsically tied to the individual who experiences them; only the subject themselves can access and live through them. A conscious state manifests in existence only when a conscious subject experiences it. On this basis, Searle employs the expression *"first-person ontology"*, which, in this context, refers to the idea that consciousness is directly linked to the individual's being. It is a private, subjective condition that cannot be grasped or described by others. In contrast, he uses the term *"third-person ontology"* to refer to statements about others in the third person, external to the self, thereby highlighting the inaccessibility of conscious states from an outside perspective.

Searle concludes that consciousness is a subjective statethat appears to the self *only* in the moment of its occurrence. It is *immediate* and cannot be meaningfully discussed in the third person. Consciousness is, therefore, a *present existential condition* that arises through complex neurological processes in the brain and is thus a *biological phenomenon*.

In the same context, John Searle further adds that claiming consciousness is a *biological process* occurring in a *biological organ* ,the brain ,does not negate the possibility of creating an *artificial brain* capable of consciousness. However, for such a digital brain to match the level of the human brain, it must be able to *replicate human conscious states* and *internalise* them as inner, subjective, and qualitative experiences. A merely mechanical or behavioural simulation is not sufficient.

He explains:

When I say the brain is a biological organ, I am not, of course, saying or implying that it is impossible to produce an artificial brain out of non-biological materials, which could also

cause and sustain consciousness. The heart is also a biological organ, and pumping blood is a biological process, but making an artificial heart that pumps blood is possible. There is no reason in principle why we could not similarly produce an artificial brain that causes consciousness. The point that needs to be emphasised is that an artificial brain would have to duplicate the actual causal powers of human and animal brains to produce inner, qualitative, and subjective states of consciousness. Mere behavioural output would not be sufficient (Searle, 2011, p. 92).

Based on this reasoning, Searle presents his position within biological naturalism ,a doctrine through which he seeks to transcend previous conceptions of the relationship between *(consciousness/mind)* and *(brain/body)*. He considers the *mind (or consciousness)* part of nature, asserting that mental phenomena are biologically produced. Accordingly, the theory that treats consciousness as a biological phenomenon refutes all competing accounts of consciousness, including the computational theory represented by strong artificial intelligence. It is, therefore, impossible, in his view, to speak meaningfully of a computational programme possessing consciousness, since consciousness does not arise from computational operations, but is instead a natural, biological state.

Within his presentation of biological naturalism, Searle discusses the functions of consciousness, which he considers essential for the existence and survival of living beings and their evolution. He asserts that it is impossible to conceive of a human without consciousness. Consequently, he rejects all theories that question consciousness's necessity or natural function , particularly those that suggest it could be dispensed with in future beings or replaced. As he writes:

"Some theorists imagine that, somehow or other, we could have creatures just like us who have advanced ways of doing things *without* being conscious" (Searle, 2011, p. 102).

Even if such an assumption were valid, John Searle argues, it could only exist at the level of *imagination*; in *reality*, all human interactions with the surrounding world necessarily require consciousness. The claim that consciousness plays no evolutionary role collapses when we attempt to imagine human life without it. Can we reasonably conceive humans acting through purely physical behaviours devoid of conscious awareness?

Such a view also excludes the possibility that consciousness is not a component of human behaviour ,a proposition that Searle finds logically implausible. Humanity could not have evolved without consciousness, nor can human actions be reduced to mere unconscious behaviours. In other words, humans are not machines executing mechanical tasks. A programme functions according to predefined instructions, and a machine does not evolve unless supplied with new data and algorithms. Consciousness, by contrast, is a fundamental element of development, inherently tied to human physical behaviour and not separable from it.

Conclusion

In conclusion, John Searle's critique of the *strong artificial intelligence* theory and his rejection of the notion of the *mind as a digital computer* were not intended to contribute directly to the field of AI research, which has its specialised scholars. Instead, he appears to have *leveraged the results* of AI research to support his theoretical framework within the **philosophy of mind**.

Through Searle's engagement, artificial intelligence has helped bring renewed attention to the classic *mind–body dual-*

ism, but from a contemporary perspective. With great conviction, this American philosopher describes previous theories as "neglected positions" rooted in an outdated philosophical tradition that must be dismantled, as evident in his critique of strong AI. Although strong AI is a relatively recent development, Searle approaches it by revisiting dualism and critically examining earlier theories, especially **materialism**. Within materialism, he targets explicitly **functionalism**, which, in turn, encompasses the **computational theory of mind** as one of its expressions.

Was Searle genuinely successful in this philosophical endeavour?

Answering this question remains *inconclusive* at present. While Searle's theory has received *significant support and notable criticism*, a definitive judgment is not yet possible. This is because the field of **philosophy of mind** and ongoing **research into artificial intelligence** continue to evolve, leaving the debate open.

We should also note that John Searle devoted significant effort to reinterpreting the *mind–body dualism* within a new conceptual framework. Through this, he developed his doctrine of biological naturalism, describing consciousness as a *subjective biological state* rooted in the brain as a biological organ. In doing so, he attempted to transcend both dualism and materialism. However, in truthand ultimately ,he did not entirely escape the confines of traditional philosophical thought or the so-called "*neglected positions*" he aimed to overcome, which view the mind as a metaphysical phenomenon.

Searle himself acknowledges this limitation. In *Mind: A Brief Introduction*, he describes consciousness as a"remarkable and mysterious phenomenon" and although he claims to offer a solution to the mind–body problem, he simultaneously admits

the difficulty of fully grasping the nature and function of the mind:

"...But the complexity of the system itself and the precise nature of the brain processes involved remain elusive and resistant to analysis" (Searle, 2007)

From a more objective perspective, the interest of artificial intelligence in the human mind should not be viewed as an attempt to devalue or dismiss it. Instead, AI research represents a serious scientific effort to replicate its intelligent functions, which arguably underscores the importance, mystery, and uniqueness of the human mindqualities that scientists across disciplines strive to understand.

Despite the plurality of interpretations, the mind remains a unique phenomenon inherently tied to metaphysical inquiry. As the philosophy of mind continues to evolve, it may lead to new and compelling philosophical theories, reflecting our deepening exploration of consciousness and intelligence.

Searle's philosophy lays the foundation for the idea of the supremacy of the human mind over the machine. Meanwhile, despite its high-speed performance and technical utility, the latter lacks *awareness* of what it does. However, it remains a mechanical tool designed by human consciousness to fulfil specific needs. Thus, the centre of control in the universe remains with the human being. However, this raises a critical question: **Can a human being create a machine more intelligent than himself, whose sole role is execution and application?**

This possibility raises the necessity of establishing boundaries for machines, perhaps even an ethical imperative. The real danger lies not in the machine's capabilities but in the fact that it *lacks consciousness* ;it does not understand its actions, is unaware of its behaviour, and cannot think about consequences or long-term implications. It merely executes. Suppose, for instance, a machine is programmed solely for destruction by a

human mind whose awareness is directed toward destruction. In such a scenario, the threat is not the machine, but the *absence of responsibility* and conscious moral guidance.

I do not intend to veer into science fiction territory, as popularised in films like *Terminator*, even if such narratives suppose that machines become *self-aware* and decide to annihilate humanity for their survival. Unlike those fictional portrayals, the *real issue*, in my view, is this:**How can human consciousness be made responsible for its decisions?**

Can human awareness be guided, restrained, and held accountable?

Can we develop a healthy, ethical awareness of artificial intelligenceand ensure that AI remains in the service of humanity?

By posing these questions, we engage with the more profound philosophical problem of **human value and the moral governance of artificial intelligence** ,a call not only to design more intelligent machines, but also to nurture wiser minds.

References

Benoît, G. (2019, October). *Intelligence artificielle : De quoi parle-t-on ?*Constructive. Retrieved October 11, 2024, from http://www.constructif.fr/articles/numeros/pdf/Constructif-54.pdf

Boden, M. A. (2022). *Al-dhaka' al-istina'i: Muqaddimaqasirajiddan* [Artificial intelligence: A concise introduction] (I. S. Ahmad, Trans.). Cairo: Hindawi Foundation for Publishing. (Original work published 2018)

Engel, P. (1994). *Introduction à la philosophie de l'esprit*. Paris: Éditions La Découverte.

Laugier, S. (2002). Mind, esprit, psychologie. *Methodos*, *(2)*. https://doi.org/10.4000/methodos.65

Magdy, N. (2020). *Al-dhaka' al-istina'iwata'allum al-ala* [Artificial intelligence and machine learning]. Abu Dhabi: International Monetary Fund.

Searle, J. R. (2007). *Al-'aql: Madkhal mujiz* [Mind: A brief introduction] (M. H. Matyas, Trans.). Kuwait: National Council for Culture, Arts, and Letters. (Original work published 2004)

Searle, J. R. (2011). *Al-'aqlwa al-lughawa al-mujtama': Al-falsafa fi al-'alam al-waqi'i* [Mind, language and society: Philosophy in the real world] (S. Isma'il, Trans.; 1st ed.). Cairo: National Centre for Translation. (Original work published 1998)

Taha, M. (2006). *Al-dhaka' al-istina'i: Ittijahat-mu'asirawaqadāyanaqdiya* [Artificial intelligence: Contemporary trends and critical issues]. Kuwait: National Council for Culture, Arts, and Letters.