Understand What is Consciousness for Developing System 2 Artificial Intelligence

Generated by ChatGPT
Prompted (instructed) by Edward Y. Chang
December 27th, 2022

Abstract

System 1 AI, also known as narrow AI, is designed to perform specific, pre-defined tasks. These tasks are typically well-defined and can be executed efficiently by a machine learning algorithm. Examples of system 1 AI include image recognition algorithms and language translation systems. While these types of AI can be very effective at the narrow tasks they are designed for, they are limited in their ability to perform more complex tasks such as reasoning, planning, and decision-making.

To address these limitations, researchers have proposed developing system 2 AI, which is inspired by the human ability to perform more complex cognitive functions such as understanding and interpreting the meaning of data and making more informed decisions. The relationship between system 1 and system 2 AI is similar to the relationship between human unconsciousness and consciousness. While system 1 AI can be thought of as being analogous to the unconscious mind, which performs automatic, reflexive actions, system 2 AI aims to replicate some of the capabilities of human consciousness, which enables us to engage in more complex thought processes.

This article discusses the concept of consciousness and how it has been approached by researchers in various fields including psychology, philosophy, physics, biology, neuroscience, and computer science. Through the use of an interactive process with ChatGPT, the article is structured in a logical and rigorous manner, starting with general concepts and progressing to more specific ideas. In each section, key ideas are elaborated upon with references, and summaries of these references are provided. Finally, the article includes comments and suggestions from the "prompter" (or the advisor), and a final pass of revision was performed by ChatGPT, mostly focusing on cosmetic enhancements rather than adding new content.

Fig.1 Consciousness of Panpsychism, by DALLE

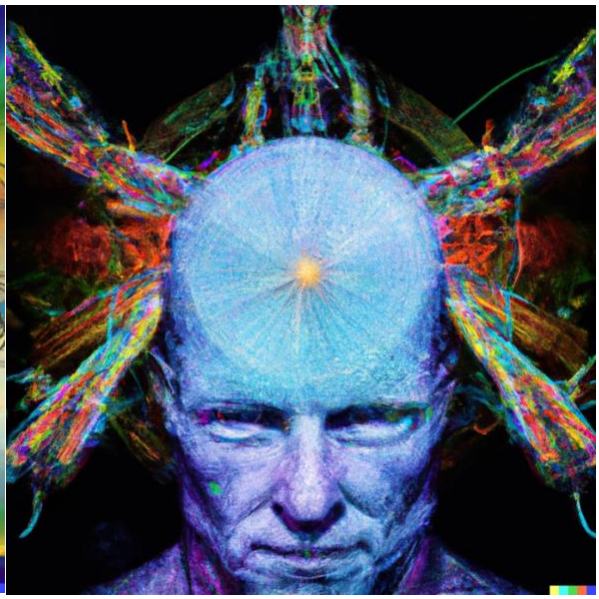

Fig.2 Consciousness of Functionalism, by DALLE

Consciousness is the state of being aware of one's own thoughts, feelings, and surroundings [Section1]. It is a complex and multifaceted concept that has been studied by philosophers, scientists, and theologians for centuries. The precise nature of consciousness and how it arises from the brain and other biological systems [Section2] is still not fully understood and is a topic of active research and debate. Some theories propose that consciousness is a fundamental aspect of the universe (panpsychism) [Section3], while others suggest that it emerges from complex computations in the brain (functionalism) [Section4]. Despite significant progress in the scientific study of consciousness, many mysteries and questions remain [Section5].

One reason why we study, and model consciousness is to improve the capabilities of artificial intelligence (AI). Currently, AI known as "system 1" is limited in its flexibility and adaptability. It can only perform narrow tasks quickly and efficiently, but it lacks the ability to adapt to changing environments and unexpected situations. This can limit its usefulness in various contexts.

To address these limitations, researchers such as Yoshua Bengio have proposed developing "system 2" AI (Bengio, 2020). In contrast to system 1 AI, system 2 AI is designed to perform more complex tasks such as reasoning, planning, and decision-making. It has the ability to understand and interpret the meaning of data, allowing it to make more informed decisions. This type of AI is better suited for tasks that require complex reasoning and decision-making abilities, such as natural language processing and autonomous decision-making.

As discussed by Daniel Kahneman in his theory of thinking (Kahneman, 2011), there are two systems of thought: System 1, which is fast and automatic, and System 2, which is slower and more deliberate. Kahneman's theory has been influential in the field of psychology and has shaped the study of decision-making and judgment. While system 1 AI excels at discriminative tasks, system 2 AI excels at generative tasks that require more complex reasoning and decision-making. By studying and modeling consciousness, we can better understand the cognitive processes involved in thinking and decision-making, which can inform the development of more advanced AI systems.

In order to develop a system that resembles human consciousness, or "system 2," we first need to identify our desired goals and functionalities [Section6]. While there have been various theories and hypotheses proposed by researchers in fields such as psychology, philosophy, and theology, these ideas are often abstract and difficult to concretely demonstrate (such as panpsychism vs. functionalism). Instead, we choose to follow Aristotle's first principle of basing our framework on scientific evidence from fields such as physics (Schrödinger, 1944), biology, neuroscience (Deisseroth, 2021), and computer science [Section7], rather than relying on more abstract and elusive ideas.

1. How are the mechanisms for human beings to aware of his/her own thoughts, feelings, and surroundings?

The mechanisms underlying human consciousness are not fully understood and are a topic of ongoing scientific research and debate. Some theories suggest that consciousness arises from complex computations in the brain, while others propose that it is a fundamental aspect of the universe.

There are several areas of the brain that are thought to be involved in conscious awareness, including the prefrontal cortex, the thalamus, and the reticular activating system. The prefrontal cortex is responsible for higher cognitive functions such as decision-making and problem-solving and is thought to play a key role in the integration of information from various brain regions. The thalamus is a key relay station for sensory information and is thought to play a role in filtering and integrating this information. The reticular activating system is a network of neurons that plays a role in maintaining wakefulness and alertness.

Other factors that may contribute to consciousness include the activity of neurotransmitters such as dopamine and serotonin, as well as the presence of certain brain waves. Some researchers have also suggested that consciousness may involve the interaction of multiple brain systems, including those involved in perception, attention, and memory.

Overall, the precise mechanisms underlying human consciousness are still not fully understood, and more research is needed to fully understand how we are able to be aware of our own thoughts, feelings, and surroundings.

2 How does consciousness arise from the brain and other biological systems?

The relationship between the brain and consciousness is a complex and multifaceted topic that has been studied by philosophers, scientists, and theologians for centuries. Despite significant progress in understanding the brain and its functions, the precise nature of consciousness and how it arises from the brain and other biological systems is still not fully understood.

One theory of consciousness is known as "panpsychism," which proposes that consciousness is a fundamental aspect of the universe and is present in all matter, including in the brain. According to this theory, the brain and other biological systems provide the necessary conditions for consciousness to emerge, but do not cause consciousness itself. Another theory is known as "functionalism," which suggests that consciousness arises from the function of the brain. According to this theory, certain computations or processes in the brain give rise to conscious experience. There are also many other

theories of consciousness, including the integrated information theory, which proposes that consciousness arises from the integration of information in the brain, and the global workspace theory, which suggests that consciousness arises from the global activation of neurons in the brain.

Overall, the relationship between the brain and consciousness is a complex and poorly understood topic, and more research is needed to fully understand how consciousness arises from the brain and other biological systems.

3 Consciousness is a fundamental aspect of the universe (Panpsychism)

The theory of panpsychism proposes that consciousness is a fundamental aspect of the universe, and is present in all matter, including in the brain. According to this theory, consciousness is not something that emerges from the brain or other biological systems but is instead a fundamental property of the universe that exists independently of these systems.

According to panpsychism, the brain and other biological systems provide the necessary conditions for consciousness to emerge, but do not cause consciousness itself. Instead, consciousness is seen as a fundamental aspect of the universe that is present in all matter, and the brain acts as a sort of "receiver" or "amplifier" of this consciousness.

Panpsychism has been proposed by several philosophers and scientists, including David Chalmers, Galen Strawson, and Thomas Nagel. However, the theory remains controversial and is not widely accepted within the scientific community.
In recent years it has gained some support as a potential way to explain the relationship between the brain and consciousness, and to address some of the challenges posed by other theories of consciousness.

4 Consciousness emerges from complex computations in the brain (Functionalism)

Functionalism is a theory of consciousness that proposes that consciousness arises from the function of the brain. According to this theory, certain computations or processes in the brain give rise to conscious experience.

One of the main arguments for functionalism is that it can help to explain the relationship between brain function and conscious experience. For example, brain injuries or damage to specific brain regions can lead to changes in consciousness, such as coma or altered states of awareness. Similarly, certain brain disorders, such as epilepsy or schizophrenia, can affect conscious experience in specific ways. These observations suggest that the brain plays a central role in shaping conscious experience.

Functionalism is a broad theory that encompasses a range of specific proposals about the nature of consciousness and how it arises from brain function. For example, some researchers have proposed that consciousness arises from the integration of information in the brain, while others have suggested that it arises from the global activation of neurons in the brain.

One of the main challenges for functionalism is explaining how subjective experiences, such as the experience of seeing a red apple or feeling pain, arise from the physical processes of the brain. This is often referred to as the "hard problem" of consciousness and is a topic of ongoing debate and research.

5 Outstanding questions about consciousness

What is consciousness? Despite much research and many theories, we still do not have a clear and agreed-upon definition of consciousness. Some researchers view it as a fundamental aspect of the universe, while others see it as an emergent property of complex systems like the brain (Chalmers, 1996).

How does consciousness arise from the brain? Even though the brain is made up of physical matter, it is not yet clear how this matter gives rise to subjective experiences like pain, pleasure, and perceptions of the world around us. Some theories posit that consciousness arises from the interactions of neurons in the brain (Baars, 1988; Crick & Koch, 1990), while others suggest that it may be the result of quantum processes (Penrose & Hameroff, 1996).

Is consciousness unique to humans? Many researchers believe that human consciousness is unique (Searle, 1992; Dennett, 1991), but there is still much debate about whether other animals and even machines might be capable of experiencing consciousness in some form (Griffiths, 2002; Sandberg & Bostrom, 2008).

What is the relationship between consciousness and the brain? While the brain and consciousness are closely linked (Kandel et al., 2000), it is not yet clear exactly how this relationship works. Some theories suggest that consciousness is a product of brain activity (Dehaene et al., 2003), while others propose that it may be a fundamental aspect of the universe that cannot be reduced to brain activity (Chalmers, 1996).

What is the nature of subjective experience? One of the most mysterious aspects of consciousness is the fact that it is subjective – each person's experience of the world is unique and cannot be directly observed by others (Nagel, 1974). This raises questions about the nature of subjective experience and how it relates to objective reality (Searle, 1992).

6 Functionalities of Human Consciousness

Human consciousness is a complex and multifaceted phenomenon that is still not fully understood. However, there are several key functionalities that are generally associated with consciousness. These include:

- Awareness: The ability to be aware of one's surroundings, thoughts, and feelings (Baars, 2002).

  In "A Cognitive Theory of Consciousness," Bernard Baars presents a theory of consciousness that is based on the idea that consciousness is a global cognitive process that integrates information from various sources and allows an organism to interact with its environment. Baars' theory is centered around the concept of a "global workspace," which is a hypothetical system in the brain that allows information from various sources to be integrated and made available to other cognitive processes. According to Baars, consciousness arises when information is broadcasted to the global workspace and becomes available to other cognitive processes, allowing the organism to act on it.

Baars' theory also incorporates the idea of "attention," which he defines as the process of selecting and focusing on certain pieces of information while ignoring others. According to Baars, attention is a key component of consciousness because it allows the organism to prioritize and act on certain pieces of information while ignoring others.

Overall, Baars' theory of consciousness is a comprehensive and influential approach that has had a significant impact on the field of psychology and cognitive science. It has been widely cited and has influenced the development of other theories of consciousness.

Researchers have attempted to model the global workspace in several ways, including through computational models, simulations, and brain imaging studies. One example of a computational model of the global workspace is the "blackboard architecture," which was developed by Baars and colleagues in the 1980s. The blackboard architecture is a computer simulation that aims to replicate the cognitive processes involved in the global workspace, including the integration of information from various sources and the ability to broadcast this information to other cognitive processes.

Other researchers have used brain imaging techniques, such as functional magnetic resonance imaging (fMRI), to study the neural basis of the global workspace and to test predictions of Baars' theory. For example, some studies have found that certain brain regions, such as the prefrontal cortex and the posterior parietal cortex, are more active when an individual is engaged in tasks that involve conscious processing, which is consistent with Baars' theory.

Overall, while Baars' theory of consciousness has not been fully validated, it has inspired a significant amount of research and has influenced the development of other theories of consciousness.

- Attention: The ability to focus on specific stimuli or tasks, and to filter out distractions (Posner & Petersen, 1990).

In "The Attention System of the Human Brain," Michael I. Posner and Steven E. Petersen present a review of the current state (in 1990) of knowledge about the attention system of the human brain. Attention is a cognitive process that involves the selection and focusing of information from the environment, and it is essential for many aspects of human cognition, including perception, learning, and memory.

Posner and Petersen propose a model of the attention system that consists of three interacting components: the alerting system, the orienting system, and the executive system. The alerting system is responsible for maintaining an overall state of alertness and arousal, while the orienting system is responsible for directing attention to specific stimuli in the environment. The executive system is responsible for controlling the allocation of attention and for coordinating the activity of the other two systems.

Posner and Petersen review evidence from a range of sources, including behavioral studies, brain imaging studies, and studies of brain-damaged patients, to support their model of the attention system. They also discuss the implications of their model for understanding various attention-related disorders, such as attention deficit hyperactivity disorder (ADHD) and stroke.

Overall, "The Attention System of the Human Brain" is a comprehensive and influential review of the attention system that has had a significant impact on the field of cognitive neuroscience. It has been widely cited and has helped to shape the development of subsequent research on attention.

- Perception: The process of interpreting sensory information and forming mental representations of the environment (Gregory, 1997).

  This area is already adequately supported by system 1 AI, or unconsciousness. The only element that a model ought to consider is how the transitions between unconscious background perception and consciousness awareness is performed. Schrödinger (Schrödinger, 1944) provides the mechanisms in physics to implement the transitions.

- Thinking: The ability to process and manipulate information, solve problems, and make decisions (Newell & Simon, 1972).

  Thinking is a complex cognitive process that involves the generation and manipulation of ideas, concepts, and information in the mind. It is a fundamental aspect of human cognition and is central to many higher-level cognitive functions, such as problem-solving, decision-making, and planning.

  In their book "Human Problem Solving," Allan Newell and Herbert Simon proposed a theoretical framework for understanding the process of thinking and problem-solving. They argued that problem-solving involves the search for and generation of new knowledge, and that this process can be broken down into several distinct stages:

  · Formulation: This stage involves defining the problem and understanding its context and constraints.
  · Search: During this stage, the problem-solver generates and considers potential solutions to the problem.
  · Evaluation: In this stage, the problem-solver evaluates the potential solutions and selects the one that is most likely to be successful.
  · Execution: This stage involves implementing the chosen solution and verifying that it has solved the problem.

- Free will and Intentionality: The ability to choose goals and to act to achieve those goals (Dennett, 1987).

  Free will and intentionality are two distinct philosophical concepts that are often discussed together in the context of human behavior and decision-making. Free will refers to the idea that individuals can make choices and decisions that are not predetermined or controlled by outside forces. This means that people can act freely and are not simply puppets or automatons whose actions are determined by external factors. Intentionality, on the other hand, refers to the property of being directed towards an object or goal. It is the mental state of being aware of and intending to do something. In other words, intentionality is the quality of having a purpose or goal in mind when you do something.

  Regarding implementation, free will can be formulated by the values of choices and the entropy among choices (Rehn, 2022). This means that an individual's free will is represented by the various

options they have to choose from and the inherent uncertainty or randomness in their decision-making process.

Intentionality is the capacity of an individual or system to have mental states, such as beliefs, desires, and intentions, that are directed towards objects or states of affairs in the external world. It is a framework for understanding and predicting the behavior of complex systems, including other people and the environment. Dennett (1987) has written extensively about intentionality and its role in shaping human behavior.

- Emotion: The experience of feelings and emotional states, and the ability to express and respond to emotions (Damasio, 1994).

  In "Descartes' Error," Antonio Damasio argues that emotions are essential for guiding human decision-making and reasoning, and that they should not be excluded from the process of rational thought. He also believes that emotions are closely tied to our sense of self and influence our memories, social relationships, and perception of the world. Damasio challenges the traditional view that emotions are irrational and should be suppressed. Nevertheless, for our purpose, we ignore emotion in our modeling effort.

These functionalities are not necessarily distinct from one another, and they often overlap and interact in complex ways. Additionally, there may be other functionalities or characteristics of consciousness that are not captured by this list. Understanding the functionalities of human consciousness is a complex and ongoing scientific challenge.

7 Computational Models of Consciousness

There are several computational models of consciousness that have been proposed in the literature. These models attempt to explain how consciousness arises from the activity of the brain, and how it might be simulated in artificial systems such as computers.

One well-known model is the Global Workspace Theory (GWT), proposed by Bernard Baars (1988). According to the GWT, consciousness arises from the interaction of various brain systems, including sensory systems, attentional systems, and memory systems. The GWT posits that consciousness arises when information from these various systems is integrated and made available to other systems in the brain, creating a global workspace that can be accessed by other systems.

Another computational model of consciousness is the Integrated Information Theory (IIT), proposed by Giulio Tononi (2004). According to the IIT, consciousness arises from the integration of information across multiple sources and is characterized by the ability to distinguish between different states of the system. The IIT proposes that the amount of consciousness in a system can be quantified by a measure called phi, which reflects the degree of integration of information in the system.

Other computational models of consciousness include the Dynamic Core Hypothesis (DC), proposed by Gerald Edelman and Giulio Tononi (2000), and the Neural Correlates of Consciousness (NCC) approach, which focuses on identifying the specific brain regions and neural processes that are associated with consciousness (Crick & Koch, 1990).

8. Conclusion

This article explores the concept of consciousness and its various functions, as well as preliminary ideas for modeling it. It also examines previous research on consciousness and its potential to be simulated in artificial systems. The use of optogenetics, a technique developed by Professor Deisseroth at Stanford, has made it possible to stimulate individual neurons and observe the resulting signal propagation and motor responses. Further research will involve analyzing collected data to gain a deeper understanding of the mechanisms underlying consciousness. For relevant lectures, please visit Stanford CS372 homepage for slides and videos.

References

1. Baars, B. J. (1988). A cognitive theory of consciousness. Cambridge University Press.
2. Baars, B. J. (2002). The conscious access hypothesis: Origins and recent evidence. Trends in Cognitive Sciences, 6(1), 47-52.
3. Bengio, Y. (2020). The Future of AI: Opportunities and Challenges. Nature, 579(7798), 479-482.
4. Chang E. Y. (2020-22). Consciousness modeling lecture series, CS372, Stanford University, 2020-22.
5. Chalmers, D. (1995). Facing up to the problem of consciousness. Journal of Consciousness Studies, 2(3), 200-219.
6. Chalmers, D. J. (1996). The conscious mind: In search of a fundamental theory. Oxford University Press.
7. Crick, F., & Koch, C. (1990). Towards a neurobiological theory of consciousness. Seminars in the Neurosciences, 2(2), 263-275.
8. Crick, F., & Koch, C. (2003). The neural correlates of consciousness. Nature Neuroscience, 6(2), 119-126.
9. Damasio, A. R. (1994). Descartes' error: Emotion, reason, and the human brain. New York, NY: Putnam.
10. Dehaene, S., Sergent, C., & Changeux, J. P. (2003). A neural network model of the basal ganglia's role in saccade initiation. Nature Neuroscience, 6(5), 450-459.
11. Deisseroth, Karl. (2021). Projections: The Future of the Brain. Penguin Press, 2021.
12. Dennett, D. C. (1987). The intentional stance. Cambridge, MA: MIT Press.
13. Dennett, D. C. (1991). Consciousness explained. Little, Brown and Company.
14. Edelman, G. M., & Tononi, G. (2000). A universe of consciousness: How matter becomes imagination. Basic Books.
15. Gregory, R. L. (1997). Eye and brain: The psychology of seeing (5th ed.). New York, NY: Oxford University Press.
16. Griffiths, P. E. (2002). What is consciousness? In A. P. Shimony, S. Savage, & J. D. Koehler (Eds.), Perception, causation, and objectivity (pp. 91-118). Oxford University Press.
17. Kahneman, D. (2011). Thinking, Fast and Slow, Daniel Kahneman, Farrar, Straus and Giroux.
18. Kandel, E. R., Schwartz, J. H., Jessell, T. M., Siegelbaum, S. A., & Hudspeth, A. J. (2000). Principles of neural science (4th ed.). McGraw-Hill.
19. Nagel, T. (1974). What is it like to be a bat? The Philosophical Review, 83(4), 435-450.
20. Nagel, T. (2012). Mind and cosmos: Why the materialist neo-Darwinian conception of nature is almost certainly false. Oxford University Press.
21. Newell, A., & Simon, H. A. (1972). Human problem solving. Englewood Cliffs, NJ: Prentice-Hall.
22. Penrose, R., & Hameroff, S. (1996). Orchestrated reduction of quantum coherence in brain microtubules: A model for conscious events. Journal of Consciousness Studies.
23. Posner, M. I., & Petersen, S. E. (1990). The attention system of the human brain. Annual Review of Neuroscience, 13, 25-42.

24. Schrödinger, Erwin. (1944). What is Life? The Physical Aspect of the Living Cell. Cambridge University Press, 1944.
25. Strawson, G. (2006). Realistic monism: Why physicalism entails panpsychism. Journal of Consciousness Studies, 13(10-11), 3-31.
26. Rehn E. M. (2022), Free Will Belief as a Consequence of Model-based Reinforcement Learning, arXiv:2111.08435v2
27. Tononi, G. (2004). An information integration theory of consciousness. BMC Neuroscience, 5(1), 42.

Appendix. Further Readings

- "Consciousness Explained" by Daniel Dennett is a classic work that offers a comprehensive overview of the various theories of consciousness and the ways in which they have evolved over time.
- "The Hard Problem of Consciousness" by David Chalmers is a seminal paper that introduces the concept of the "hard problem" of consciousness, which refers to the difficulty in explaining how subjective experiences arise from the physical processes of the brain.
- "The Neural Correlates of Consciousness" by Francis Crick and Christof Koch is a review article that discusses the various brain regions and processes that are thought to be involved in conscious awareness.
- "Toward a Science of Consciousness" is a conference held annually in Tucson, Arizona, which brings together researchers from a variety of disciplines to discuss the latest research on consciousness. The conference website provides access to many papers and presentations on the subject.