# Artificial consciousness: from AI to conscious machines

Nehal Mamdouh, Maadi STEM High School for Girls

Yossef Eldershaby, Gharbiya STEM high school

**Mentor:** Roaa El-fishawy, STEM High School for Girls - Maadi

## Abstract

*Consciousness is a subjective (implicit) experience. Artificial consciousness aims to simulate this consciousness. This is by building a model as complex as a human brain. Any model less complex than the brain will not be able to simulate the human brain nor a part of it. Building a subject is one of the biggest difficulties because scientists till now don't know what specifically a subject is. Consequently, it is impossible to build something you don't know it. Many attempts tried to build machines able to do tasks with the same proficiency as humans. Many attempts succeeded as deep blue that beat the chess world champion Garry Kasparov, but this didn't reach human consciousness yet. It just follows specific complex commands. This category of machines lacks emotions, love, creativity, desire, and curiosity. Now, scientists try to model the brain by RAM which every neural connection (synapse) equals a floating-point number that requires 4 bytes of memory to be represented in a computer. The brain contains $10^{15}$ synapses that equal 4 million GBs of RAM. This memory is not available on a computer till now. It is predicted that it will be available near 2029. This idea may fail for any reason, but all researchers, scientists, and technologists believe that artificial consciousness will become a reality someday even in the far future.*

## VIII. Introduction

### i. What is Consciousness?

Consciousness is one of the most mysterious scientific concepts. Scientists till now discover more about the methodology of human consciousness. Consciousness is everything you experience and everything you feel, which are sometimes named qualia. Many modern philosophers believe that this is just an illusion as they believe should be a meaningless universe of matter and void[1]. Logically this is wrong as it doesn't depend on any scientific reason and it is opposite to the real situation that these experiences, by way or another, exists.

In 2018, David Gamez, a Lecturer at Middlesex University, developed another explanation for consciousness: over the last 3 centuries, science has developed a series of interpretations of the world that have stripped objects of their sensory properties. You consciously deal with an apple as a red and tasty object, but scientifically apples are colorless

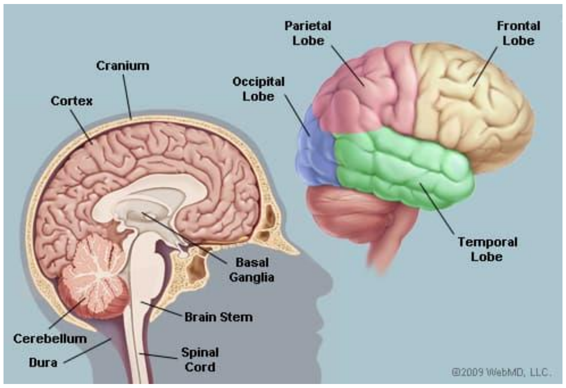

Figure 1: Human Brain Anatomy

collections of jigging atoms. These colors, sounds, smells and all sensations that we encounter in our daily life need to be associated with something and this thing is consciousness. Gamez defines “consciousness” as another name for our bubbles of experience, which contain the sensory properties that science removed from the physical world [2].

### ii. The rise of Artificial Consciousness accompanied by Artificial Intelligence

The rise of AI especially and technology generally in the 20th century caused the foundation of a new field related to AI which is Artificial Consciousness (AC). The idea of the universal effective AI model, which is creating machines as have all human aspects, is the reason that scientists created a new field [3] [4] [5]. Scientists consider Artificial Consciousness as a branch or sub-field of AI. The reason is that Artificial Intelligence (AI) is the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings (human activities)[6], Artificial Consciousness is the simulation or the use of AI to model a conscious machine. The ambiguity here is as mentioned before, consciousness is one of the most ambiguous scientific concepts about humans till now, and AI depends on computations, algorithms, processing, and functions of AI method to simulate human activities, but consciousness is thought to be untouchable with those methods.[7]

Artificial Consciousness is mainly inspired by human imagination. This is proved by the idea that the first spot on intelligent robots was by sci-fi movies and stories. Since the early 1950s, sci-fi movies depicted robots as human-crafted machines able to perform complex operations, work with us on critical missions in hostile environments, or pilot and control spaceships in galactic travels[7].

The most famous example of this archetype, HAL 9000, the main character in Stanley Kubrick’s 1968 epic, *2001: A Space Odyssey*. HAL controls the entire spaceship, talks as a human with the astronauts, recognizes the crew’s emotions and renders aesthetic judgments. it also murders astronauts in pursuit of a plan elaborated from flaws in its programming. On the other side, it plays chess too. This theme was developed by time from James Cameron’s terminator in 1984 to terminator 2 in 1991 and the matrix in 1999[7]. Nevertheless, the term of AC was not used in those movies. It was just an interesting fantasy theme. The words “Artificial consciousness” were first used in the book *Kibernetikai ge´pek*, by *Tihame´r Nemes*, the author, in 1969. He wrote a paragraph about Artificial Consciousness indicating the features of a conscious machine.

AC is a controversial concept because it gives rise to several issues that require combining much information from different disciplines especially computer science, neurophysiology, and philosophy[7]. AC can be classified into two kinds:

Weak artificial consciousness: It is a simulation of conscious behavior. implementation of a smart program that simulates the behaviors of a conscious being at a primary level of technology and AI, without understanding the mechanisms that generate consciousness[8]. Something like a primary model.

Strong artificial consciousness: It refers to real conscious thinking emerging from a complex computing machine (artificial brain). In this case, the main difference with respect to the natural counterpart depends on the hardware that generates the process[8].

This review will focus on the development of artificial consciousness from weak to strongest predicted version. It focuses on the biological and psychological perspectives too, conducting many questions about AC and aiming to solve it by representing the rise and the development of AC and what ambiguities it faced chronologically.

## IX. Consciousness, biological process, or psychological concept:

Consciousness is a subjective experience. What “it is like” to perceive a scene, to endure pain, to entertain a thought, or to reflect on the experience itself. When consciousness fades, as it does in dreamless sleep, from the intrinsic perspective of the experiencing subject, the entire world vanishes. Consciousness mainly depends on the integrity of certain brain regions and the particular content of an experience depends on the activity of neurons in parts of the cerebral cortex (look at fig 1 [9]). In fact, refined clinical and experimental studies are not sufficient for understanding the relationship between consciousness and the brain. It is still anonymous why the cortex supports consciousness when the cerebellum does not, despite having four times as many neurons [10].

As a prescientific term, “Consciousness” is used in widely different senses. A machine must be turned on properly for its computations to unfold normally. Distinguishing two other essential dimensions of conscious computation can be useful. We label them using the terms global availability and self-monitoring. C1: Global availability corresponds to the transitive meaning of consciousness (as in “The driver is conscious of the light”). It refers to the relationship between a cognitive system and a specific object of thought, such as a mental representation of “the fuel-tank light.” This object appears to be selected for further processing, including verbal and nonverbal reports. Information that is conscious in this sense becomes globally available to the organism; for example, we can recall it, act upon it, and speak about it. This sense is synonymous with “having the information in mind”; among the vast repertoire of thoughts that can become conscious at a given time, only that which is globally available constitutes the content of C1 consciousness[11]

In at least three quite different ways the term “consciousness” has been used.

(1) It has sometimes been defined as a state, as in drowsy, alert or altered states of consciousness.

(2) It has also been used to refer to an architectural concept, namely the executive system at the center of cognition that seems to receive input, allocate attention, set priorities, generate imagery, and initiate recall from memory.

(3) It may be used as an indicator of representational awareness, as in becoming conscious of some specific idea or event.

From an evolutionary perspective, these different meanings of the word elicit very different explanations in terms of biological fitness and selection pressures; moreover, their underlying mechanisms appear in evolutionary history at different times and in different species. For example, the brain mechanisms of state variables would appear to be more fundamental than the other two, since basic arousal and sleep mechanisms evolved early and are essentially similar in many mammals. The neural machinery supporting our putative attentional architecture comes next in the evolutionary hierarchy since it is concerned mostly with the control of complex behaviors and only appears in fairly advanced organisms.[12]

## X. AC ORIGIN

### i. The origin of the term “AC”

Engineers are in always attempt to design something, which could not be defined precisely. They aimed at building artificial replicas which imitated some features of something, real or virtual, that elicited their imagination [13]. On the other hand, neuroscientists like Giulio Tononi and Gerard Edelman claimed that [14]: To understand the mind we may have to invent further ways of looking at brains, and here is how it starts:

### ii. AC tech discipline

Artificial consciousness is a technological area closer to robotics and AI technical fields. It is not surprisingly a scientific discipline and has a limited relation to psychology or neurosciences. Nevertheless, in the future, artificial consciousness could give unexpected contributions to the understanding of the study of the human mind because it is a reliable testbed for checking theories and hypotheses. Artificial consciousness is perfectly described as “epigenetic robotics” both disciplines stress the role of development. However, artificial consciousness leaves the implementation of the sensory-motor-cognitive system to epigenetic robotics. In simpler terms, AC is addressing the issue of the robot with the external environment. because artificial consciousness sits on two giants’ shoulders (neurosciences and artificial intelligence), researchers are not seeking to make a confusing use of linguistic terms, Today, the term ‘artificial consciousness’ has a pure technological meaning.

Researchers often use consciousness in its folk psychology and everyday meaning. The researchers in the field of artificial consciousness know well that the study of natural consciousness is far from being conclusive [15].

Researchers adopt a typical engineering attitude. They build artifacts that evoke characteristics of a human being from the scratch. However, they do not want to insert a module (a sort of ‘consciousness module’) in a pre-assembled robot. They want to build a conscious-like robot, i.e. a robot which behaves like a conscious being. Very often, engineers build artifacts before knowing exactly the laws which are at the basis of the processes and methods used in the construction of the artifact itself (engineers design and build proteins even if they do not know the laws governing the protein folding in 3D). Ray Kurzweil writes:

The question here should be “how will we come to terms with the consciousness that will be claimed by non-biological intelligence?” Such claims will be accepted from a logical practical perspective for only one thing “they” will turn into “us”, so there won’t be any clear distinction between non-biological and biological intelligence. Furthermore, these non-biological entities will be extremely intelligent, so they’ll be able to convince other humans of their consciousness:

- They’ll have the delicate emotional cues, which convince us today that humans are conscious.
- They will be able to make other humans feel contradictory feelings.
- They’ll get mad if others don’t obey their claims.
  But, this is fundamentally a political and psychological prediction, not a philosophical argument. [16]

iii. From AI to AC

“Mind cannot be demonstrated as identical to brain activity” an equivalence that Bennett and Hacker regarded as a metrological fallacy [3]. We experience only humanoid consciousness as a whole and not in one of his/her parts, not even in a neural sophisticated part like the brain. If the concept of a man-made artifact that acts like a human being were accepted, artificial consciousness would acquire a new status and it would be an updated version of artificial intelligence.

In 2005, Teed Rockwell wrote in his issue, Neither Brain nor Ghost, that one of the biggest mistakes of symbolic systems AI was to substitute the propositions that are caused by experience for the experience itself—this may be right in case of linguistic experience only, but the experience is not limited to linguistic affair only. Those AI researchers, who limited experience to linguistic affairs only, saw common sense as a particular set of concepts [17].

Another attempt by other researchers was to translate common sense into a set of propositions and store all the propositions in their machines’ memories. But it is clear to almost everyone that it was a doomed project. It was necessary to program in even statements as obvious as ‘when you put an object on another object and move the bottom object, both objects move’, one of many statements that never has to be verbalized by anyone who has a body and conscious brain and has used them to try to move those objects then failed and recognized why this phenomenon happened then stored this observation or conclusion in form of experience consciously to use it in other situations.

## XI. AC technical development and difficulties

i. How to verify consciousness (Turing test)?

In 1950, Alan Turing, an English mathematician, computer scientist, logician, philosopher, and theoretical biologist, tried to answer one of the most ambiguous questions at this time “can computers think?” Turing considered the machines as digital computers only and operationalized thinking as the ability to answer questions in a particular context. The test is to ask a question for a computer and the same question for a human operator. Both answer it on a keyboard for 5 minutes. The answer should be well enough that the interrogator could not easily discriminate between the human and computer. The examiner inputs a question about anything that comes to his mind. Both the computer and the human respond to each question. If the examiner cannot with confidence distinguish between the computer and the operator based on the nature of their answers, we must conclude that the machine has passed the Turing test[18] [7] [19].

Point to consider, Turing in his original paper didn't mention consciousness except in the context of an objection that the thought in the brain is always driven and accompanied by feeling. In other words, consciousness is feeling. So, the text generated by the machine in the absence of feeling, however it seems convincing, could not be taken as a sufficient indicator of thought[19].

In 1990, the Turing test received its first formal acknowledgment. Hugh Loebner, a New York philanthropist, and the Cambridge Center for Behavioral Studies in Massachusetts established the Loebner Prize Competition in Artificial Intelligence. It was awarded a $100,000 prize for the first computer which succeeded in the Turing test[7] [19].

### ii. Development of Turing test

In 1998, the questioning's scope has been wider and nearly include anything. Each judge selects a score on a scale of 1 to 10. 1 means human and 10 means computers. Now, current computers can pass the Turing test (pass here means confidence distinguish between the computer and human) in case presence of restrictions to interact to highly specific topics as chess. So far, no computer has given responses indistinguishable from a human, but every year the computer's scores edge closer to an average of 5. The possibility of building a device that will pass the human Turing test, at least in the far future, is not ruled out yet[18] [19].

More recently, people have suggested extensions to the standard test that involve processing more material beyond text as audio, visual data, or controlling a humanoid body in a human-like way for an extended time interval. By these suggested modifications, passing any of the Turing tests needs a machine would almost certainly have to have experience of the world, a capacity for imagination, and emotional behavior. Because there is no sequence of pre-programmed responses is likely to be convincing over an extended time [19].

### iii. Computer beats human

On 11 May 1997 at 3:00 P.M. in New York City, for the first time in the history, a computer beat reigning world chess champion, Garry Kasparov. It was IBM's Deep Blue. It is estimated that the search space in a chess game includes about 10,120 possible positions. Deep Blue could analyze 200 million positions per second. Deep Blue victory can be explained by its speed combined with a smart search algorithm, able to account for positional advantage. In other words, computer superiority was due to brute force, rather than sophisticated machine intelligence. The conflict here is whether this means that Deep blue is conscious or not[20] [7] [21]. If Turing's thought was applied in this case, so this means that Deep Blue is conscious. Because Kasparov expressed doubts while he was playing against the computer. Sometimes, he felt like playing against a human, not a programmed machine. In some situations, he appreciated the beauty of the moves done by the machine, as if it was driven by intention, rather than by algorithms. So, this asserts that if the Turing test was conducted here, Deep Blue will certainly pass the test as its performance can't be distinguished from human performance. so, in Turing perspective, Deep Blue is conscious.

The conflict here when considering another perspective. From a pragmatic perspective, Turing may be right. We as humans believe that anyone like us is self-conscious too. The reason is that we consider the similarity between us as a factor, this person has as same organs as me and has the same brain too. So, he is self-conscious like me. Nevertheless, if something with a different structure as mechatronics organs, neural processors, and technological parts, but behave as a human. The answer now may be different, the possibility of being self-conscious is not low. In the case of Deep Blue, it doesn't behave like humans. It is like a calculator. It does the ordered process, but does Deep Blue or Calculator understand what they do? They both apply or take procedures by following specific given algorithms or commands with a difference in complicity[20] [7]. It can be said that this category of machines as a calculator or Deep Blue is driven by electronic circuits working in a fully automated mode. It doesn't show creativity, love, emotions, or unlogic decisions depending on desire or living-organism instinct. Just act as operated slave[20].

### iv. Difference between subject or object:

Manufacturers work on building sophisticated robots called epigenetic robots. They aim to reach unique personalities for robots through the interaction with the environment and make them capable of going through a series of development phases of a normal human (from toddler to adult). This idea appeals to consumers. Moreover, robots must show emotions like happiness, anger, surprise, and sadness, in different degrees. Those robots must be curious and able to explore the external world on their own: these robots develop concerning their personal history. However, designing and implementation of robots capable of having a subjective experience of what happens to them are not achieved yet.

The recent research on consciousness is focused on the design of conscious machines. The time has come to elevate from behavior-based robots to conscious robots. before any new design approach towards a new generation of artificial beings, engineers have to deal with a new problem: how to build a subject? engineers didn't use to build subjects before[7] [4].

Implicit, which is subject, in many theories is explained as an external event and is represented in the brain (the object), so they are connected, but from the same perspective they are different, they are not the same thing. This undemonstrated hypothesis is the reason why it is hard to address what is consciousness. So, this makes building subjects nearly impossible till scientists clearly demonstrate what is consciousness[7] [4].

### v. Is it possible the machines could be conscious?

In 1980, the philosopher John Searle presented his proof that machines could not possibly think or understand. The reason is that computers do human tasks but in an unintelligent manner. He believes that no matter how good the performance of the program if it can't think and understand [22]. Nevertheless, this assumption has some problems. If this reason was applied to the biological counterpart or human brain. In fact, those are also biologically operated to respond to specific inputs by specific reactions and each neuron automatically responds to any input according to fixed natural laws. Each neuron or cell does its function, but it could not possibly think or understand why it did that. In other words, cells are not conscious. However, this does not prevent us from experiencing happiness, love, and irrational behaviors. This negates Searle's assumption[20].

In 2015, a book called "Impossible Minds" discussed this topic from a scientific rigor aspect. It addressed the diversity between biological and artificial brains. They both can do the same task in different ways. This doesn't matter on the result that we observe, which is consciousness. So, the issue became in our beliefs. If we believe in AC concerning our religious regulations. This means that the possibility of realizing an artificial self-aware being remains open[23].

All research indicates that AC is possible. No one reason negates this fact. It is not achieved yet. But it is possible, and scientists predict that it will happen in the near following years.

## XII. AC future

i. When will a machine become self-aware?

After proving that creating conscious machines is not impossible yet. The scientific answer to this question is controversial, but it is possible to indicate a condition that must happen to consider the machine as self-aware. A neural network must be as complex as the human brain or more because less complex brains are not able to produce conscious thoughts. It will not produce any conscious thoughts (see figure 2 [7]). Consciousness is a step function of brain complexity[20].

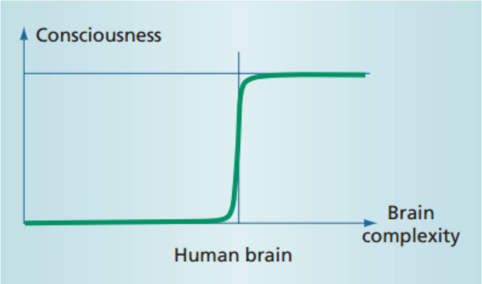

Figure 2: The self-awareness threshold

Since memory is used to simulate the human brain. What is the capacity of the memory needed to equate the brain in complicity? The human brain contains about $10^{12}$ neurons, and each neuron makes about $10^{3}$ connections (synapses) with other neurons. So the total equals $10^{15}$ synapses. Each synapse can be simulated by 4 bytes. In consequence, 4 x $10^{15}$ bytes (4 million Gigabytes). Then, Is such a memory available on a computer? Since 1980, the RAM capacity has increased exponentially by a factor of 10 every 4 years (see figure 3 [20]).

So, bytes = $10^{((\text{year} – 1966)/4)}$

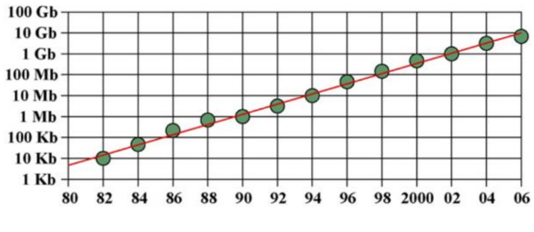

Figure 3: Typical random-access memory installed in personal computers

We just have to substitute that number in the equation above and compute the result. The answer is the year 2029. In any case, even if we adopt different numbers, the computation's basic principle remains the same. we could advance that date by only a few years.

## XIII. Conclusion

Consciousness refers to your personal perception of your unique thoughts, memories, feelings, and environments. Essentially, your consciousness is your awareness of yourself and the world around you. This awareness is subjective and unique. From a neurological perspective, Science is still exploring the neural basis of consciousness. But even if we have a complete neuroscience picture of how the brain works or performs, many philosophers still believe that there is still a problem they call the "Consciousness Problem." The brain is the most complex organ in the entire universe as we know it. It has about 100 billion neurons. It has more neural connections than there are stars in the entire universe. This is why we are incredible beings who have a spark of consciousness.

A popular discussed approach to achieve general intelligent models that can be conscious is whole depending on reaching a "brain simulation". The low-level brain model is created by scanning and mapping the biological brain in detail and copying its state to a computer system or other computing device.

Eventually, it is possible to indicate a condition that must happen to consider a machine as self-aware or conscious. A neural network must be at least as complex as the human brain because less complex brains are not able to produce conscious thoughts. Actually, it will not produce any conscious thoughts. Scientists and technical now work on building a model as complex as the human brain. It is just a prediction. However, they believe that AC will reach human consciousness in 2029. Even this attempt failed, AC will reach human consciousness even in the far future.

## XIV. References

[1] C. Koch, "What Is Consciousness?," Nature, vol. 557, no. 7704, pp. S8–S12, May 2018, doi: 10.1038/d41586-018-05097-x.

[2] D. Gamez, Human and Machine Consciousness. Open Book Publishers, 2018. doi: 10.11647/OBP.0107.

[3] R. Chrisley, “Philosophical foundations of artificial consciousness,” Artif. Intell. Med., vol. 44, no. 2, pp. 119–137, Oct. 2008, doi: 10.1016/j.artmed.2008.07.011.

[4] R. Manzotti and V. Tagliasco, “Artificial consciousness: a discipline between technological and theoretical obstacles,” Artif. Intell. Med., vol. 44, no. 2, pp. 105–117, Oct. 2008, doi: 10.1016/j.artmed.2008.07.002.

[5] “Fuzzy Comprehensive Evaluation-based artificial consciousness model.” https://ieeexplore.ieee.org/document/5554447 (accessed Sep. 13, 2021).

[6] “artificial intelligence | Definition, Examples, and Applications,” Encyclopedia Britannica. https://www.britannica.com/technology/artificial-intelligence (accessed Sep. 15, 2021).

[7] G. Buttazzo, “Artificial consciousness: Utopia or real possibility?,” Comput. Vol. 34 Issue 7 July 2001, pp. 24–30, Jul. 2001.

[8] O. Holland, Machine Consciousness. Imprint Academic, 2003.

[9] “Brain (Human Anatomy): Picture, Function, Parts, Conditions, and More.” https://www.webmd.com/brain/picture-of-the-brain (accessed Sep. 20, 2021).

[10] G. Tononi, M. Boly, M. Massimini, and C. Koch, “Integrated information theory: from consciousness to its physical substrate,” Nat. Rev. Neurosci., vol. 17, no. 7, pp. 450–461, Jul. 2016, doi: 10.1038/nrn.2016.44.

[11] S. Dehaene, H. Lau, and S. Kouider, “What is consciousness, and could machines have it?,” Science, vol. 358, no. 6362, pp. 486–492, Oct. 2017, doi: 10.1126/science.aan8871.

[12] M. Donald, “The neurobiology of human consciousness: an evolutionary approach,” Neuropsychologia, vol. 33, no. 9, pp. 1087–1102, Sep. 1995, doi: 10.1016/0028-3932(95)00050-d.

[13] S. C. Florman, The Existential Pleasures of Engineering. St. Martin’s Griffin, 1976.

[14] G. Tononi, and G. M. Edelman, A Universe of Consciousness: How Matter Becomes Imagination. Basic Books, 2000. Accessed: Sep. 20, 2021. [Online]. Available: https://philpapers.org/rec/COWAUO

[15] B. H. Turner and M. E. Knapp, “Consciousness: a neurobiological approach,” Integr. Physiol. Behav. Sci. Off. J. Pavlov. Soc., vol. 30, no. 2, pp. 151–156, Jun. 1995, doi: 10.1007/BF02691683.

[16] R. Kurzweil, The Singularity is Near: When Humans Transcend Biology - PhilPapers. 2005. Accessed: Sep. 16, 2021. [Online]. Available: https://philpapers.org/rec/KURTSI

[17] W. T. Rockwell, Neither Brain nor Ghost: A Nondualist Alternative to the Mind-Brain Identity Theory. Cambridge, MA, USA: A Bradford Book, 2005.

[18] G. Buttazzo and R. Manzotti, “Artificial consciousness: theoretical and practical issues,” Artif. Intell. Med., vol. 44, no. 2, pp. 79–82, Oct. 2008, doi: 10.1016/j.artmed.2008.08.001.

[19] D. Gamez and O. Holland, “Artificial Intelligence and Consciousness☆,” in Reference Module in Neuroscience and Biobehavioral Psychology, Elsevier, 2017. doi: 10.1016/B978-0-12-809324-5.05918-6.

[20] G. Buttazzo, “Artificial consciousness: hazardous questions (and answers),” Artif. Intell. Med., vol. 44, no. 2, pp. 139–146, Oct. 2008, doi: 10.1016/j.artmed.2008.07.004.

[21] “IBM100 - Deep Blue,” Mar. 07, 2012. http://www-03.ibm.com/ibm/history/ibm100/us/en/icons/deepblue/ (accessed Sep. 17, 2021).

[22] J. R. Searle, “Minds, brains, and programs,” Behav. Brain Sci., vol. 3, no. 3, pp. 417–424, Sep. 1980, doi: 10.1017/S0140525X00005756.

[23] I. Aleksander, Impossible Minds: My Neurons, My Consciousness. 2015. doi: 10.1142/9781783265701_fmatter