Journal of Data Science 22 (1), 1–9
January 2024
DOI: 10.6339/24-JDS1117
Philosophies of Data Science

Decoding Consciousness in Artificial Intelligence

Momiao Xiong[1,2,*]

[1] *Department of Biostatistics and Data Science, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA*
[2] *Society of Artificial Intelligence Research, Houston, TX 77030, USA*

Abstract

The exploration of whether artificial intelligence (AI) can evolve to possess consciousness is an intensely debated and researched topic within the fields of philosophy, neuroscience, and artificial intelligence. Understanding this complex phenomenon hinges on integrating two complementary perspectives of consciousness: the objective and the subjective. Objective perspectives involve quantifiable measures and observable phenomena, offering a more scientific and empirical approach. This includes the use of neuroimaging technologies such as electrocorticography (ECoG), EEG, and fMRI to study brain activities and patterns. These methods allow for the mapping and understanding of neural representations related to language, visual, acoustic, emotional, and semantic information. However, the objective approach may miss the nuances of personal experience and introspection. On the other hand, subjective perspectives focus on personal experiences, thoughts, and feelings. This introspective view provides insights into the individual nature of consciousness, which cannot be directly measured or observed by others. Yet, the subjective approach is often criticized for its lack of empirical evidence and its reliance on personal interpretation, which may not be universally applicable or reliable. Integrating these two perspectives is essential for a comprehensive understanding of consciousness. By combining objective measures with subjective reports, we can develop a more holistic understanding of the mind.

Keywords *artificial intelligence; artificial general intelligence; consciousness; mortal computation*

Consciousness in Artificial Intelligence

The question of whether AI can be evolved into having consciousness remains a hot topic of ongoing debate and research in philosophy, neuroscience, and artificial intelligence (Witzel, 2023; Lenharo, 2023a,b). There is no consensus on the theory of consciousness (Fleming, 2023). Since we cannot directly observe consciousness by looking inside someone's brain, it is difficult to define and study empirically (Goff, 2023). Additionally, consciousness defies reductionist approaches, which means that it cannot be fully understood by breaking it down into simpler, constituent parts alone. Instead, it requires a more holistic view that considers the complex interplay of various elements and processes.

On August 17, 2023, Butlin et al. (2023) released a long preprint "Consciousness in Artificial Intelligence: Insights from the Science of Consciousness". The authors stated that their methods for studying consciousness in AI are based on three principles:

[*] Email: momiao.xiong@gmail.com.

Received December 5, 2023; Accepted January 16, 2024

- computational functionalism,
- neuroscientific theories of consciousness, and
- a theory-heavy approach.

Computational functionalism is a theory in the philosophy of mind and cognitive science which posits that mental states are constituted solely by their functional role — that is, they are causal relations to other mental states, sensory inputs, and behavioral outputs. This theory is analogous to the way a computer works; it is not the material substance of the components which defines them but rather how they function in relation to the system as a whole.

Computational functionalism of consciousness is a request for the set of properties of systems with which consciousness is associated. Both philosophy of the mind and cognitive science view the theoretical "framework of the mind" as a computational system. This perspective suggests that mental states and processes can be understood and explained in terms of computational operations, similar to the way a computer processes information.

Functionalism of consciousness suggests that it has a certain functional organization. In other words, we can assert a certain range of claims, which stand in certain causal relations to each other and to the environment.

The second principle is neuroscientific theories of consciousness. Its goal is to identify functions that are necessary and sufficient for consciousness in humans. Computational functionalism implies that cognitive processes and states can be understood in terms of abstract computational operations and algorithms, independent of the physical medium in which they are instantiated. This means that mental states are akin to software processes, which can theoretically be realized in multiple hardware systems, whether biological brains or artificial machines.

The third principle is a theory-heavy approach. The theories of consciousness can help us to understand consciousness and assess whether the AI has consciousness or not. The theory-heavy approach is most suitable for investigating consciousness in AI.

Five Types of Theories of Consciousness

Theories of consciousness include

- Integrated Information Theory (IIT) (Albantakis et al., 2022),
- Global Workspace Theory (GWT) (Mashour et al., 2022),
- Higher-order theories (Michel and Lau, 2021),
- Attention Schema Theory (Wilterson et al., 2020), and
- Recurrent Processing Theory (RPT) (Negro, 2023), which are compatible with their computational functionalism framework.

IIT is a theoretical framework for understanding consciousness, proposed by neuroscientist Giulio Tononi. In contrast to certain theories that concentrate on particular brain structures (such as the hippocampus) or specific brain functions (such as memory processing), IIT endeavors to establish a foundational and measurable framework for understanding the essence of consciousness in a system.

GWT is a cognitive neuroscience theory proposed by psychologist Bernard Baars. It proposes a framework for understanding how consciousness arises in the brain, in which the brain operates like a theater. GWT likens the brain to a global workspace where information is broadcasted and made available for a wide range of cognitive processes.

Higher-order theories of consciousness are a class of theories that propose that consciousness is not solely a matter of having mental states but involves having higher-order thoughts or perceptions about those mental states. These theories suggest that what makes a mental state conscious is the subject's awareness or representation of that mental state. In other words, consciousness arises from the capacity to reflect on or be aware of one's own mental experiences.

Attention Schema Theory (AST) is a cognitive neuroscience theory proposed by neuroscientist and philosopher Michael Graziano. The theory seeks to explain consciousness by focusing on the role of attention and the brain's ability to attribute attention to certain information.

Finally, RPT suggests that the unconscious functions, such as feature extraction and categorization of basic sensory information, are mediated by the feedforward sweep. This initial processing stage rapidly transmits sensory data from lower to higher neural regions. In contrast, conscious functions, particularly those involving the complex integration and organization of perceptual information, are facilitated by recurrent (feedback) cortico-cortical connections. These connections, which link higher and lower processing areas, allow for a more detailed and contextually informed interpretation of sensory input.

Easy Problem and the Hard Problem

There is no consensus definition of consciousness. In general, consciousness involves sensory events (vision, hearing, taste, touch, and smell), and psychological and neurological (thoughts, emotions, desires, beliefs) processes. It consists of both phenomenal experiences and subjective experiences, which are not separated. Like phenomenal experience, subject experience is also an essential component of consciousness.

The theory of "neural correlate of consciousness assumes that consciousness is neurological processes. For years, Christof Koch, a renowned neuroscientist known for his significant contributions to neural and cognitive science, collaborated with Francis Crick, a biologist who co-won the Nobel Prize for discovering the structure of DNA. Together, they developed the theory of the 'neural correlate of consciousness' (NCC). This theory posits that every conscious experience is associated with specific neuronal activity, which is not only essential but also fundamentally measurable. In essence, their collaborative work aimed to identify the precise brain mechanisms underlying conscious experiences, marking a pivotal advancement in the understanding of the brain's role in consciousness.

David Chalmers (Chalmers, 2023), an influential philosopher and cognitive scientist, renowned for his work in the area of philosophy of mind, particularly in the study of consciousness, rebutted that he did not believe that science would be able to measure the association between neural correlates in the brain and the subjective experience of consciousness. He defined this as "the hard problem of consciousness". Chalmers made a distinction between the "easy problems" of consciousness, which involve explaining cognitive functions and behaviors in terms of brain processes, and the "hard problem," which pertains to understanding the subjective, qualitative nature of conscious experience itself. The hard problem raises questions about why certain patterns of neural activity give rise to specific subjective feelings or qualia. Chalmers' assertion that the hard problem is "sufficiently challenging to clearly explain consciousness for at least a quarter of a century" reflects the ongoing complexity and depth of the philosophical and scientific inquiry into consciousness.

In 1998, at the conference of the Association for the Scientific Study of Consciousness (ASSC), Koch made a bet with his friend Chalmers: that scientists would find an NCC within 25 years. That is, Koch, a neuroscientist, was betting that within a quarter of a century, researchers would identify the precise brain mechanisms responsible for consciousness. This would entail not just understanding which brain areas are active during conscious states, but also elucidating how this activity leads to the subjective experience of being conscious. The bet reflects a significant challenge in neuroscience and philosophy, as finding the NCC would represent a major advancement in our understanding of the brain and the nature of consciousness. On June 30, 2023, At the 26th ASSC conference, 25 years after the initial bet, they announced that the results: the two conscious theories, IIT and GWT, were still in the conflict. Koch had lost the bet. Unfortunately, despite years of scientific effort, scientists still do not know how or why the experience of consciousness arises. We still do not have fundamental theory on the science of consciousness (Jarow, 2023; Zimmer, 2023).

This problem comes from the flaw of the scientific research method proposed by Galileo (Ellia et al., 2021). Galileo considered subjective properties to be outside the scope of science, and there for deliberately overlooked subjective properties, relegating scientific research to objectively explaining objective. In contrast, it is consistent with Chalmer's views to argue that correct scientific methods should aim to objectively explain both objective and subjective properties. Chalmers' view implies that solving the hard problem requires a profound shift in our understanding of consciousness, and acknowledges that this challenge may persist for an extended period, emphasizing the depth of the mystery surrounding the nature of subjective experience. The quest to unravel the hard problem remains a dynamic and interdisciplinary endeavor involving contributions from philosophy, neuroscience, psychology, and other fields.

Objective Point of View

The current research in consciousness uses behavioral observations and carefully designed stimulus–response tests to assess the presence or absence of consciousness, taking everything in between as a black box. Subjective experience or the way things feel, is excluded as outside the scope of objective scientific explanation. To make a black box explainable, scientists have begun to open the black box and replace it with internal states and operations, which are defined as cognitive functions. To reveal neural mechanisms of consciousness, functional Magnetic Resonance Imaging (fMRI), electroencephalogram (EEG), and other neural signals are used to measure the neural activity and search for NCC. To objectively study consciousness through NCC, we need to investigate content-specific NCC by a combination of recording, stimulation, and Lesion data (information and observations derived from studying the effects of lesions, or damage, to specific areas of the brain). This type of data is crucial in the field of neuroscience, particularly in understanding brain function and the relationship between specific brain regions and various cognitive abilities or behaviors. Lesions can be due to various causes, such as strokes, traumatic brain injuries, surgical removals, tumors, or degenerative diseases, revealing which neural mechanisms contribute to experience (and how) to identify the full NCC by comparing brain conditions in which consciousness as a whole is present. In other words, scientists try to collect enough empirical evidence to uncover objectively which brain areas and neural activities are involved in consciousness and its contents.

Advances in neural science research open the doors to use neuroimaging techniques such as fMRI, EEG, and deep learning for decoding the contents of experience from distributed neural activities in the brain (Du et al., 2022; Yen et al., 2023). The emphasis on decoding the contents of experience implies an interest in understanding how specific mental states, perceptions, or thoughts are represented in the patterns of neural activity. The ability to decode the contents of experience from distributed neural activities opens avenues for a deeper understanding of cognition, perception, and consciousness.

The success in decoding the contents of consciousness from brain activity patterns encourages some neuroscientists to pursue NCC research with the hope to objectively investigate the subjective properties of experience. This cognitive paradigm takes the external as an objective existence and overlooks internal (subjective) consciousness (Chen and Chen, 2022). This paradigm considers attention, working memory, and decision-making as functions and only values functions since they can be studied objectively by independent observers following the Galilean notion of science (Ellia et al. 2021).

In general, the objective point of view has the following strengths:

(1) Empirical Rigor: The objective approach relies on empirical methods, such as neuroscientific measurements and behavioral experiments, providing a high degree of empirical rigor and replicability.

(2) Scientific Advancements: Advances in neuroimaging technologies and experimental designs have allowed researchers to uncover correlations between brain activity and conscious states, contributing valuable insights.

However, the objective point of view also has the following limitations:

(1) Reductionism Concerns: Some objective approaches may be criticized for adopting reductionist stances, focusing on isolated neural correlates while potentially overlooking the holistic nature of consciousness.

(2) Subjective Gap: Objective methods often face challenges in capturing the richness and diversity of subjective experiences, leading to what is known as the "hard problem" of consciousness.

Subjective Point of View

From a subjective point of view, consciousness is experienced directly and uniquely by each individual. This perspective is inherently private, encompassing personal experiences, thoughts, sensations, emotions, and perceptions, and remains inaccessible and incomprehensible to others. However, when analyzing consciousness, certain cognitive paradigms often commit essential mistakes. The first is inverting the epistemic order, meaning they prioritize external, objective measurements over the internal, subjective experience. The second mistake is misplacing objectivity, where the inherently subjective nature of consciousness is overlooked in favor of a purely objective viewpoint. These errors highlight the challenges in bridging subjective experiences with objective analysis in the study of consciousness.

The epistemic order is that first we are conscious, then we can usually engage in various cognitive functions such as attending to contents, memorizing, and manipulating them in memory, and finally we can translate some contents into words. Conscious is first, whereas reporting our experiences is second. The Galilean notion of consciousness changes this epistemic order. It proposes that experience and function are first, and consciousness is second.

From a scientific perspective, it is crucial to correctly identify where objectivity is applicable and where it is not, i.e., where the things need to be explained. If we misplace objectivity and replace experience with functions, we will only explain those functions and not explain experience (Ellia et al. 2021). To correctly place objectivity, Ellia et al. (2021) argue that the intrinsic, phenomenal structure of an experience should have a physical explanation and take a cause–effect structure as the physical explanation. However, cause-effect structure does not consider all possible confounders and there is a lack of methods to distinguish causation with association (Negro, 2023). The essential problem is the current theories of consciousness are systematically underdetermined. Different models with equal explanatory and predictive power can fit the same data. We do not have a well-defined methodology for deciding which model of consciousness fits best with currently available data.

This viewpoint asserts that subjective experience, unique to humans, transcends mere information processing or neuronal activity and cannot be replicated by machines. The key components of subjective experience theory are as follows:

(1) Uniqueness of Human Subjective Experience:
The assertion that subjective experience is "owned by humans" emphasizes the belief in the exclusivity of conscious, subjective awareness to the human species. This aligns with a philosophical perspective that consciousness is a distinctive feature of human existence.

(2) Irreducibility to Information Processing or Neuronal Behavior:
This component contends that subjective experience cannot be transformed or reduced to the mere processing of information or the behavior of neurons. This position is in line with the idea that consciousness involves more than just the functional aspects of information processing in the brain or the activities of individual neurons.

(3) Incapability of Machine Generation:
By stating that subjective experience cannot be generated by machines, the perspective implies a skepticism about the ability of artificial intelligence or computational systems to truly replicate or emulate human-like consciousness. This skepticism might be rooted in the belief that machines lack certain essential qualities or aspects that contribute to subjective experience.

This perspective reflects a position often associated with certain philosophical viewpoints, such as dualism or property dualism, which posit that consciousness is not reducible to physical processes. It also aligns with arguments against the possibility of achieving true artificial consciousness in machines.

It's important to note that there are alternative perspectives in philosophy and cognitive science, such as functionalism which we have discussed or panpsychism, which propose different ways of understanding consciousness and the potential for its emergence in non-biological systems.

Subjectivist views underscore the enduring mystery and complexity surrounding subjective experience, as well as the challenges and debates inherent in attempts to characterize, explain, or replicate it, particularly in the realm of artificial intelligence and machine consciousness. Subjective experience is a fundamental aspect of the human universe that cannot be fully explained or simulated by scientific or mathematical models. The difficulty of explaining subjective experience is the hardest problem of consciousness (Gülen, 2023).

In general, the subjective point of view has the following strengths:

(1) Richness of Experience: The subjective viewpoint is directly concerned with the qualitative, first-person aspects of consciousness, acknowledging the rich and diverse nature of individual experiences.

(2) Personal Insight: Subjective viewpoint maintains that subjective reports and introspection can provide unique insights into the nuances of consciousness that may not be accessible through external measurements.

Limitations of the subjective point of view include:

(1) Subjective Variability: Individual experiences can vary widely, making it challenging to establish generalizable principles or create a unified theory of consciousness.

(2) Reliability Issues: Subjective reports are subject to biases, memory distortions, and other cognitive limitations, raising questions about the reliability of introspective data.

Note how these limitations are inherent to the subjectivity of consciousness itself.

Conclusions

In November of 2023, Google DeepMind proposed "Levels of AGI" based on depth (performance) and breadth (generality) of capabilities, and a framework for identifying the capabilities and behavior of Artificial General Intelligence (AGI) models (Morris et al., 2023). Less than 2 weeks later, Meta developed a benchmark for evaluating General AI Assistants (Mialon et al., 2023). The purpose of both papers was to objectively evaluate a set of fundamental abilities and performance of AGI systems. They did not attempt to test for the consciousness of AGI. From an objective perspective on consciousness, focusing on observable and measurable aspects rather than subjective experiences or personal introspection, Kleiner (2023) presented a significant finding on November 27, 2023. He demonstrated that if Computational Functionalism is valid, consciousness necessitates material computation. Consequently, this implies that none of the contemporary AI systems, as of that date, possess consciousness.

In summary, the question of whether artificial intelligence can have consciousness is a complex and unsolved problem. While some scientists and philosophers argue that AI may have subjective experience and consciousness in the future, others raise arguments against AI consciousness and believe that machines are fundamentally incapable of having these experiences. We expect that while AI technologies continue to rapidly develop, the debate over the possibility of AI consciousness will not stop.

Attaining a thorough understanding of consciousness typically necessitates the amalgamation of both objective and subjective viewpoints. Integrating objective metrics with subjective experiences can foster a more comprehensive grasp of the mind's workings. This interplay between objective data and subjective insights prompts profound philosophical inquiries about the essence of consciousness, the mind-body conundrum, and the inherent boundaries of scientific exploration.

To effectively bridge the gap between objective and subjective perspectives, interdisciplinary collaboration is essential, involving fields such as neuroscience, psychology, philosophy, and cognitive science. For a deeper insight into AI consciousness and to ensure its healthy evolution, it is crucial to integrate diverse computational algorithms and neural models. These models should be adept at deciphering neural representations of various types of information — including linguistic, visual, acoustic, emotional, and semantic — using data from non-invasive neuroimaging technologies like electrocorticography (ECoG), EEG, and fMRI. It's important to thoroughly examine the conscious implications of these technologies.

Acknowledgement

The author thank the editor, Glen Wright Colopy, for his helpful suggestions and edits to the manuscript.

References

Albantakis L, et al. (2022). Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms. arXiv preprint: https://arxiv.org/abs/2212.14787.

Butlin P, et al. (2023). Consciousness in artificial intelligence: Insights from the science of consciousness. arXiv preprint: https://arxiv.org/abs/2308.08708.

Chalmers DJ (2023). David J. Chalmers. *Neuron*, 111(21): 3341–3343. https://doi.org/10.1016/j.neuron.2023.10.018

Chen J, Chen L (2022). The hard problem of consciousness — A perspective from holistic philosophy. *Front. Neurosci.*, 25 October 2022. Sec. Perception Science, 16.

Du B, Cheng X, Duan Y, Ning H (2022). fMRI brain decoding and its applications in brain–computer interface: A survey. *Brain Sciences*, 12: 228. https://doi.org/10.3390/brainsci12020228

Ellia F, et al. (2021). Consciousness and the fallacy of misplaced objectivity. *Neuroscience of Consciousness*, 2021(2): niab032. 2021. https://doi.org/10.1093/nc/niab032

Fleming S (2023). The integrated information theory of consciousness as pseudoscience. https://osf.io/preprints/psyarxiv/zsr78/.

Goff P (2023). Understanding consciousness goes beyond exploring brain chemistry. https://www.scientificamerican.com/article/understanding-consciousness-goes-beyond-exploring-brain-chemistry/.

Gülen K (2023). Exploring the mind in the machine. While some argue that AI can be capable of subjective experience and consciousness, others believe that machines are fundamentally incapable of having these experiences. https://dataconomy.com/2023/03/23/can-artificial-intelligence-have-consciousness/.

Jarow O (June 30, 2023). Why scientists haven't cracked consciousness. The science of consciousness still has no theory. *Vox*. https://www.vox.com/future-perfect/2023/6/30/23778870/consciousness-brain-mind-hard-problem-neuroscience-koch-chalmers.

Kleiner J (2023). Consciousness requires mortal computation. https://philarchive.org/archive/JOHCRM.

Lenharo M (2023a). Consciousness theory slammed as 'pseudoscience' — sparking uproar. Researchers publicly call out theory that they say is not well supported by science, but that gets undue attention. *Nature NEWS*, 20 September 2023.

Lenharo M (2023b). Decades-long bet on consciousness ends — and it's philosopher 1, neuroscientist 0. Christof Koch wagered David Chalmers 25 years ago that researchers would learn how the brain achieves consciousness by now. But the quest continues. *Nature. NEWS*, 24 June 2023.

Mashour GA, Roelfsema P, Changeux JP, Dehaene S (2022). Conscious processing and the global neuronal workspace hypothesis. *Neuron*, 105(5): 776–798. https://doi.org/10.1016/j.neuron.2020.01.026

Mialon G, Fourrier C, Swift C, Wolf T, Scialom T LY (2023). GAIA: A benchmark for general AI assistants. arXiv preprint: https://arxiv.org/abs/2311.12983.

Michel M, Lau H (2021). Higher-order theories do just fine. *Cognitive Neuroscience*, 12(2): 77–78. https://doi.org/10.1080/17588928.2020.1839402

Morris MR, et al. Levels of AGI: Operationalizing progress on the path to AGI. arXiv preprint: https://arxiv.org/abs/2311.02462.

Negro N (2023). Can the integrated information theory explain consciousness from consciousness itself? *Review of Philosophy and Psychology*, 14: 1471–1489. https://doi.org/10.1007/s13164-022-00653-x

Wilterson AI, Kemper CM, Kim N, Webb TW, Reblando AMW, Graziano MSA (2020). Attention control and the attention schema theory of consciousness. *Progress in Neurobiology*, 195: 101844. https://doi.org/10.1016/j.pneurobio.2020.101844

Witzel MJ (2023). My debate with AI about it being conscious. https://www.linkedin.com/

pulse/my-debate-ai-being-conscious-mark-j-witzel/.

Yen C, Lin C-L, Chiang M-C (2023). Exploring the frontiers of neuroimaging: A review of recent advances in understanding brain functioning and disorders. *Life (Basel)*, 13: 1472.

Zimmer C (2023). Leading theories of consciousness square off. *The New York Times*. https://www.nytimes.com/2023/07/01/science/consciousness-theories.html?auth=login-google1tap&login=google1tap.