İSTANBUL UNIVERSITY PRESS

Felsefe Arkivi - Archives of Philosophy
**ARCP** 2024, 60 : 15–23
DOI: 10.26650/arcp.1417812

Research Article / Araştırma Makalesi

# On the Relationship Between General Artificial Intelligence and Consciousness*

## Genel Yapay Zeka ve Bilinç İlişkisi Üzerine

Ferhat Onur[1]

[1]Karadeniz Teknik Üniversitesi, Edebiyat Fakültesi Sosyoloji Bölümü, Trabzon, Türkiye

**Corresponding author / Sorumlu yazar :** Ferhat Onur

**E-mail / E-posta :** ferhatonur@ktu.edu.tr

*This article is a revised and expanded version of the paper written in Turkish titled "Gerçek Yapay Zekanın Önündeki Engel Bilinç midir? / Is Consciousness an Obstacle to Real Artificial Intelligence", which I presented at the 1st International Congress on Artificial Intelligence and Data Science in 2021, İzmir/Türkiye.

**ABSTRACT**

Artificial intelligence research continues at full speed with countless applications where its effects can be directly observed. However, the ultimate goal of these efforts is not just to produce technologies that make our lives easier but also to achieve the kind of intelligence that humans possess—namely, artificial general intelligence and beyond. We want AI to not only be a chess master, drive a car, or serve as an assistant but also to engage in meaningful communication, come up with creative ideas, have common sense, and think about the world. When discussing the possibility of such intelligence., philosophical problems arise. Prominent among these is the epistemological problem of how we can know whether AI has real intelligence like humans. At this point, the behaviorist approach, which seeks signs of intelligence in exhibited behaviours, may be justified; however, there is a catch. AI must convince us that it indeed possesses general intelligence. The fact that consciousness could play a role in this attempt at persuasion would imply that AI and consciousness are more strongly interconnected than previously thought. In the goal of becoming more autonomous beings/individuals, the path of AI may intersect with the existence of consciousness.

**Keywords:** Consciousness, general artificial intelligence, Turing test, behaviourism, abductive reasoning

**ÖZ**

Yapay zekâ araştırmaları doğrudan etkilerini gördüğümüz sayısız uygulamayla birlikte tüm hızıyla sürüyor. Ancak bu araştırmaların nihai amacı hayatımızı kolaylaştıran teknolojiler üretmekten ziyade insanın sahip olduğu türden bir zekâyı, genel yapay zekâyı, hatta daha fazlasını elde etmektir. Yapay zekânın sadece bir satranç ustası olmasını, araç kullanmasını veya asistanlık yapmasını değil, anlamlı iletişim kurabilmesini, yaratıcı fikirler ortaya koyabilmesini, sağduyu sahibi olabilmesini, dünya hakkında düşünebilmesini istiyoruz. Böyle bir zekânın olabilirliği üzerine konuşulduğunda felsefi problemler baş gösteriyor. Bunlardan öne çıkanı yapay zekânın insanlar gibi gerçek bir zekâya sahip olup olmadığını nasıl bilebileceğimiz ile ilgili epistemolojik problemdir. Bu noktada zekânın belirtisini sergilenen davranışlarda arayan davranışçı yaklaşım haklı görülebilir ancak bunun bir koşulu vardır: Yapay zekâ bizi kendisinin gerçekten de genel bir zekâya sahip olduğuna ikna etmelidir. Bilincin bu ikna girişiminde bir rol oynuyor olabilmesi yapay zekâ ve bilincin sanılandan daha güçlü bir şekilde birbirine bağlı olduğunu imleyecektir. Daha otonom bir varlık/birey olma hedefinde yapay zekânın yolu bilincin varlığı ile kesişiyor olabilir.

**Anahtar Kelimeler:** Bilinç, genel yapay zeka, Turing testi, davranışçılık, savavarımlı çıkarım

| | |
|---|---|
| **Submitted/Başvuru** | : 10.01.2024 |
| **Revision Requested/ Revizyon Talebi** | : 22.01.2024 |
| **Last Revision Received/ Son Revizyon** | : 22.05.2024 |
| **Accepted/Kabul** | : 14.06.2024 |

## Introduction

The stages of development of artificial intelligence (AI) are generally characterised by three distinct periods[1]. In the first period, which dominated AI research from the 1950s to the 1990s, programmers translated their knowledge of specific fields into algorithms or code to solve problems within those fields. The resulting programmes were rule-based and focused on representing human knowledge through these rules, which is why such systems are said to produce “handcrafted knowledge”. Examples of such systems include widely used software like chess-playing programmes, smartphone applications and logistics programmes for scheduling.

In the second wave of AI, dating from the 2010s to the present, rule-based, symbolic representation was replaced by statistical learning. The distinguishing feature of this second period, during which a type of machine learning called “deep learning/thinking” is carried out, is that it makes it possible for machines to program themselves. Programmers create statistical models for specific problem domains and expose them to extensive datasets. By utilising these models, machines can interpret data, expand their knowledge, and make predictions and decisions. Current AI technologies exemplify the second wave with applications such as virtual assistants, chatbots, driverless vehicles, customized experiences, facial recognition, coloring, translation, and computer-aided disease diagnosis.

The third wave, which is still in its early stages, pushes us to see AI systems as more than just tools that apply rules programmed by humans (first wave) or make generalizations from human-arranged datasets (second wave), and invites us to think of them as partners who have acquired human-like communication and reasoning skills[2]. At this stage, AI is expected to not only make decisions in response to the events and situations it encounters, but also to explain those decisions and reveal the underlying reasons in a "contextually adaptable" manner. The contextual adaptability of third-wave systems means that they can also communicate naturally. An important step towards this natural communication can be seen in recently developed systems such as ChatGPT. However, some believe that the intended result in AI will not be achieved even if the third period or wave is successfully realised.

The ultimate goal of AI research is to achieve *artificial general intelligence* (AGI) or *strong artificial intelligence*[3]. Presently, the predominant focus of research in developing thinking machines is on *weak* or *narrow artificial intelligence*, which involves designing programmes tailored to solve specific problems. In contrast, AGI will be capable of doing all the intellectual work humans can do and possibly even surpass them[4].

A person’s intellectual activity in everyday life is based on an extensive knowledge of the world. While some of this knowledge is rational and calculated, most of it is rooted in common sense and intuition. Because common sense and intuitive knowledge cannot be expressed formally, one of the major challenges for AI is to encode this informal knowledge in computers[5]. This problem was first addressed by Alan Turing (1912-1954), who is widely regarded as one of the founding fathers of AI. Turing discussed it as the eighth argument against AI in his renowned article entitled “Computing Machinery and Intelligence”. In this argument, known as "The Argument from Informality of Behavior", Turing stated, “it is not possible to produce a set of rules purporting to describe what a man should do in every conceivable set of circumstances.”[6]

We live with numerous moral, social, and legal rules governing our behavior, yet we frequently encounter situations where these rules prove ineffective or inapplicable for various reasons. For instance, while we impose various penalties for murder in civilian life, such penalties may not hold during times of war. Even in the context of war, precise rules governing the permissibility of taking a life can be elusive. Our personalities, psychology, the social conditioning of the group we belong to, and many other factors can all impact our decisions under such circumstances. In essence, we are beings capable of adapting our behaviours based on the context in which we find ourselves.

The question that arises then is whether there is an artificial way to achieve this flexibility and adaptability exhibited by human intelligence. Turing acknowledged that it seems impossible to provide rules of behavior that would cover every possibility, but he added that this does not mean endorsing the idea that humans are not machines. He suspected

[1] This classification was made by the Defense Advanced Research Projects Agency (DARPA), a government agency affiliated with the US Department of Defense, which has made many innovations in the field of technology and has received the approval of most researchers as it is found to be heuristically useful. DARPA’s presentation on the subject by John Launchbury can be accessed at https://www.darpa.mil/about-us/darpa-perspective-on-ai.

[2] Kristian Kersting, “Rethinking Computer Science Through AI”, KI – *Künstliche Intelligenz* (2020) 34, 435.

[3] While some authors differentiate between artificial general intelligence and strong artificial intelligence, what is meant by these two terms here is "a human-like cognitive capacity" in its broadest sense, so the distinction in question is not relevant to our discussion. It can also be argued that the ultimate goal of AI research is super-AI, which is defined as “an intellect that is much smarter than the best human brains in practically every field, including scientific creativity, general wisdom, and social skills” (Nick Bostrom, “How Long Before Superintelligence?”, *Linguistic and Philosophical Investigations* (2006)). However, since the existence of such hypothetical intelligence does not significantly impact the argument presented in this paper, it is reasonable to omit it from our consideration.

[4] Toby Walsh, *Machines That Think: The Future of Artificial Intelligence* (New York: Prometheus Books, 2018). Nicolas Sabouret, *Understanding Artificial Intelligence* (Boca Raton: CRC Press, 2021).

[5] Ian Goodfellow, Aaron Courville & Yoshua Bengio, *Deep Learning* (Cambridge: MIT Press, 2016), 2.

[6] Alan M. Turing, “Computing Machinery and Intelligence”, *Mind* (1950) 49, 452.

that humans were indeed machines. If human beings were machines, it might one day be possible for machines to think: "I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted."[7]

For AGI programmes, just as in Turing, the acknowledgment that machines can think is only a matter of time. Turing's prophecy was not proven. We are not yet talking without any hesitation about machines thinking. So what went wrong? Is this still a matter of time, or is there another obstacle? I argue that if we adopt a behaviourist standpoint (we infer consciousness from behaviour that we label as conscious), which Turing seems to endorse, consciousness may pose a problem for the idea that machines can possess general intelligence. But first let's delve a bit deeper into the connection between behaviorism and machine intelligence.

## 1. Can Machines Think?

What conditions or criteria are required for us to say that a machine is thinking? Turing thought that he had established a criterion with the "imitation game" he put forward in the article we mentioned. According to this game, also known as the "Turing test", if a machine can provide answers to questions asked under certain conditions (such as interrogator stay in a room apart from the other two speakers), in a manner indistinguishable from a human being, we can conclude that the machine exhibits thinking.

Thus, Turing's criterion is behaviourist, focusing solely on the responses (answers) to stimuli (questions) rather than the material structure (biological or synthetic) or states (mental or physical) of the entities involved. The reason Turing and other behaviorists hold this view is simple: There is no other way to attribute thoughts and other mental properties to any entity. The determination of whether an entity thinks ultimately hinges on the observation of appropriate behavior.

Imagine we land on a planet where we discover that there is life, and we are greeted by some life forms of that planet that are unlike us and are trying to communicate with us. How can we ensure the effectiveness of our communication and their understanding of our messages? Typically, we gauge this by assessing the adequacy of their responses to our questions and the dialogues we engage in. In fact, we also come to understand that other people have thoughts and minds by observing their verbal or physical behavior. So much so that if an event akin to the one depicted in the 1988 science fiction movie "They Live" were to occur and extraterrestrial beings resembling us in appearance and behaviour were living among us, distinguishing them from humans solely based on their appearance and behaviour would be impossible.

So why should we not say that a machine that passes the Turing test actually understands, thinks, and communicates with us? As with any other philosophical problem, objections and arguments have been raised, stating that the issue cannot be solved easily. The first serious objection was voiced by Claude Shannon (1916-2001), the father of information theory, and computer scientist John McCarthy (1927-2011), who coined the term "artificial intelligence":

> A disadvantage of the Turing's definition of thinking is that it is possible, in principle, to design a machine with a complete set of arbitrarily chosen responses to all possible input stimuli. Such a machine, in a sense, for any given input situation (including past history) merely looks up in a "dictionary" for the appropriate response. With a suitable dictionary such a machine would surely satisfy Turing's definition but does not reflect our usual intuitive concept of thinking. This suggests that a more fundamental definition must involve something relating to the manner in which the machine arrives at its responses – something which corresponds to differentiating between a person who solves a problem by thinking it out and one who has previously memorized the answer[8].

Various mental activities can be considered forms of thinking. Reasoning, understanding, judging, planning, problem-solving, remembering, imagining are all appear to be part of our thought processes. If the difference between human and machine cognitive processes corresponds to the difference between solving a problem and remembering an answer, as Shannon and McCarthy noted, then thinking must involve the active participation of the subject rather than merely accessing pre-existing knowledge as machines do. In other words, thought is a mental phenomenon that the subject consciously engages in, which is what the machine supposedly lacks.

If we consider remembering as the brain retrieving information about past experiences, it would not be categorized as a form of thinking according to this perspective. Thus, just as the machine responds to stimuli by scanning its database as if looking in a dictionary, the brain searches its memory and retrieves the desired memories into consciousness. But how do we know that the brain does not operate in the same way as other thinking processes? Shannon and McCarthy's words imply that the subject participates in the routine functioning of the brain in thought production and that it is this participation that makes thought non-algorithmic or non-computational. However,, they do not explain what the subject's contribution might be.

[7] Turing, "Computing Machinery and Intelligence", 442.

[8] Claude E. Shannon & John McCarthy, "Preface", in *Automata Studies*, Ed. Claude E. Shannon & John McCarthy (Princeton: Princeton University Press, 1956), vi.

So, perhaps there is no contribution from the subject, and what is happening is that the brain automatically conducts its thought processes. The conclusion to be drawn from this line of reasoning is as follows: "The obvious inability of present-day computer science to account for such events is no reason at all for doubting that they can be accounted for by computer science."[9] Not everyone agrees.

In parallel with Shannon and McCarthy, philosopher John Searle attempts to demonstrate that the concept of thought cannot be ascribed to machines through his thought experiment known as the "Chinese Room"[10]. According to Searle, what computers (smart devices) do can be compared to a person who does not speak Chinese but still answers questions posed to him/her in a closed room by looking at a rule book. Even though the person may have no knowledge or understanding of Chinese, he/she can respond to questions and engage in a conversation following specific instructions. Similarly, a computer executes the user's commands as long as it adheres to the instructions in its programmes, but it lacks an understanding of the meaning behind those commands. In Searle's words, computers only manipulate symbols and cannot generate meaning in the same way as the human mind[11].

It is worth noting that Searle does not claim that machines cannot think; rather, he argues that thoughts cannot be generated solely by programmes or that the programmes alone are insufficient to produce thoughts and other mental phenomena. This means that we cannot attribute thinking not only to old-fashioned first-wave AI but also to second- and third-wave AI technologies, as long as they rely on programmes to attain human-like cognitive abilities. Searle believes that the brain produces thoughts by causing them to occur. However, computational processes executed by programmes don't have causal powers because they are abstract mathematical processes.

But how do we know that the brain causes thoughts? Perhaps when the brain is in a certain state some thought occurs without being caused. In other words, it is possible that neuronal activity does not cause thoughts but *constitutes* them. If this is the case, can we also say that programme states constitutes mental states? Searle says no: "The implemented syntactical or formal program of a computer is not constitutive of, or otherwise sufficient, to guarantee the presence of semantic content."[12] Syntax is different than semantics. Syntax is all about the arrangement of symbols, whereas semantics is the meaning derived from these arrangements.

If we think about how modern chatbots software works, a programme like ChatGPT, for example, can provide appropriate responses to user input by employing a language model that has been pre-trained on a large collection of data, allowing it to recognize patterns in the language it is communicating with. Searle would tell us that even such a sophisticated programme performs nothing more than a syntactic operation. Why? Because, unlike a human, it has no idea what this vast collection of data or the language it uses actually means. What it knows is how to conduct a dialogue as it statistically learns which questions or words are responded to with which answers or words in a conversation. The programme's situation is no different than the person in the room who does not understand Chinese but is able to communicate properly.

Searle's take on the issue shows that his view about the possibility of AGI is based on a certain assumption: that understanding requires consciousness. The term "consciousness" can have various meanings[13]. It can refer to awakeness, awareness, attention, experience, and is sometimes used interchangeably with the term "mind." In philosophy of mind, it has been argued that there are multiple types of consciousness[14,15,16]. Here, I use the term in its simplest sense, consistent with its etymology, to mean "the state of being aware". The word consciousness is derived from the Latin verb *conscio*, which means "I know together with" in the sense of sharing knowledge[17]. Just as you can share knowledge with someone else, you can also share it with yourself. In this regard, the term consciousness can also mean "being aware of what goes on in one's own mind." When we consider the term in this sense, we can say that Searle's Chinese Room argument makes the existence of consciousness essential for understanding. Although Searle does not state it explicitly, it is not difficult to infer this from his words: "I do not have any understanding of Chinese because I do not know what any of the words mean, I have no way to attach meaning to any of the symbols.[18]" The machine cannot attach meaning to the symbols because it does not have the knowledge that the relationship in question can be established

[9] Eric B. Baum, *What Is Thought?* (Cambridge: MIT Press, 2004), 1.

[10] John R. Searle, "Minds, Brains and Programs", *The Behavioral and Brain Sciences* (1980) 3.

[11] John R. Searle, *Minds, Brains and Science* (Cambridge: Harvard University Press, 1984), 30.

[12] John R. Searle, *Philosophy in a New Century* (New York: Cambridge University Press, 2008), 68.

[13] Adam Zeman, *Consciousness: A User's Guide* (New Haven and London: Yale University Press, 2002).

[14] Ned Block, "On a Confusion about a Function of Consciousness", *The Behavioral and Brain Sciences* (1995), 18(2): 247-287.

[15] David M. Rosenthal, *Consciousness and Mind* (New York: Oxford University Press, 2005).

[16] Christopher Hill, *Consciousness* (Cambridge: Cambridge University Press, 2009).

[17] Zeman, *Consciousness: A User's Guide*, 15.

[18] Searle, *Philosophy in a New Century*, 69.

(or cannot share this knowledge with itself). Therefore, it unconsciously or unknowingly matches symbols with other symbols.

The process of understanding, then, is inherently a conscious process. We cannot claim to understand something unless we are aware of the object of our understanding. Hence, the primary reason we do not attribute thoughts to ChatGPT is its lack of awareness of its processes. Consequently, consciousness might pose an obstacle to the development of thinking machines. Nevertheless, a behaviorist might raise an objection, contending that the indication of consciousness is nothing more than the manifestation of appropriate behavior. I will now argue that, even if we accept the behaviorist objection, consciousness can still be a potential impediment to our creation of thinking machines.

## 2. Artificial Intelligence and Consciousness

Although artificial intelligence studies are not being conducted for this purpose, as of 2023, no machine has passed the Turing test. For philosophers like Searle, it would not matter even if the test had been passed. Because the machine appears to us to be thinking, but in reality, it would not be thinking. This appearance-reality distinction lies at the basis of philosophical considerations about artificial intelligence. We have seen that this distinction is not well justified for behaviorists when it comes to attributing thoughts to artificial intelligence. Yet, there is one thing that prevents us from seeing the behaviorists as completely right and thus declaring humans as machines based only on appearance or behavior: It is humans who make the machine, but what makes humans is the evolutionary process, which spans millions of years.

Whether it operates through rule-based methods, relies on statistical modelling, or is the product of a new third-wave approach, we know that artificial intelligence technologies are ultimately based on algorithms, no matter how elegant or complex they may be. However, we do not know whether or to what extent the human mind is algorithmic. Artificial intelligence experts, who care about neurobiology or not, operate under the assumption that the brain fundamentally functions as an algorithm-based system in their quest for AGI. Those who find neurobiology important believe that uncovering hidden algorithms in the brain will pave the way for humanoid machines. Others believe that the complexity of the programmes alone will be sufficient.

Nonetheless, there is a chance that this endeavour or project may not produce the desired outcomes. Views have been advanced that address the informal knowledge problem mentioned by Turing, asserting that intuitive behaviors and common sense knowledge, resulting from evolutionary processes and our interaction with the world, are not a matter of computation. It has been argued that there is no conceivable algorithm for this aspect of intelligence, which needs to be understood on the path to AGI, and it may not be attainable in the near future with current approaches[19].

Common sense knowledge is ordinary knowledge shared by all, derived from experiencing the world with all its events and facts and serving as useful shortcuts in regulating our behavior. Babies can't talk; if you touch a hot pan, you'll get burned; you can't break the wall with your fist; you get wet when it rains, etc. are all examples of common sense knowledge. As we grow older, we acquire a huge amount of this type of information without even realizing it. It might seem extremely simple; yet, one of the most significant obstacles for artificial intelligence research to achieve human-level intelligence is to impart common sense knowledge to machines for effective use in practical matters. As G. Marcus and E. Davis observed, “the great irony of common sense—and indeed AI itself—is that it is stuff that everybody knows, yet nobody seems to know what exactly it is or how to build machines that have it.”[20]

More important than the amount of knowledge acquired is the ability to apply this knowledge to everyday situations, including new and unexpected events. For example, if someone were to ask us, "Is a cat or a dog a better driver?" even though we encounter this question for the first time, we can automatically provide an answer. We can do this not because we have never seen cats or dogs drive, but because we can reason from common sense knowledge that driving requires advanced cognitive skills and that cats and dogs do not have such skills. We somehow combine these two pieces of knowledge to produce an inference that cats and dogs cannot drive vehicles. Ronald J. Brachman and Hector J. Levesque called this “commonsense reasoning”[21]. Therefore, to say that a machine has common sense, it must possess common sense knowledge and be able to use common sense reasoning based on that knowledge.

For Erik J. Larson, a sceptical writer on AGI, common sense reasoning utilizes a different mode of inference than deduction and induction[22]. Deduction is what rule-based first-wave AI systems are good at, and induction is what

[19] Hubert L. Dreyfus, *What Computers Still Can't Do: A Critique of Artificial Reason* (Cambridge: MIT Press, 1992). Brian C. Smith, *The Promise of Artificial Intelligence: Reckoning and Judgment* (Cambridge: MIT Press, 2019). Herbert L. Roiblat, *Algorithms Are Not Enough: Creating General Artificial Intelligence* (Cambridge: MIT Press, 2020). Erik J. Larson, *The Myth of Artificial Intelligence: Why Computers Can't Think the Way We Do* (Cambridge: Harvard University Press, 2021).

[20] Gary Marcus & Ernest Davis, *Rebooting AI: Building Artificial Intelligence We Can Trust* (New York: Pantheon Books, 2019).

[21] Ronald J. Brahman & Hector J. Levesque, *Machines Like Us: Toward AI with Common Sense* (Cambridge: MIT Press, 2022).

[22] Larson, *The Myth of Artificial Intelligence: Why Computers Can't Think the Way We Do*.

machine learning second-wave AI systems can implement successfully. Both have limitations that make them unsuitable for AGI. The limitations of deduction stem primarily from its nature, as it is unable to generate new infromation about the world. All deduction can do is to ultimately reveal the knowledge implicitly present in the premises. Second, deduction can be perfectly valid for wrong reasons, or the reasons may be irrelevant to the desired outcome. Third, deduction aims at certainty, whereas real-world situations almost always contain contingent factors.

Induction is based on the assumption that future experiences will be similar to past experiences or that the continuity of past experiences will likely extend into the future without any unexpected occurrences. This approach may be useful for specific applications or programmes where it is important for experiences to match expectations; however, it falls short when dealing with unpredictable and constantly changing, dynamic real-world environments. In a real-world environment where surprises do not lack and which confronts us with novel experiences, we mostly rely on common sense.

According to Larson, in everyday life situations, we reason from common sense by leaning not on deduction or induction but on another category of logical reasoning known as abduction. Abduction is the inference to the best explanation that we make when we are dealing with a particular event that requires an account to understand. For example, if I find the tire of my car, which was in good condition the day before, is now flat, my first guess would be that I ran over something sharp. While there could be multiple explanations or reasons for the flat tire (there may be a manufacturing defect in the tire; someone may have punctured my tire on purpose; tires may burst on their own because they do not obey the laws of physics, etc.), my previous life experiences or common sense knowledge suggest that this is the most plausible explanation.

Although such conjectural procedures are conducted automatically by our brains, the representation of this process of automatization in computing machines appears to be quite problematic. Larson calls this *the selection problem*:

> The selection problem is finding the operative or best or plausible cause, given all the possibilities real or imagined. So the core problem of automating abductive inference can be recast as this problem of selection, which helps expose the difficulty inherent in the required inference, but in the end it's really the same problem. To abduce we must solve the selection problem among competing causes or factors, and to solve this problem, we must somehow grasp what is relevant in some situation or other. The problem is that no one has a clue how to do this. Our actual inferences are often guesses, considered relevant or plausible—not deductions or inductions. That's why, from the standpoint of AI, they seem magical[23].

It is possible to see Larson's argument as an elaboration of Searle's. While Searle claims that computers cannot understand natural language due to their reliance on syntactic operations, he does not explain why they cannot extract semantics from syntax, merely noting that brains have this capability. According to Larson, on the other hand, computers cannot do this because they do not have the common sense knowledge to derive meaning, nor do they have a method to use this common sense knowledge in common sense reasoning, which is mostly based on abductive inference. The key difference between the two is that while Searle assumes that understanding natural language requires consciousness and thus introduces an ontological dimension to the issue, Larson views the problem as fundamentally epistemological, citing the lack of knowledge and methods for representing common sense in machines, which he sees as a prerequisite for understanding.

Thus, if a route is established for machines to possess common sense, Larson would accept that they understand and therefore think, regardless of whether they are aware of it or not. Accordingly, Larson's position, unlike Searle's, is compatible with behaviorism. This means that, if a machine can convince us that it is conscious with its newly acquired commonsensensical communication and action skills, there would be no valid reason to deny the attribution of consciousness to that machine. However, we may still need consciousness, not as imitation but as a real phenomenon. Let me explain.

Even though defining intelligence precisely or delineating its components is a problem in itself[24], it can be generally understood as encompassing the adept utilization of knowledge and skills to address challenges or accomplish objectives, as well as the capacity to make sound and fitting decisions in the face of difficulties[25]. Since transitioning machine intelligence from its current weak version to a possible strong version means transforming or replacing specific problem solvers into/with general problem solvers, and since intuitive-common sense behavior is a fundamental feature of general problem solvers like humans, it seems crucial to identify a way or method for incorporating this information into potential AGI systems. This is what Larson's argument entails.

[23] Larson, *The Myth of Artificial Intelligence: Why Computers Can't Think the Way We Do*, 83.

[24] David Cycleback, *Philosophy of Artificial Intelligence* (London: Bookboon, 2018), 19-20.

[25] Daeyeol Lee, *Birth of Intelligence: From RNA to Artificial Intelligence* (New York: Oxford University Press, 2020).

Given that common sense behavior heavily relies on the subject's experience of the world, it may be necessary to provide a machine with a physical presence or body and situate it in the world. This would enable the machine to learn through interactions with its environment, form beliefs, and acquire diverse cognitive skills[26]. This is where consciousness comes into the picture. It can be claimed that consciousness is an integral part of systems that develop their own intelligence[27]. Simply put, a person develops a mind during the growth process, and consciousness and intelligence are complementary elements of this mind, which appears as a holistic phenomenon. Just as we do not see an intelligent person who is not conscious, we also do not see a conscious person who does not have intelligence.

In addition, consciousness is the gateway through which humans open up to the world. Without consciousness, experiencing the world would not be possible. This aspect (or, for some, type) of consciousness is commonly known as *phenomenal consciousness*. Phenomenal consciousness is particularly related to the feelings generated by experiences in the subject. Thus, someone who claims that intelligence has nothing to do with consciousness will have to argue that subjective experience does not contribute to a person's cognitive development, which is not a task that most people would want to undertake. As a matter of fact, even watching a baby's development provides strong evidence against this claim. Still, the fact that intelligence is something to do with consciousness does not mean that it cannot be achieved without consciousness.

Alternatively, we can continue to take a behaviorist position and consider the machine-human analogy in terms of behavior and function rather than structure. In this way, the question of whether the human mind is algorithmic or not becomes irrelevant to the goal of achieving AGI. If we manage to develop artificial intelligence systems that demonstrate cognitive capabilities on par with humans through observable behaviors, it would be sufficient to declare that machines are thinking. Nevertheless, even within this framework, we may have to take consciousness into account.

Let's go back to the Chinese Room. The reason why we could not attribute thoughts to the person in the room even though he was conducting the dialogue was because he did not understand the elements of the dialogue (words, sentences, phrases, expressions, etc.). The act of understanding necessitates awareness of what is understood, and being aware of something means being conscious of it. If understanding is considered a marker of intelligence, an unconscious machine would not be expected to achieve AGI.

A behaviorist could counter this argument by asserting that, just as the internal states of humans are not directly accessible to others, the same holds true for machines. So, it does not matter whether the machine *actually* understands and therefore is conscious. It is enough for it to appear to understand or be conscious. Ray Kurzweil, a renowned futurist, expresses this point in his latest book: "I will accept nonbiological entities that are fully convincing in their emotional reactions to be conscious persons, and my prediction is that the consensus in society will accept them as well."[28] Kurzweil may be right, but for this to happen, machines must be, as he says, "fully convincing." But it's unclear how a machine can be fully convincing without feeling anything.

A fully convincing machine that feels nothing is equivalent to a philosophical zombie. A philosophical zombie is quite different from the fictional zombies we encounter in horror movies. Fictional zombies are typically depicted as reanimated corpses or infected individuals who lack higher cognitive functions and are driven by a primal urge to consume living flesh. In contrast, philosophical zombies refer to beings that is behaviorally and functionally indistinguishable from a normal human but lack conscious experiences or phenomenal consciousness. While it is widely acknowledged that a philosophical zombie is not physically possible, it is debatable whether it is logically or even metaphysically possible[29]. If it is metaphysically possible, then, a machine lacking phenomenal consciousness could fully convince us that it is conscious, just like any other human being.

The idea of a philosophical zombie seems counterintuitive. If such a being is not possible, then, the only option for machines to convince us is imitation. Accordingly, a machine that can imitate all aspects of human behavior can convince us that it is conscious. The problem here is that there may be aspects of human behavior that cannot be replicated through imitation. For example, subjective experiences of humans, with various subtleties and nuances, can have a significant influence on their behavior. It is doubtful how a machine that does not feel anything could exhibit the same type of behavior through programmes instead of feelings and emotions and thus convince us that it has the

[26] This approach is known as "situated AI" in the philosophy of artificial intelligence (Rodney A. Brooks, "Intelligence without Reason", in *The Artificial Life Route to Artificial Intelligence: Building Embodied, Situated Agents*, Ed. Luc Steels & Rodney A. Brooks (New Jersey: Lawrence Erlbaum Associates, 1995). In the article we mentioned, Alan Turing proposed this approach as an alternative to the traditional approach to artificial intelligence, which focuses on specializing in certain intellectual activities (such as playing chess or quizzes), and suggested that the development of artificial intelligence "could follow the normal teaching of a child" (Turing, "Computing Machinery and Intelligence", 460).

[27] Klaus Henning, *Gamechanger AI: How Artificial Intelligence is Transforming our World* (Cham: Springer, 2021).

[28] Ray Kurzweil, *How to Create a Mind: The Secret of Human Thought Revealed* (New York: Viking Penguin, 2012).

[29] Robert Kirk, "Zombies", *The Stanford Encyclopedia of Philosophy* (Fall 2023 Edition), Ed. Edward N. Zalta & Uri Nodelman, URL = https://plato.stanford.edu/archives/fall2023/entries/zombies

experiences in question. If consciousness cannot be faked, and if a philosophical zombie is not possible, then achieving true artificial intelligence without creating conscious machines may remain an unrealized dream.

## Conclusion

From a broad perspective, we can talk about two primary approaches to the development of artificial intelligence. The traditional approach, currently in practice, focuses on creating sophisticated algorithms with expanding problem-solving capabilities and demonstrating intelligent behavior in certain areas without paying much attention to how the brain can acquire intelligence. However, it remains unclear how these algorithms will evolve into the singular or multiple algorithms necessary for achieving artificial general intelligence.

On the other hand, the alternative approach takes seriously the human brain's ability to produce intelligence and posits that AI systems should undergo a similar process to achieve AGI considering that human life experiences are key to achieving general intelligence. Two problems arise here. First, considering the development of intelligence in all its aspects—since human evolutionary history cannot be excluded from this process—the alternative approach may have to face the challenge of going back to the beginning of life and deciphering the first algorithms of life. Second, while humans can effortlessly use a vast amount of common sense knowledge in common sense/abductive reasoning in their intelligent behavior, it remains a mystery how this knowledge and skill will be integrated into artificial intelligence systems.

For both approaches, consciousness—both in terms of awareness and its phenomenal aspect—seems to pose a problem. One could argue that consciousness is what makes cognitive operations genuine. If understanding, decision-making, planning, and thinking in general inherently involve awareness, then a machine without consciousness would not be expected to have real (human-like) intelligence. Considering the impact of subjective experiences on cognitive processes, this expectation becomes even weaker.

Still, there is the possibility that even in the absence of genuine intelligence, a machine could convincingly simulate it. Unless an algorithm for feeling can be devised (if such an algorithm exists at all), this probability seems exceedingly slim. However, if unexpected happens and the machines somehow manage to deceive us, the boundary between appearance and reality would blur, and then we will begin discussing machine rights.

**Peer Review:** Externally peer-reviewed.
**Conflict of Interest:** The author has no conflict of interest to declare.
**Grant Support:** The author declared that this study has received no financial support.

**Hakem Değerlendirmesi:** Dış bağımsız.
**Çıkar Çatışması:** Yazar çıkar çatışması bildirmemiştir.
**Finansal Destek:** Yazar bu çalışma için finansal destek almadığını beyan etmiştir.

## ORCID ID of the author / Yazarın ORCID ID'si

Ferhat Onur 0000-0001-7052-2881

## REFERENCES / KAYNAKLAR

Baum, Eric B. *What Is Thought?* Cambridge: MIT Press, 2004.

Block, Ned. "On a Confusion about a Function of Consciousness". *Behavioral and Brain Sciences* 18(2) (1995): 247-287.

Bostrom, Nick. "How Long Before Superintelligence?" *Linguistic and Philosophical Investigations* 5(1) (2006): 11-30.

Brachman, Ronald J. & Levesque, Hector J. *Machines Like Us: Toward AI with Common Sense*. Cambridge: MIT Press, 2022.

Brooks, Rodney A. "Intelligence without Reason." *The Artificial Life Route to Artificial Intelligence: Building Embodied, Situated Agents*. Editor Luc Steels & Rodney Brooks, 25-81. New Jersey: Lawrence Erlbaum Associates, 1995.

Cycleback, David. *Philosophy of Artificial Intelligence*. London: Bookboon, 2018.

Dreyfus, Hubert L. *What Computers Still Can't Do: A Critique of Artificial Reason*. Cambridge: MIT Press, 1992.

Goodfellow, Ian, Bengio, Yoshua & Courville, Aaron. *Deep Learning*. Cambridge: MIT Press, 2016.

Henning, Klaus. *Gamechanger AI: How Artificial Intelligence is Transforming our World*. Cham: Springer, 2021.

Hill, Christopher. *Consciousness*. Cambridge: Cambridge University Press, 2009.

Kersting, Kristian. “Rethinking Computer Science Through AI.” *KI – Künstliche Intelligenz*, 34 (2020): 435-437.

Kirk, Robert. “Zombies”. *The Stanford Encyclopedia of Philosophy* (Fall 2023 Edition). Editor Edward N. Zalta & Uri Nodelman, URL = https://plato.stanford.edu/archives/fall2023/entries/zombies/

Kurzweil, Ray. *How to Create a Mind: The Secret of Human Thought Revealed*. New York: Viking Penguin, 2012.

Larson, Erik J. *The Myth of Artificial Intelligence: Why Computers Can’t Think the Way We Do*. Cambridge: Harvard University Press, 2021.

Lee, Daeyeol. *Birth of Intelligence: From RNA to Artificial Intelligence*. New York: Oxford University Press, 2020.

Marcus, Gary. & Davis, Ernest. *Rebooting AI: Building Artificial Intelligence We Can Trust*. New York: Pantheon Books, 2019.

Roitblat, Herbert L. *Algorithms Are Not Enough: Creating General Artificial Intelligence*. Cambridge: MIT Press, 2020.

Rosenthal, David M. *Consciousness and Mind*. New York: Oxford University Press.

Sabouret, Nicolas. *Understanding Artificial Intelligence*. Boca Raton: CRC Press, 2021.

Searle, John R. “Minds, Brains and Programs.” *The Behavioral and Brain Sciences*, 3 (1980): 417-424.

Searle, John R. *Minds, Brains and Science*. Cambridge: Harvard University Press, 1984.

Searle, John R. *Philosophy in a New Century*. New York: Cambridge University Press, 2008.

Shannon, Claude. & McCarthy, John. “Preface.” *Automata Studies*, Editor Claude E. Shannon & John McCarthy, v-viii. Princeton: Princeton University Press, 1956.

Smith, Brian C. *The Promise of Artificial Intelligence: Reckoning and Judgment*. Cambridge: MIT Press, 2019.

Turing, Alan M. “Computing Machinery and Intelligence.” *Mind*, 49 (1950): 433-460.

Walsh, Toby. *Machines That Think: The Future of Artificial Intelligence*. New York: Prometheus Books, 2018.

Zeman, Adam. *Consciousness: A User’s Guide*. New Haven and London: Yale University Press.

**How cite this article / Atıf Biçimi**

Onur, Ferhat. “On the Relationship Between General Artificial Intelligence and Consciousness” *Felsefe Arkivi-Archives of Philosophy, 60* (2024): 15-23. https://doi.org/10.26650/arcp.1417812