Connectionism's new wave. Reexamining two theories of machine consciousness

Alexandra CHIRILĂ
Doctoral School of Filosofie și Științe Social-Politice
"Alexandru Ioan Cuza University of Iași, Romania
chirila.alexandraa@gmail.com

Abstract

This paper analyses the possibility of machine consciousness concerning Daniel C. Dennett and John R. Searle, in the light of the most recent technological advancements. We will consider the opposing views that each author has of qualia and how it either widens or narrows the possibility of machine consciousness. While Dennett strongly believes that human-like artificial intelligence is just a matter of time, Searle thinks that until we solve the problem of how the mind works, the idea of an AI that does everything humans are capable of is far from reality. Dennett reinforces his beliefs about this kind of AI, based on the concept of strange inversion of reason. On the other hand, Searle states that consciousness can be causally reduced to neural activity, but not ontologically. Even though connectionism used to be popular two decades ago, recent developments in Deep Learning have made connectionism return to the spotlight, with software like AlphaZero, Chat GPT, and DALLE. Artificial neural networks are becoming more complex and nuanced, and this calls for a revisit of the classic arguments mentioned above to determine their relevance in light of the most recent developments. This paper aims to determine whether any of these views can still help researchers and engineers when considering state-of-the-art technologies and future developments in AI.

Keywords: *Connectionism, Dennett, Searle, machine consciousness, qualia.*

Introduction

This paper examines the new wave of connectionism in the field of artificial intelligence (AI) and its implications for two of the classic theories of machine consciousness: the theories of Daniel Dennett and John Searle. We start with an overview of connectionism, followed by its modern developments in the light of technological advancements, such as deep learning architectures. The second part of the paper concerns the relevance and compatibility of Dennett's and Searle's positions with regards to connectionism and AI. The goal is to assess whether these theories hold their place in current debates in the field of AI.

Early connectionism and Deep Learning

Connectionism was the dominant framework within the philosophy of mind during the 1980s. Connectionism theorists believed intellectual abilities could be developed in artificial agents using artificial neural networks. Connectionism represented the alternative to the classical theory of mind, which modeled cognition as symbolic computation.

Computational and technological power was scarce during that time, which means that scientists had no means to put their ideas into practice. Early connectionist models relied on artificial neural networks (ANN) composed of input units, hidden units, and output units. The input units received the information, the hidden units processed said information, and the output units gave a response. These units were inspired by their biological counterparts. The input units are analogous to sensory neurons, output units to motor neurons, while hidden units represent all other neurons. These initial artificial networks worked in a binary system, meaning that they could be only activated or deactivated, without any degree, like switching a light on and off without being able to change its intensity. Furthermore, these early networks were limited in depth, meaning the architecture only allowed for one or two layers of hidden units.

During the 1990s, the field of machine learning witnessed an exponential growth. As Tom Michael Mitchells explains, "the field of machine learning is concerned with the question of how to construct computer programs that automatically improve with experience" (Mitchell, 1997, xv). However, recent developments have enabled the emergence of deep learning, a class of machine learning. According to Deng and Yu, deep learning is "a class of machine learning techniques that exploit many layers of information processing for supervised or unsupervised feature extraction and transformation, and for pattern analysis and classification." (Deng & Yu, 2013, pp. 199-200). Deep learning represents connectionism's new wave for at least two reasons, as follows. Firstly, deep learning architectures overcome a significant limitation of earlier ANN, reaching a far greater number of hidden layers, ranging from five to hundreds of layers (Buckner & Garson, 1997). This increase allows for more complex and accurate information processing. Moreover, the great number of hidden layers translates to the possibility of creating a better copy of the biological neural network. Secondly, modern artificial systems no longer operate on a binary model. The newer ANN can represent different levels of activation, an important feature in convolutional neural networks (CNN), which concern visual image analysis. AlphaGo and its better version AlphaZero have demonstrated the power of deep networks. While

AlphaGo required little human input and data, AlphaZero achieved superhuman performance with no human input, training entirely through self-play, becoming the best go player in roughly 24 hours. Other examples of successful deep networks are DeepSpeare, which learned to write poems and sonnets after analyzing vast amounts of data and DALLE or MidJourney, generative CNNs capable of generating images based on text prompts. Deep Learning is of utmost importance for large companies. As Cameron Buckner stated, by 2016, Google had approximately one thousand deep learning projects in development (Buckner, 2019, p. 2).

Despite these advances, one problem remains, the anthropomorphisation of AI systems. In his paper, *Artificial Intelligence meets natural stupidity*, Drew McDermott articulates this issue that is just as relevant today as it was back in 1976 when the paper was published. First, many researchers have the tendency to describe internal representations of artificial systems using natural language terms. In other words, developers and engineers adopt everyday language to describe machine operations. This leads to wishful thinking by overestimating the capabilities of deep networks. This linguistic approach to code writing and development has led us to believe that "the human use of language is a royal road to the cognitive psyche" (McDermott, 1976, p. 7). This kind of anthropomorphism is often used in day to day discourse and, as such, we often attribute understanding to objects around us. We say that the thermometer *feels* the temperature; the door *knows* to open itself when a person is around, but neither the thermometer nor the door feels or knows. Searle states (1980, p. 419) that we use verbs like these because we have the tendency to extend our intentionality and impose it on these artifacts. However, there are no instances of intentionality in any AI or day-to-day artefacts. The apparent intelligence of artificial systems is a product of computational power, depth of layers, and statistical correlations. The distinction between weak AI (systems that simulate reasoning) and strong AI (systems that possess consciousness) maintains its importance. Considering the contemporary artificial intelligence systems, no architecture meets the criteria for strong AI. With this context established, it is time to turn to Dennett and Searle, whose theories offer accounts of consciousness and intentionality. Their relevance will be established in the following sections.

Searle's biological naturalism

John Searle supports the biological naturalism view (Searle, 2004, p. 113), which places consciousness in the field of biological phenomena. However, Searle asserts that consciousness is superior to the physical level (Chua, 2017, p. 46). In

Mind, the author distinguishes between two types of reductions: causal and ontological (Searle, 2004, p. 119). When it comes to causal reductions, an X phenomenon is causally reduced to a Y phenomenon if and only if X entirely results (causally) from Y's behavior. On the other hand, ontological reduction means that phenomenon X is ontologically reduced to Y if and only if X is nothing else other than Y.

The most important aspect of biological naturalism is that conscious and subjective states are part of actual phenomena. This means that consciousness is not an illusion, nor can it be ontologically reduced to the neurobiological dimension. Consciousness cannot be ontologically reduced to neurobiology because this would mean a shift from *first person* to *third person*. Consciousness has, by definition, a first-person ontology, and as Searle put it, the meaning of the concept is lost if we redefine it in third-person terms (Searle, 2004, p. 123). Searle mentions identity theorists, which identify consciousness with a neurobiological process in order to consolidate his argument. He denies this identity and states that "we can treat one and the same event as having both neurobiological features and phenomenological features. One and the same event is a sequence of neuron firings and is also painful" (Searle, 2004, pp. 124-125).

John Moses Chua notes how Searle explains consciousness by dividing it into three moments. The first moment is physical, in which receptors react to an external stimulus, like the pinch of the skin, and send information to the brain, then the neural activity begins. This process is observable and can be measured. After neurons process the input, the second moment begins: the sensation (in this case, pain) or qualia, the subjective phenomena, immeasurable and only sensed by the experiencing person. The second moment can only exist if we take the first-person ontology into account because this phenomenon cannot be discussed in third-person terms. The third moment is the reaction, which is observable and measurable, such as muscles tensing, rapid heart beating (Chua, 2017, p. 49).

Searle maintains that consciousness cannot be compiled by software alone because programs cannot possess consciousness. To defend this statement, Searle proposes the already well-known thought experiment called the Chinese Room Argument. This argument underlines that no computational system can prove it possesses intentionality (Boyles, 2012). We bring up a concept discussed in the first section, specifically the classical theory of mind, which Searle implicitly agrees to. This is because the CRA concerns how an AI would not be able to *understand* Chinese, but this reduces the AI problem to a symbolic processing issue, just as classical theorists suggest. Deep learning systems are reducible to

computation, but they do not rely on symbolic architectures that were present at the time when Searle constructed the CRA.

This raises the question: is Searle's biological naturalism compatible with deep artificial neural networks (ANNs)? The answer is ambivalent. On one hand deep networks are capable of syntax in natural languages, but they do not have access to the semantic dimension of language. In this matter, Searle's critique still stands: artificial systems do not *know* Chinese, they merely act as if they do. Apart from this, there is a more nuanced aspect to be discussed. Suppose qualia can be causally reduced to neurobiological activity. This means that our neural network is the cause of subjective phenomena and consciousness, which implies that if we copy every aspect and detail of our neural network, in principle, we would end up with a system that possesses qualia. However, this approach would lead to two functioning systems that we cannot understand. If we cannot understand human qualia, then we will not be able to understand or even recognize artificial qualia. In other words, we lack the means and criteria to determine whether the system possesses qualia or simulates it. Therefore, even if we had the means to recreate the human neural network to the smallest detail, we would not undoubtedly know if we were standing in front of an actual artificial agent.

However, developing such an agent might help us better understand how our minds work. If Searle is correct and consciousness emerges in a 1:1 ANN, the system would serve as a mirror which allows investigation of the mind. Dennet notes in *From Bacteria to Bach and Back* that we are having trouble explaining how our mind works because it is so very close and personal (Dennett, 2017, p. 18). Maybe by externalizing consciousness, we can study it as an impersonal object with no connection to ourselves whatsoever. Dennett also offers a different approach to consciousness and qualia, as it will be discussed in the following section.

Dennett's computational functionalism

Daniel Dennett has a reductive computationalist view of consciousness, stating that all mental phenomena can be reduced to computational processes. This makes Dennett a reductionist because he rejects the existence of any ontologically superior domain, holding that there is nothing more to the mind than neural activity (Chua, 2017, p. 46). In *From Bacteria to Bach and Back*, Dennett introduces the concept of strange inversion of reasoning. One good example of such inversion is the transition from geocentrism to heliocentrism. Specifically, this transition is a strange inversion of reasoning because scientific discoveries overturned common sense. Dennett believes that competence without

comprehension is such an inversion. In other words, the generally accepted opinion is that comprehension in a particular domain precedes competence. However, the author believes that systems often exhibit competence without any understanding. Following Darwin and Turing, Dennett believes that the simplest life forms and machines are competent to do what they must do, but they do not possess any comprehension whatsoever. Just like McDermott, Dennett acknowledges the dangers of anthropomorphism. We anthropomorphize plants and bacteria to understand their behavior. The problem with anthropomorphizing is that we credit them not just with competencies but also with comprehension of actions, benefits, or reasons behind behaviors (Dennett, 2017, p. 85).

In his essay *Quining Qualia*, Dennett denies the existence of qualia. In order to consider this subjective feature as a component of experience, Dennett argues that we would need to be able to recognize when a change in qualia has occurred (just as we can recognize changes in other types of states), or that there must be a detectable difference between experiencing a change in sensation and not possessing the sensation at all. Dennett concludes the essay by asserting that we can demonstrate neither of these conditions, therefore, he maintains that qualia cannot be a component of experience (Dennett, 1988). As previously mentioned, Dennett opposes the naturalistic perspective. To support his claim that qualia do not exist and are nothing more than an illusion, he introduces the concept of Cartesian gravity. Specifically, imagine a mind explorer who wants to begin the investigation of consciousness starting with his own mind. This explorer is at home, on Planet Descartes, contemplating the task of uncovering the mysteries of consciousness. He observes the external universe from a first-person point of view. Cartesian gravity is what keeps him fixed in an egocentric position. His internal monologue, reminiscent of Descartes, goes: here I am, a conscious being, intimately familiar with the idea in my mind, which I know better than anyone else, simply because they are mine. In the meantime, from afar, another explorer approaches, a scientific one, equipped with maps, theories, and models, and so on, determined to also discover the nature of consciousness. The closer this scientific explorer gets, the more uneasy he feels. He is drawn into a perspective he should avoid, but the gravitational pull is too strong. Upon landing on Planet Descartes, his third-person, objective perspective is transformed into a first-person perspective. Moreover, the scientific explorer finds himself unable to use the tools he brought along. Cartesian gravity cannot be resisted once one is on this planet. This shift from the third-person to first-person is inevitable and represents a clear case of strange inversion of reasoning. Two contradictory points are revealed, but they cannot coexist (Dennett, 2017, p. 20, adaptation).

Dennett's notion of Cartesian gravity makes use of the same concepts employed by Searle in arguing for the phenomenological dimension of consciousness, which, in Searle's view, coexists with its neurobiological basis. However, Dennett contends that these two aspects cannot coexist simultaneously. The feature that appeals to neuronal activity represents the scientific approach to explaining consciousness, while the phenomenological trait corresponds to the individual's own perspective, the one experienced on Planet Descartes. Thus, qualia, the phenomenological trait of consciousness is, according to Dennett, only an illusion, one that is particularly difficult to abandon because the gravitational pull of Cartesian subjectivity is too strong.

Instead of a centralized system, Dennett proposes the multiple drafts model (Dennett, 1991, pp. 111-115). Specifically, ideas or perceptions are multiple narrative fragments, also called drafts, which exist in various stages of editing. In his view, these drafts serve different specific functions. They do not arrive in a single location, such as a central processing unit (CPU) in the brain, but are edited throughout the brain in ways that shape cognition. The question which drafts are conscious, or capable of leading to a conscious state inevitably places us once again on Planet Descartes. For Dennett, consciousness is a kind of virtual machine, an evolved computer that gives structure to brain activity (Chua, 2017, p. 48). The multiple drafts model has lead Dennett to assert:

> *if the self is just the Center of Narrative Gravity, and if all the phenomena of human consciousness are explicable as just the activities of a virtual machine realized in the astronomically adjustable connections of a human brain, then, in principle, a suitably programmed robot, with a silicon-based computer brain, would be conscious, would have a self.* (Dennett, 1991, p. 431)

Conclusion

From an engineering perspective, Dennett's theory appears more aligned with current views in AI than Searle's. Although more instinctively plausible, Searle's theory puts more obstacles in the way of engineers. However, if Dennett is correct and constructing a silicon brain is what it takes for a conscious machine to exist, then it means we could bypass the Cartesian gravity. Studying an external consciousness would allow the usage of every scientific tool we possess without us being pulled into the first-person perspective. Yet the same problem remains: even if a machine exhibits human capacities, how could we determine whether it truly possesses a subjective point of view? We don't know if qualia exist, but there is a practical nature to it. As Raúl Arrabales et al. put it, on one hand, a comprehensive understanding of qualia might make possible the building of conscious machines;

on the other hand, the path to a complete understanding of qualia in biology might lay through the research on new computational models (Arrabales *et al*., 2010). Specifically, even though Dennett's perspective remains controversial, his theory allows the possibility of artificial consciousness without any biological constraints. By using Dennett's strange inversion of reasoning, maybe we need not understand human qualia to construct artificial qualia, but instead perhaps developing more advanced deep networks and computational models is what will lead to understanding consciousness.

References

1. Arrabales, R., Ledezma, A., & Sanchis, A. (2010). On the practical nature of artificial qualia. *Proceedings of the 2010 Annual Convention of the Society for the Study of Artificial Intelligence and Simulation of Behavior (AISB 2010)*.
2. Boyles, R. J. (2012). Artificial Qualia, Intentional Systems and Machine Consciousness. *Proceedings of the Research@DLSU Congress 2012: Science and Technology Conference*, 110a–110c.
3. Buckner, C. (2019). Deep learning: A philosophical introduction. *Philosophy Compass*, *14*(10), 1-19.
4. Buckner, C., & Garson, J. (1997, May 18). Connectionism. In E. N. Zalta & U. Nodelman (Eds.), *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/connectionism/.
5. Chua, J. M. (2017). Formulating Consciousness: A Comparative Analysis of Searle's and Dennett's Theory of Consciousness. *Talisik Undergraduate Journal of Philosophy*, *IV*(1), 43-62.
6. Deng, L., & Yu, D. (2013). Deep Learning: Methods and Applications. *Foundations and Trends® in Signal Processing*, *7*(3-4), 197-387.
7. Dennett, D. C. (1988). Quining Qualia. In A. Marcel, & E. Bisiach (Eds.), *Consciousness in Modern Science*. Oxford University Press.
8. Dennett, D. C. (1991). *Consciousness Explained*. Back Bay Books.
9. Dennett, D. C. (2017). *From Bacteria to Bach and Back: The Evolution of Minds*. W.W. Norton.
10. McDermott, D. (1976). Artificial intelligence meets natural stupidity. *ACM SIGART Bulletin*, (57), 4-9.
11. Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill Education.
12. Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417-424.
13. Searle, J. R. (2004). *Mind: A Brief Introduction*. Oxford University Press: USA.