Can AI Really Understand? Semantic Limitations and the Boundaries of Algorithmic Cognition

Mariusz Mazurek

DOI: 10.18355/XL.2025.18.03.06

Abstract
This paper undertakes an interdisciplinary analysis of the problem of understanding and knowledge in the context of contemporary artificial intelligence models. The starting point is the distinction between the weak and strong versions of the AI hypothesis, which reflect two fundamentally different visions of machines' cognitive potential. The article focuses on the question of whether algorithmic information processing can lead to genuine semantic and conceptual understanding. Against the backdrop of classical epistemological theories and the philosophy of language, the paper examines the limitations of the syntactic approach to cognition, in which knowledge is reduced to operations on symbols devoid of reference to the world or meaning. Special attention is given to John Searle's "Chinese Room" argument, which illustrates the fundamental difference between formal symbol manipulation and conscious understanding. The author argues that while artificial intelligence can efficiently perform complex cognitive tasks, it lacks intentionality and the capacity to transcend the symbolic domain.
Key words: artificial intelligence, language, semantics, strong and weak AI, Chinese Room, intentionality, cognition, meaning

Introduction

Contemporary research in artificial intelligence (AI) has a significant impact on how we understand cognitive processes and the role of language in the acquisition and transmission of knowledge. Breakthroughs in fields such as machine learning and natural language processing prompt the question of to what extent computational information systems can imitate, simulate, or fully replicate the cognitive and linguistic abilities characteristic of human beings. The debate centers on two fundamental positions: the weak and strong versions of artificial intelligence.

In the weak AI perspective, computers are regarded as effective tools for studying the human mind because they can model or simulate certain aspects of cognitive processes. According to this view, even if an algorithmic computer system exhibits functional similarities to human cognitive behavior, it does not necessarily imply that the machine actually "thinks" or "understands" in the same way humans do. The primary role of computers, in this case, is to test hypotheses about the nature of the mind and verify theoretical models, which in turn allows researchers to draw methodological inspiration from computer science and apply it to fields such as psychology or linguistics.

In contrast, the strong AI position maintains that sufficiently complex and advanced computer programs can not only simulate but genuinely exhibit mental states, and even be considered "minds" in the full sense of the term. Proponents of strong AI assume that the human brain can be treated as a kind of biological computer (hardware), and the mind as the "software" running on it. Therefore, if an information system could be made sufficiently complex and properly programmed, it could—at least in theory—develop consciousness and understanding comparable to that of a human being.

One of the most influential critiques of the strong AI thesis was presented by John R. Searle. In his thought experiment known as the "Chinese Room," Searle argues that a computer engaged solely in syntactic operations on symbols cannot achieve genuine

understanding (semantics) or intentionality—both of which are, in his view, essential to human cognition. The Chinese Room example illustrates that the mere "correct" manipulation of sets of symbols does not amount to authentic grasp of meaning, which constitutes the foundation of conceptuality in traditional epistemology. A computer, confined to the syntactic level, does not gain access to meaning and does not establish a relationship with the referent that symbols are purported to represent.

The debate between syntax and semantics gains particular relevance in the context of language—both natural and artificial. Within AI-based cognitive science, there is a prevailing view that language can be treated solely as a structure of symbols subject to algorithmic transformations, without the need to refer to the level of meaning. This purely formal model of communication—prominent in many AI applications such as translation systems or chatbots—leads to the question of whether it is appropriate to speak of a machine "understanding" language at all. In contrast, the tradition of the philosophy of language, from antiquity to contemporary theories of communication, holds that language primarily serves as a carrier of meaning and reference, rather than merely a set of syntactic rules. This raises the concern that while computers may be remarkably effective at processing data, they may still fail to grasp the true nature of meaning and context.

This fundamental tension also gives rise to broader epistemological issues. Traditional models of cognition presuppose a subject–object dichotomy, in which the human being, by transcending their own mental states, relates them to external reality. Strong AI, by reducing cognition to symbolic manipulation of data within the machine, seemingly bypasses the external world and any meaningful referential relationship. Critics argue that such a detachment from assigning real-world "reference" to computational outputs contradicts the classical conception of knowledge as always contextually embedded and infused with meaning.

Consequently, these reflections open up a broad spectrum of questions at the intersection of linguistics, philosophy of mind, and computer science. Can a computer operating solely on the level of syntax ever acquire the capacity for semantic understanding? Can a machine, equipped with vast datasets and sophisticated algorithms, produce something that could meaningfully be called "conceptual knowledge"? Or does the essence of cognition, in the philosophical sense, transcend formal procedures and require a subject capable of interpretation, intentionality, and reference to the world?

The aim of this study is to provide an in-depth analysis of these issues, with particular emphasis on the linguistic dimension of artificial intelligence.

The discussion will encompass both the arguments of critics of strong AI and various attempts to defend the thesis that a sufficiently advanced computational system could develop genuine cognitive competencies, including language abilities approximating those of humans. These reflections will be grounded in the contexts of the philosophy of language, epistemology, and cognitive science, in order to demonstrate how the shift toward purely syntactic models of information processing alters traditional conceptions of cognition and communication. Underlying these inquiries is the open question of whether, and how, the reductionist paradigm of artificial intelligence—centered on symbols and algorithms—can be reconciled with the centuries-old philosophical tradition in which conceptual understanding, meaning, and the subject–object relation play a central role.

The Syntactic Approach to Cognitive Processes

Cognitivists primarily view cognition as a process of information transformation. Within this framework, thinking is reduced to computation, and all cognitive activities are understood as the manipulation of symbols according to predefined algorithms (Newell & Simon, 1976). A typical example is the operation of machine translation

programs, in which individual words and grammatical structures are treated as sequences of symbols processed according to clearly defined linguistic and statistical rules. The algorithm—in this case, a computational model—does not "understand" the meaning of words in the human sense but instead transforms them in accordance with a specified set of instructions.

In cognitive science, it is crucial that both human minds and artificial intelligence systems are seen as engaging in algorithmic processes characterized by the sequential execution of instructions. From the perspective of AI, these instructions (or procedural rules) are designed by programmers—humans who define the task and anticipate possible variations of its solutions. In relation to the human mind, cognitivists propose that, just as software governs the operations of a computer, there exists an innate "program" that directs our cognition. This program is believed to be encoded in the mind and to originate externally (e.g., from evolutionarily developed mechanisms, the environment, or some other yet-undefined force). However, this idea remains largely hypothetical and is rarely elaborated upon in mainstream cognitive science research (Thagard, 2005).

A well-known example that illustrates this perspective is John Searle's famous thought experiment, the "Chinese Room." It demonstrates that algorithmic manipulation of symbols—even if it leads to correct responses in the Chinese language—does not entail actual understanding. From the cognitivist point of view, however, this external lack of "meaning" does not pose a fundamental problem; what matters most is that the rules for data processing are executed according to a given algorithm. In other words, the cognitivist model places its emphasis on the computational structure of cognitive processes, largely setting aside questions about the origin of those rules or the deeper (semantic) dimension of meaning (Searle, 1980).

This conception of cognition—where syntactic information processing plays a central role—marks a clear departure from traditional philosophical accounts of understanding. Whereas classical epistemology focused on how the human mind imparts meaning to experienced phenomena, cognitivists focus on a system of operations whose primary function is to transform symbolic configurations into new arrangements (Card, Moran & Newell, 1983). Whether the subject is computer programs or human thinking, the core concern lies in the formal mechanism of information transformation. In this sense, cognitivism offers a radically different explanatory model of cognition, one that makes no room for tracing individual sources of meaning or considering acts of autonomous rule-generation by the subject—instead, the emphasis is placed on the fact that rules already exist and enable algorithmic operations on symbols.

In the cognitivist view of cognition, the key element is the abstract, logical structure of the algorithm, not the specific physical medium in which that algorithm is implemented. This means that a similar set of computations—that is, symbol manipulations—can be performed on a classical computer, a quantum machine, or any other system capable of executing operations in accordance with defined rules (Nilsson, 2007). For example, a text-processing program that detects grammatical errors and suggests corrections can run on multiple platforms (macOS, Windows, Linux), because the core of its functioning—the algorithm—is not tied to any one hardware architecture.

For cognitivists, the essence of this process lies in syntactic order: symbols are processed strictly according to formal rules, irrespective of their potential semantic content (Gentner, 1983). In practice, this means that a system can operate on sequences of signs in any way, as long as it follows the proper instructions. Thus, all cognitive processes are reduced to computations at the syntactic level—what matters is the correct transformation of symbols according to specified procedures, not their interpretation in terms of meaning or reference to the external world. From this

perspective, the ability of a mind or machine to "understand" is replaced by proficiency in information processing—achieving outputs that match expected patterns or goals (Laird, Lebiere & Rosenbloom, 2017).

Two Levels of Information Processing

Within cognitive science, the concept of "information processing" is significantly narrower than its everyday usage. In common understanding, "processing information" includes all manners of gathering, analyzing, and interpreting data in order to gain new knowledge or understanding. However, in the cognitivist model, it is reduced to the manipulation of symbols according to strictly defined rules—that is, to purely syntactic operations devoid of reference to their potential meaning (Piccinini & Scarantino, 2011).

In practice, this means that a system—whether it be the human mind or artificial intelligence—"moves" or "transforms" sequences of symbols in a manner dictated by algorithms. Each symbol is treated as a formal token that need not carry any semantic content. For example, when a machine translation system converts a sentence from Polish to English, it merely performs symbolic operations. To an external user, the output may appear as a meaningful English message, but the internal process within the computer involves matching patterns, phrases, or character sequences without any "awareness" of the intended meaning.

A key characteristic of this conception of information processing is that it does not engage the semantic level and therefore requires no interpretation or understanding of what occurs at the symbolic level. In traditional epistemology, however, assigning meanings, establishing references to reality, and situating knowledge within cultural contexts were considered essential components of cognition. In cognitive science—particularly in approaches influenced by artificial intelligence—the problem of interpretation is either omitted or deliberately relegated to the background.

Moreover, the origin or nature of the algorithms used for information processing is often treated as a secondary issue. It is assumed that such rules are somehow "given" or "pre-existing"—in artificial systems, provided by programmers; in the human mind, resulting from evolution, innate predispositions, or environmental influences. Cognitivists typically do not engage in deep reflection on how these rules actually come into being, from what level of reality they originate, or how they might be modified during the operation of the system. What matters is simply that these rules exist and allow for symbol manipulation in a formally correct way.

This perspective, focused exclusively on the syntactic aspects of processing, effectively reduces the very notion of information in cognitive science to sets of symbols transformed by predefined procedures. The reference to the external world or the realm of meaning—central to everyday conceptions of knowledge and understanding—either disappears or is significantly diminished. For this reason, some scholars criticize this approach as overly simplifying the complexity of human cognition. They argue that true understanding requires going beyond "pure syntax" to include semantics and contextual (cultural or personal) dimensions.

Nevertheless, it is precisely this "purely syntactic" perspective that enables cognitivists to construct precise, formal models of mental function or artificial intelligent systems. This allows for the development of advanced computational theories that can be tested via computer simulations. These models are capable of efficiently recognizing patterns, classifying data, and performing complex logical operations. For many researchers, this is sufficient evidence that information processing in the cognitivist sense is the driving force behind thinking—both in artificial systems and, possibly, in the brain itself.

The Creation of New Knowledge in AI Models

Within the framework of artificial intelligence (AI), the process of "learning" and "inference" is often understood as the application of logical rules—primarily deductive reasoning—to preexisting bodies of knowledge. As a result, new configurations of information are generated that do not introduce genuinely original content but rather reveal aspects that were already implicitly present in the input data. For instance, a text analysis system may identify an idiomatic expression hidden within a large volume of text. While this may appear to us as a "discovery," strictly speaking, the information required for this discovery was already embedded in the dataset. The AI system merely processed it according to predetermined algorithms.

From an epistemological standpoint, new knowledge emerges when something is revealed that goes beyond the available premises—for example, through inductive reasoning or acts of creative understanding and interpretation. In models that rely purely on deduction, which are characteristic of certain streams of cognitive science, such creative processes are significantly limited. Deduction can uncover hidden consequences of assumptions or data, but it does not generate meanings that extend beyond what is already contained in the premises (Kase, Slocum & Zhang, 2011).

For this reason, deductive operations—while essential for automated reasoning—are often described as "informationally empty." This type of processing can be extremely useful for reorganizing knowledge or optimizing practical solutions (e.g., finding the shortest route in GPS systems), but it does not provide radically new insights. Similarly, recommendation systems for films or music operate on data already present in the system (user profiles, the preferences of others), and the algorithms merely reconfigure this information to suggest curated content. Although the result may appear novel from the user's perspective, in cognitive terms it is still derived from information already embedded in the system, rather than being truly "invented."

In this view of cognition, there is no room for genuinely new information to arise—everything is reduced to the recombination of already available data or knowledge. In other words, the cognitive process consists of a series of "content-empty" operations, since all potential A comparable example is the computational search of a large text database for keywords: while the system may "discover" certain connections between documents, it ultimately operates on information that was already present within the dataset from the beginning. No genuinely new meaning or cognitive quality emerges—what occurs is rather an automatic reorganization of preexisting elements.

As a consequence, such a model of cognition does not account for deeper processes of meaning-making or concept formation. The activities of the mind (or of an artificial intelligence system) are not interpreted as processes through which concepts emerge or are constituted, but merely as acts of symbolic manipulation. Therefore, there is no room in this framework for genuine acts of understanding, which in traditional epistemology are associated with the transition from data to meaning. In the approach centered on "purely syntactic" information processing, the issue of interpretation or meaning attribution is marginalized.

For cognitivists inspired by AI models, this is not considered a fundamental problem—the key concern is the system's capacity to execute correct computational operations, not whether it involves genuine understanding or the creation of new knowledge. As a result, the vision of cognition deliberately bypasses questions related to concepts, the reflection on meaning, or interpretation. The very notion of "novelty" is restricted to the revelation of hidden consequences of already established assumptions and rules, rather than the emergence of a radically new cognitive quality (Piccinini & Scarantino, 2011).

Thus, one could argue that within this approach, the cognitive process is devoid of a strictly creative, conceptual dimension. All "creativity" is reduced to the discovery of structures that were previously encoded within the system, without extended

reflection on what true understanding entails or how genuinely new ideas arise in the mind—or in a computational system.

Moreover, in AI models, it is difficult to speak of any form of transcendence, understood in the traditional sense as the capacity to go beyond a fixed set of symbolic configurations or beyond one's own states of consciousness. Intelligent systems operate within a single, strictly defined "reality"—that of formal language, in which symbols are transformed according to predetermined algorithms. There is no "second dimension" or alternate level of reality to which the system could refer or with which it could meaningfully engage.

An example would be an algorithm playing chess: it has at its disposal a set of rules and answers to dependent data, and its "world" is limited to virtual chessboards and the functions of the pieces. The machine is not accessible to the meaning of the king or bishop in any context of the "real" world - it does not consider the historical origins of chess, there is no risk of occurrence, and it is also impossible to define itself in a perspective "off the chessboard". Tools for operating on the appropriate symbols (moves, positions, values), and all its "strategies" are reduced to a method of calculating at the highest level the move in the current situation. This is the case with advanced language processing systems that answer questions or extract text. They seemingly do not have a single access to the real reality that has been created encoded in their parameters (statistical models) or in the information layer of symbolic knowledge. Such systems can get their answers "transcending" from the symbol system, but in fact still in the case of data processing. Their "world" is in the context of the legal complex, which do not "refer" to the real experience of the human subject – there is no introspection, no sensory experience, and no reflection on one's own state of being (Lee and Lim, 2024).

In this extension of artificial intelligence, neither classical introspection (the ability to use the hardware of mental states from an "internal" perspective) nor extraspection (understood as the experience of the world of consciousness and awareness) is available. Everything is limited to calculating monetary means or probabilities within a ready-made set of procedures. From these procedures, there is no self-knowledge or the ability to evaluate one's own actions in an introspective way – a response in the same way as people do when they reflect on the motives of their thoughts or the conditions of their behavior.

Moreover, in this symbolic reality, there are no separate "objects" or "entities" to which the system can refer in a direct and conscious way. For a chatbot, the description of a flower in a garden is a set of characters (sequences of letters) and specific semantic labels that can be applied (added, removed, replaced) according to rules. However, there is no room here for perceiving the actual beauty of the flower or experiencing the scent. From the perspective of artificial intelligence, all these features remain internally consistent, but enclosed in a one-dimensional symbolic - they do not affect the crossing of this plane, as it happens in humans, which can cause effects or reflections on them. Depending on what the whole idea of knowledge is: it becomes a "package" of symbols that can be transported from one place to then be processed in different ways, but without applying it, common meaning or connection with sensory experience. This is not knowledge "about the world" in the classical sense - it is actually a systematized set of formal content that can be summarized, ordered or verified in terms of compliance with specific rules. This type of knowledge is not related to an empirical context, is not referred to a real concept or to the concept of a subject.

In such a case, AI can of course generate extended analyses or conclusions from large data, but they still remain tightly bound to formal meaning. The lack of transcendence or introspection means that there is no transcendence of the language of symbols itself. And since this is a legitimate (computational) transfer of data, there is no room

for either a real “I” (capable of self-reference, reflection) or direct contact with empirical reality.

According to this vision, the cognitive act is reduced to the algorithmic manipulation of data – specifically symbols – which consequently makes the mind similar to a specific processor or computer executing a series of commands. The ideas of “reason” and “consciousness” accepted in cognitive science do not overlap with concepts known from modern philosophy (for example, Cartesian philosophy), where the emphasis was on the subject’s ability to doubt, self-reflect, or consciously experience the surrounding reality. There is no room for this in the cognitivist trend, because according to the logic of symbolic processing, it is enough for the system to act in accordance with the rules. This, in turn, means that “reasoning” is treated as a formal process, and “consciousness” is often reduced to an internal computational state, devoid of rich, philosophical depth. This difference in the understanding of key concepts is most clearly visible in the perspective on the essence of knowledge and the way of knowing. The cognitively oriented subject is seen primarily as a mechanism for managing symbols, which implies that knowledge is purely formal, almost “mathematical” in nature – it can be rearranged, ordered, combined, but lacks reference to any specific meaning or significance. Instead of deeper understanding, we get an efficiently functioning algorithm: input data is entered, processed, and then output data is obtained, which may give the impression of “in-depth analysis,” although in fact it is the product of a set of predetermined instructions. Furthermore, since algorithms are rarely (if ever) created by the subject itself, but rather received from the outside (e.g. by programmers in the case of AI, or by evolution in the case of the human mind), we are dealing with another problem. If the cognizing subject is not an active creator of its own rules for processing information, what does its actual independence or creativity consist in? This lack of transparency regarding the source and nature of algorithms reinforces the impression that the subject understood in this way is largely “locked” into formal patterns that it cannot transcend. As a result, the entire cognitive process in such a model is deprived of many features that have traditionally been attributed to a thinking and conscious person. There is no space for subjective experience, self-reflection, freedom to create new ideas or introduce innovative interpretations. If cognition is limited to processing symbols within imposed rules, then it becomes – in the eyes of critics – reductive and simplified. It is unable to meet the expectations set for it by the common understanding of the cognitive process: to help solve problems in a way that goes beyond simple deductions, to discover new areas of knowledge or to generate truly original ideas.

From this point of view, cognitive science presents a vision of the mind that, although based on certain intuitions of philosophical rationalism, simultaneously reformulates them in a radical way. Whatever was understood by the concepts of "rationality" and "consciousness" in traditional philosophy (for example in Descartes' concept of Cogito ergo sum, or in phenomenology, in which consciousness was always awareness of something), is transformed for the needs of the computational model. This "new" understanding cuts off from the key role of deeper interpretation, subjectivity and experiencing the world, and focuses on formal, algorithmic aspects. As a result, the cognitivist vision leads to a certain form of a "breakneck" - as some say - concept of knowledge, because it eliminates from it what is often considered the essence of knowledge: reference to the world, creative going beyond data, reflection on meaning and significance. It is not surprising, then, that philosophers and theorists of mind, trying to solve the riddles of consciousness or understanding, formulate a number of objections to reducing cognitive processes to the manipulation of symbols. They claim that it is difficult to speak here of any deep insight into the nature of thinking.

Weak vs. Strong Version of the Artificial Intelligence Hypothesis

The concept of weak AI assumes that the computer acts as a research tool that allows for the formulation and testing of hypotheses about the human mind. However, for this tool to be truly suitable for studying cognitive processes, it must to some extent resemble the object being studied, i.e. the brain/mind. Such convergence means, first of all, that the computer and the human mind operate according to a similar structural principle: in the case of current AI methods, this will be algorithmic information processing (Searle, 1980).

For example, if we want to use a computer to simulate the way a human solves logical tasks, we should assume that there is an analogy between the computational "steps" of the program and the sequence of mental states during the actual solution of the problem by a human. For the weak version of AI, it does not matter whether the computer really "understands" the task or only thoughtlessly applies the rules of the algorithm. What is important is that certain computational mechanisms can faithfully (or at least sufficiently accurately) reproduce human cognitive functions. If this were not the case, a computer could not be used to test hypotheses about human thinking — because its operation would be at odds with what the mind actually does. This relationship between the research tool and the object being studied is clearly visible in practical applications. Let us take examples from the field of cognitive psychology (1) Modeling short-term memory — in order to study how people store and manipulate information in short-term memory, computer programs are created that perform an analogous set of operations (e.g. storing elements in a "buffer", transferring them to another "register" after a specified time). It is believed that if computer simulations accurately predict and explain the behavior of the people being studied, then the algorithm may to some extent reflect the mechanism of human memory. (2) Simulations of learning processes — learning systems (machine learning) develop their "skills" based on input data and an appropriate error function. Cognitive psychologists often check to what extent this process resembles the learning of children or adults in specific tasks (e.g. pattern recognition). If the results achieved by machines and humans overlap, we conclude that, at least in some sense, statistical algorithms can imitate human learning patterns.

From the point of view of weak AI, both of these examples assume that the mind and the computer must share some common rules of operation, at least in the formal dimension: both the human and the program "process" information step by step, based on certain rules. If the brain did not use similar, algorithmic mechanisms to any extent, we could not use the computer sensibly to confirm hypotheses about human thinking. In other words, the tool (computer) must be in essential respects commensurate with the phenomenon (mind) that we want to study.

The conclusion that follows from this argument is quite far-reaching: (1) if we assume that the computer is indeed able to correctly simulate cognitive processes, then we also agree with the thesis that the human mind implements similar (or analogous) rules of information processing. (2) In particular, that the functioning of the human mind can be reduced (at least at some level of description) to algorithms that can be defined and executed in the computer. From a philosophical perspective, this means that weak AI introduces us to a mechanistic picture of cognition. Although it does not go as far as strong AI, which speaks of the possibility of genuine consciousness in a machine, it nevertheless demands the acceptance of a certain "algorithmic" dimension of the human brain's operation. For many critics, such a conclusion is already controversial enough, because it reduces the rich reality of the mind to formal computational schemes.

On the other hand, proponents of weak AI defend themselves by pointing out that it is not about fully explaining the entire phenomenon of consciousness or internal experience, but about the possibility of scientific, empirical study of certain aspects of

human thinking. Computers can verify to what extent various hypotheses — regarding the way we make calculations in our heads, use memory, or learn new information — are consistent with observed experimental results. If such verification is positive, it is an additional argument in favor of the algorithmic interpretation of specific functions of the mind. In order for a computer to serve as a useful tool for studying cognitive processes, there must be a similarity (at least functional) between its operation and the operation of the brain. If it turned out that the human mind works in a fundamentally non-algorithmic way, then computer simulations could not be reliably used in psychological or cognitive research. Weak AI therefore emphasizes that this similarity is a practical requirement: without it, it would be impossible to translate simulation results into an understanding of human cognitive processes. Thus, by adopting the assumptions of weak AI, we also assume that at least some aspects of the human mind are subject to algorithmic mechanisms that can be mapped in computer systems. John R. Searle maintains that computers can be used—understood as a combination of hardware and software—in research on human cognitive processes without assuming that computers actually perform all the operations that constitute the essence of human thought. According to Searle, a computer program can simulate the functions of the mind in certain areas, but it does not have to reproduce them in full or implement them in an identical way to the human brain. Thus, he rejects the logical consequence that the mere use of a computer to study the mind would prove that the basis of the mind's operation is strictly algorithmic (Mazurek, 2025, p. 105).

The Problem of Functional Similarity

However, the second part of Searle's argument raises doubts: how can hypotheses about human cognition be tested using a tool that supposedly shows no similarity at all (either in a partial sense or at the level of mechanism) to the object being studied? Since a computer is used to simulate certain mental processes, we must assume at least partial similarity to the functions of the mind. In practice, when we model short-term memory or learning processes, we count on the results of the program's operation to some extent reproducing the behaviors observed in humans. Without such functional similarity, it would be difficult to compare the results of computer simulations with data from psychological studies.

Let's assume that we create a computer program to solve simple logical tasks such as Sudoku or deductive puzzles. If the program and the human solve these tasks in a similar time and make similar types of errors (e.g., they are more likely to make mistakes when there are a large number of simultaneous dependencies), then we can conclude that certain elements of human reasoning can be interpreted in algorithmic categories. The weak version of AI says that although the computer does not have to "understand" Sudoku in the sense of human consciousness and motivation, similar operations occur in terms of structure (e.g. checking subsequent variants). However, it is difficult to maintain that there is no functional similarity between them — after all, we compare results and methods of operation precisely in order to say something about the mind.

Searle suggests that the computer can be solely a research instrument, without necessarily concluding that the mind works the same way as the machine. Yes, it is possible that the research focuses only on those areas of cognition in which the computer can perform a reasonable simulation (e.g. logical reasoning, processing of visual signals). Then there is still room for other areas (e.g. sensory experiences, emotions) in which the computer is unable to provide an analogous simulation (Searle, 1980).

Nevertheless, even in such "fragmentary" applications, the computer must function similarly to the mind in certain specific aspects (e.g. algorithmic processing of information) — otherwise the experiment and the conclusions drawn from it would be methodologically unjustified. For this reason, Searle's critics point out that weak AI is

not entirely “neutral”: it postulates at least some common algorithmic denominator between the operation of the machine and the operation of the mind.
Searle’s reservations stem from his reductive attitude toward the reductive approach that the human mind is merely a set of algorithmic operations. Nevertheless, weak AI does introduce the assumption that at least some key cognitive functions (e.g. reasoning, problem-solving, pattern recognition) can be described in terms of formal rules. As a result, it turns out that even if we only accept the role of the computer as a research tool, we still admit the existence of some structural similarity between the operation of the computer system and the operation of the mind.
Some critics point out that weak AI involves thinking in a kind of “vicious circle”: (1) we use a computer to study the human mind. (2) For this to be possible, the computer must be adapted to the subject of the study, i.e. act in a similar way (at least in part). (3) Based on the simulation results, we infer the algorithmic nature of the mind.
As a result, we assumed at the beginning (point 2) that the mind must be capable of being simulated algorithmically, and then (point 3) we prove exactly what we assumed. If research were to prove that the human mind actually processes information algorithmically, then methodology should not a priori accept such a statement as a condition for computer modeling. In other words, research results may be rather pragmatic in nature (they show the usefulness of models to a certain extent), and not decisively decisive about the nature of human cognitive processes.

Is a computer a subject of cognition in the weak version?
An additional issue is the question of whether a computer can be considered a "subject of cognition" at all. When we talk about a subject, we usually mean someone (or something) capable of experiencing, understanding, and making decisions from their own perspective. In weak AI, the computer is still more of a tool: the human researcher designs tasks, interprets results, and draws conclusions about human thinking. The machine only plays the role of an "executor" of the algorithm, without thereby becoming an independent, conscious "I".
Meanwhile, in the approach that we could call semi-official in cognitive science, it is sometimes said that "the computer takes over the function of the subject" when it simulates the process of thinking or decision-making. Two parallel interpretations therefore appear:
(1) The computer as a replica of the subject - when the emphasis is on simulating the mind and demonstrating similarity in cognitive operations (e.g. in language models or artificial neural networks).
(2) Computer as a tool – when it is emphasized that humans are the only real cognitive subject, and the computer only enables a better understanding of specific mechanisms, without having its own consciousness or intentions.
Both of these perspectives coexist in the weak version of AI, which creates a certain tension. On the one hand, for the simulation to be valuable, the computer must “behave” in some aspects like a subject; on the other hand, it is constantly denied that it actually is one in the full sense.
On the one hand, it is said that a chatbot (e.g. a program that learns to generate answers) takes over the function of an interlocutor and in this sense simulates human linguistic abilities. On the other hand, authors usually maintain that a chatbot is not a “real” speaker, because it has no consciousness, cultural context or intentions. As a result, it is sometimes treated as an “impoverished speaker” and sometimes only as a “tool” for modeling dialogue.
Algorithms that suggest movies or songs operate on the basis of certain simple assumptions about the user’s preferences, which is supposed to imitate “understanding” taste. The creators of these systems often emphasize that they are only tools that imitate certain decision-making processes of the user (e.g., associating

similar pieces of music), and at the same time admit that these mechanisms are partially analogous to human patterns. In the strong version of artificial intelligence (strong AI), it is assumed that a machine appropriately equipped with the right program not only simulates cognitive processes, but itself "possesses" a mind. In other words, the computer is no longer just a useful tool in research on human thinking, but becomes a "full-fledged subject of cognition" - capable of understanding, reasoning and having all the mental states that we typically attribute to humans. In this sense, any appropriately complex structure that implements the right program (algorithm) would be called a cognitive system identical to the human mind.

John R. Searle (Searle, 1980), explaining the strong version of the hypothesis, points out that in this approach the brain is treated as hardware and the mind as software. Therefore, if it is possible to construct a computer with a sufficiently complex architecture and equip it with appropriately developed software, it will be able to process information in an identical way to the human mind. A practical example of such an idea is the dream of creating the so-called AGI, a system that could perform all cognitive tasks in the same way as a human (or even better), including artistic creation or understanding humor.

It is important here to assume that there are no significant limits to the capabilities of a computer - the only challenge is the degree of complexity of its structure and software. In the future, strong AI supporters emphasize, computers may become so complex as to reproduce every cognitive activity of the human mind: from simple perceptual activities, through advanced forms of reasoning, to fully conscious experience of emotions.

In practice, however, there are numerous examples that constitute a flashpoint in discussions about strong AI. It is enough to mention the problem of “reflective thinking” or experiencing subjective states (qualia) (Chalmers, 1995, 1996): supporters of strong AI will claim that advanced programs based on complex calculations and hierarchical architecture of neural networks could produce equivalents of these states and processes. Their opponents, on the other hand, point out that we still do not have convincing evidence that purely syntactic operations on symbols can produce real “feeling” and “consciousness” in the sense in which human beings experience them. However, from the perspective of strong AI, the limitations of current technologies do not invalidate the principle that all cognitive processes, including feeling and understanding, can be fully replicated in a computer. This stems from the belief that both the human brain and the computer are systems that process information according to finite (although extremely complex) rules – and therefore, if thinking and consciousness occur in one of these systems (the brain), then it should also be possible in the other (the computer), provided that we meet the hardware and programming requirements. Searle’s views on artificial intelligence can be expanded by referring to more basic categories, such as the relationship between the subject and the object of knowledge and the division into two levels of knowledge – syntactic and semantic. In this broader, almost universal approach, the essence of computer programs becomes the manipulation of symbols in language understood solely in terms of its syntax, while the layer of meaning is not present at all. This means that what is usually called “language” in relation to code or computer instructions is reduced to sequences of signs, rules for combining them, and methods of transformation. An example is programs that translate from one natural language to another, which often rely on the statistical analysis of the occurrences of words and phrases in large corpora of text. Although they ostensibly “know” the meaning of words, in practice they perform only syntactic operations: they match correspondences and purely formal dependencies (e.g. frequency of collocations, grammatical patterns) on the basis of training data. Such a translator does not “understand” what a given word really means in a cultural, emotional, or social

context – all it needs is the computational ability to properly combine and sort symbols.

From this perspective, the question of "understanding" becomes inadequate, because it refers to a semantic level that a computer program does not have. Just as in the "Chinese room" thought experiment, where the person manipulating Chinese characters does not understand their meaning at all, computer algorithms remain at the level of syntax processing. They do not use content in the human sense – they do not refer to an object or idea in the physical or mental world, but operate exclusively within sequences of symbols. Extending Searle's interpretation to this broader perspective, we can say that there is no layer of meaning in "computer knowledge": it is knowledge reduced to the rules of compilation, statistical analysis, or formal conversion of one symbol into another. If a computer working with such knowledge were to be considered a "subject of knowledge," it would have to be shown that it enters into a relationship with some sphere of meaning (the external world or the world of one's own experience), which, however, contradicts the fact that all of its operations are limited to fulfilling algorithms at the syntactic level.

To further illustrate the lack of semantics in computer processing, we can cite the example of data sorting systems: software can group information into folders and hierarchies extremely effectively (e.g. by placing different types of files in appropriate directories), but it does not create any "understanding" of the content. Documents, photos, and spreadsheets in the eyes of a computer are just collections of bits and specific attributes; their actual content or meaning is the domain of a human who is able to consciously refer to what a given file represents and in what context to use it.

It follows that in a purely syntactic approach – appropriate for machines – there is no room at all for the problem of understanding in the traditional, philosophical sense. If we tried to forcefully "inject" a layer of meaning into a computer, we would additionally have to explain how the machine relates to real objects or ideas, and not just to symbols and instructions. However, as long as the actions of computer intelligence remain within the framework of syntax, the question of intentionality, meaning or understanding does not arise. We can therefore conclude that Searle, with his division into syntactic and semantic aspects, touches on the fundamental problem of the "semantic void" of all computations, and at the same time highlights the essential difference between an authentically cognizant subject and a machine operating on signs. In the concepts of strong artificial intelligence, which assume the possibility of fully explaining and reproducing cognitive processes solely through symbol processing, there is a radical break with the belief accepted in the philosophical tradition that knowledge must be conceptual and semantically embedded. From the times of ancient thinkers, through modern epistemology, to contemporary theories of cognition, there is a consensus among philosophers that the key feature of human cognition is the use of concepts, or meanings, which allow us to understand the world and make it an object of reflection. This "conceptual" nature of cognition is expressed in the constant connection of perception, thought, and language into a coherent whole, in which words are not just empty signs, but refer to content and allow it to be interpreted.

Against this background, the idea of artificial intelligence as a system operating solely at the level of syntax – analyzing sequences of symbols without reference to meaning – raises doubts as to whether we are really dealing here with "knowledge" in the philosophical sense. If a computer processes data without any internal semantic representations, then everything it generates can be treated as sets of formally correct strings of characters, without any real meaning. Translation systems are a good example: they can translate text very efficiently, but they are still limited to matching patterns and syntactic rules and statistical correlations between words in different languages. As a result, the translation may be grammatically correct and useful in

everyday communication, but from an epistemological point of view it lacks confirmation that the machine actually “understands” the content it is operating on. In other words, it is futile to look for an understanding of the meaning of concepts in it: the computer works with strings of symbols, remaining at the level of their formal properties. This reduction of language to syntax alone contradicts many classical theories of cognition, in which understanding is an inherent attribute of all knowledge. Philosophers since Aristotle have emphasized that in order to “know” and “understand” something, one must operate in a space of meanings, references, and mental categories. Therefore, when the concept of strong AI claims that syntactic manipulation of symbols is sufficient for cognition, it exposes itself to the accusation of eradicating conceptuality from mental processes. Some researchers argue that even if such syntactic intelligence is able to effectively perform tasks that humans perceive as evidence of “intelligence” (e.g., recognizing images or playing complex games), we still have no basis for considering it to be “cognition” in the full sense. There is no evidence that these systems ever cross the threshold of syntax alone and enter into a relationship with the semantic layer, where concepts acquire meaning and contextual reference. Creators and advocates of strong AI usually respond that in the future it will be possible to develop algorithms and architectures (for example, using neural networks and advanced robotics) that will be able to break down the barriers between syntax and semantics. At the same time, however, as critics rightly note, AI progress to date has focused primarily on increasingly efficient manipulation of symbols and patterns within data – not on creating a genuine “map of meaning” or giving meaning to the world, as human consciousness does. The question, therefore, of whether a computer operating solely at the syntactic level can achieve a truly conceptual state remains open, and cognitive scientists do not provide a clear answer as to whether this is even necessary to talk about “knowledge” in terms of AI. For now, the overwhelming impression remains that the concept of a purely syntactic AI contradicts the philosophical dogma that thinking requires more than the formal transformation of symbols – it requires the ability to derive meanings from them, interpret them, and arrange them into internally coherent concepts. It is difficult to say whether a technology or theory will emerge in the future that is capable of blurring this fundamental difference. At this stage, however, it seems that strong AI, understood as a complete rejection of the conceptuality of cognition and its reduction to symbolic syntax, must face the accusation of inconsistency with a centuries-old philosophical tradition – a tradition for which the semantic and conceptual aspect of knowledge constitutes the core of what we call human cognition.

In the models of cognition characteristic of artificial intelligence, there is no classical division between subject and object: the computer does not "go out" to anything in the external world, and even more so does not "transcend" its states in an act of transcendence. The operations performed by the machine are primarily transformations of symbols within a formally defined language - there is no reference relation in which expressions would refer to something beyond the symbols themselves. If we tried to ask what exactly such knowledge "concerns", it would turn out to be an irrelevant question: in its "cognitive" activity, the computer does not refer to any external reality, but only manipulates signs based on given rules. In this sense, the idea of "cognition" in artificial intelligence brings to mind certain philosophical concepts, such as Wittgenstein's language games or Richard Rorty's neopragmatism. The similarity lies in the fact that all mental (or linguistic) operations take place in a certain closed system of signs and rules for their use, without the ambition to go beyond this system. For example, in Wittgenstein’s language games, players do not necessarily “describe” some external reality, but communicate within established conventions. A computer program playing chess can be viewed similarly: its “world” is a 64-square chessboard and a set of allowed moves defined syntactically. The program does not relate these moves to anything outside the game itself. However, in

the case of artificial intelligence, there is a fundamental difference: although Wittgenstein's or Rorty's concepts emphasize that language remains locked within social conventions and does not necessarily have to reflect objective reality, there is still a participant in the language game – a specific person or community using expressions and interpreting their meanings. It is the human who decides how to understand given words, when to apply specific rules, and how to interpret the statements of others. Meanwhile, in the view of artificial intelligence, the "language" of the program is merely a set of symbols (e.g. binary strings or text characters), and their processing is strictly formal procedures. For example, a text-generating program (like a language model) can combine words into seemingly meaningful sentences, but it only operates on statistical dependencies within a data set. There is no conscious participant who assigns meaning to expressions, considers the social context, or interprets them at the level of semantics. This lack of distinction between subject and object, as well as the lack of reference to the realm of meaning, means that in the concept of artificial intelligence, the computer itself cannot be treated as a classic cognitive subject who experiences objects, makes judgments about them, or establishes relationships with them. Instead, we are dealing with a system of symbols and rules that "processes" them to achieve given computational goals, such as the optimal move in a game or translating text from one language to another. Imagine, for example, an advanced voice assistant in a smartphone that recognizes voice commands and answers questions. Although it may seem to "understand" and "relate" to the world, in fact it only analyzes the sound wave and extracts words from it, and then generates an answer based on its model. The whole thing takes place within certain syntactic and statistical patterns, without the need to assign meanings to words or refer to actual objects.

Contrary to the concept of language in neopragmatic philosophy or in Wittgenstein, where there is always a communicative community or an individual user of language, in the concept of artificial intelligence – especially one based on symbolic procedures – the entire "game" takes place in the domain of formal rules without players in the classical sense. The computer system does not communicate with other entities, does not share impressions, does not reflect on what the symbols represent. It simply processes them, just as a calculator processes digits in addition and multiplication. Therefore, there is no concept of interpretation or negotiation of meanings in the way described by philosophical trends that place the human as a conscious subject at the center.

Conclusion

The approach to artificial intelligence, especially in its strong version, proposes a radical reinterpretation of the concept of cognition and knowledge, reducing it essentially to the syntactic processing of symbols. In such a model, the traditional subject-object opposition disappears: "cognition" no longer assumes reference to anything beyond the system of symbols itself, and the relation of reference, important in many classical epistemological approaches, ceases to play any role. This scheme leads to the conclusion that the act of "understanding" is understood exclusively as the correct performance of data transformations according to predefined rules.

Such a treatment of knowledge as a purely syntactic entity contradicts the common belief in philosophy about the indispensability of the semantic layer in cognitive processes. Both in theories based on conceptuality and meaning, as well as in many approaches from the borderland of linguistics and philosophy of language, knowledge is inextricably linked to the understanding of meanings and references, which requires the existence of an interpreting subject. In turn, in the vision proposed by strong AI, there is no subject "going out" to the object, and there is also a lack of meaning in terms of the classical sign-meaning-world relationship.

Despite the radical reduction of semantics to symbol manipulation, artificial intelligence provides fruitful research tools that allow simulation and analysis of various aspects of mental processes. They also create practical solutions (for example in the field of machine learning or natural language processing), the effectiveness of which in selected tasks is undeniable. Nevertheless, the question of whether machines actually "know" remains open and is a source of philosophical disputes.

The conflict between the syntactic approach and the traditionally understood, semantic understanding of knowledge is particularly revealed in the debate around the strong version of AI. On the one hand, proponents of strong AI maintain that thanks to sufficient computational complexity, a machine can fully gain the properties of the mind, including understanding and consciousness. On the other hand, critics emphasize the lack of intentionality and the "semantic void" in machine operations, which, in their opinion, do not allow for equating human cognition with computer data processing. This discussion, taking place on the border of philosophy, cognitive science, and computer science, highlights how complex and multifaceted the issue of the nature of mind and cognition is. It turns out that transferring these categories to the sphere of artificial information systems is associated not so much with a simple extrapolation of reasoning from humans to machines, but with the need to thoroughly rethink concepts such as understanding, intentionality, meaning, and the subject-object relationship. As a result, the dispute over strong and weak AI continues to open up new areas of reflection on what constitutes the essence of cognition and whether it can be reduced to algorithmic symbol processing at all.

Declaration on AI use

AI-assisted tools were used during manuscript preparation for language editing and refinement. The author takes full responsibility for the content, interpretation, and integrity of the research.

Bibliographic references

Card, S. K., Moran, T. P., & Newell, A. (1983). The psychology of human-computer interaction. Lawrence Erlbaum Associates.

Chalmers, D. J. (1995). Facing up to the problem of consciousness. Journal of Consciousness Studies, 2(3), 200-219.

Chalmers, D. J. (1996). The conscious mind: In search of a fundamental theory. Oxford University Press.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. Cognitive Science, 7(2), 155-170. https://doi.org/10.1207/s15516709cog0702_3

Kase, K., Slocum, A., & Zhang, Y. Y. (2011). Literature review: Epistemological perspective - inductive and deductive thinking. In Asian versus western management thinking (pp. 39-72). New York and London: Palgrave.

Laird, J. E., Lebiere, C., & Rosenbloom, P. S. (2017). A standard model of the mind: Toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics. AI Magazine, 38(4), 13-26. https://doi.org/10.1609/aimag.v38i4.2744

Lee, B. W., & Lim, J. (2024). Language models don't learn the physical manifestation of language. arXiv preprint arXiv:2402.11349. https://doi.org/10.48550/arXiv.2402.11349

Mazurek, M. (2025). Limitations of artificial intelligence: Why artificial intelligence cannot replace the human mind. Filozofia i Nauka. Studia Filozoficzne i Interdyscyplinarne, 13, 97-111. https://doi.org/10.37240/FiN.2025.13.1.6

Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. Communications of the ACM, 19(3), 113-126. https://doi.org/10.1145/360018.360022

Nilsson, N. J. (2007). The physical symbol system hypothesis: Status and prospects. In B. Neumann & M. M. Richter (Eds.), 50 years of artificial intelligence (pp. 9-17). Springer.
Piccinini, G., & Scarantino, A. (2011). Information processing, computation, and cognition. Journal of Biological Physics, 37(1), 1-38. https://doi.org/10.1007/s10867-010-9195-3
Searle, J. R. (1980). Minds, brains, and programs. Behavioral and Brain Sciences, 3(3), 417-424. https://doi.org/10.1017/S0140525X00005756
Thagard, P. (2005). Mind: Introduction to cognitive science (2nd ed.). MIT Press.

Words: 9473
Characters: 62 163 (35 standard page)

Mariusz Mazurek, PhD
Institute of Philosophy and Sociology of the Polish Academy of Sciences
Nowy Świat 72, 00-330 Warsaw
Poland
mariusz.mazurek@ifispan.edu.pl
ORCID: 0000-0002-9646-6897