Consciousness: Just Another Technique?

Ulrike Barthelmeß¹ · Ulrich Furbach¹

Received: 19 June 2021 / Accepted: 10 August 2021 / Published online: 21 August 2021
© The Author(s) 2021

Abstract

This note is intended as a contribution to the discussion whether artificial systems can have consciousness. Based on the notion of Tononi's Information Integration Theory we will argue, that AI systems that have to reason with large knowledge bases need such techniques in order to handle them efficiently. We furthermore discuss mind-wandering and creativity on this basis.

What is it like to be a bat? asks philosopher Thomas Nagel in the title of an essay that has become famous [4]. We cannot know because we are fixed to an outside perspective in our observation. We may know a lot about the cognitive abilities of a bat, but we have no idea what it feels like to navigate by echolocation. We are concerned here with the concept of *qualia*, the subjective experiential content of a mental state, the *phenomenal consciousness*. The term consciousness has different meanings in usage. We sometimes mean heightened attention, controlled perception as opposed to unfocused or unconscious perception, sometimes the state of wakefulness as opposed to mental absence, sleep or unconsciousness. Unfortunately, a scientifically strict and uniform definition of the term does not exist. Philosophy, psychology and neuroscience each emphasize different aspects. The neuroscientist António Damásio defines consciousness as a state of mind in which one has knowledge of one's own existence and the existence of an environment. We are concerned with consciousness as a phenomenon of experience. The thought experiment of the blind physicist Mary may help us to understand what is meant by this: Mary is blind from birth, she has never seen colors, but she knows everything about the physics of colors. She has all of

humanity's knowledge about colors and color vision. Now, if one day Mary suddenly can see and sees the color red in front of her, that is she is experiencing red, two different reactions are conceivable: "Yes, that's red." or "Wow, that's what red is like!" Did she learn something new about red by seeing it or did she already know everything?

Consciousness and knowledge: From a materialistic or physicalistic point of view, Mary would not know—or have experienced—more after seeing the red. But if the wow-effect occurs, Mary has learned something new, has been enriched by a subjective experiential content, a *qualia*. Researchers look at consciousness and *qualia* from a variety of perspectives. According to neuroscientists Giulio Tononi or Bernhard J. Baars, humans need consciousness to efficiently cope with the immense amount of knowledge, memories and sensory information they constantly have to deal with. Tononi's so-called Integrated Information Theory [8] studies any networked physical system, which does not necessarily have to be biological. How do Mary and a photodiode or a mega-pixel camera differ when presented with the color red? Mary's brain takes on one of an extremely large number of possible states, the photodiode only one of two possible states, red or not red. The camera also has millions of possible states, but unlike Mary, these are isolated, they are not connected. In Mary's brain, neurons are activated that form an intense connection, they bundle together and form a core structure that, when it exceeds a certain threshold, conveys a certain experience of red that is linked to other concepts, memories and experiences. Tononi's theory uses a mathematical structural analysis to determine the extent to which the system can integrate information. The aim is to identify which parts of the system integrate information in a spatio-temporal framework in order to achieve a

Supported by the German Research Foundation (DFG) under the grant FU 263/19-1 CoRg—Cognitive Reasoning. "Computer auf dem Weg zum Bewusstsein", a German version of this article, was published in Frankfurter Allgemeine Zeitung on July 5th, 2021.

✉ Ulrike Barthelmeß
ubarthelmess@gmx.de

Ulrich Furbach
uli@uni-koblenz.de

¹ University of Koblenz, Koblenz, Germany

coherent and unified state. According to Tononi, these areas, so-called complexes, correspond to the “subjects of conscious experience”, this is the seat of qualia. The measure obtained also indicates the degree of consciousness that the system possesses. Baars’ Global Workspace Theory [1], also known as Baars’ Theatre, emphasises even more clearly that consciousness is indispensable for mastering large amounts of knowledge. The theatre metaphor describes a stage with a spotlight (attention) on which the actors struggle to enter the light of the spotlight. Behind the stage, i.e. in the darkness of the unconscious, a lot of people are active (technicians, authors, directors) to support the action. In the auditorium are many people representing knowledge, skills and experience. This total system is consciousness. Together with Tononi’s considerations, it becomes clear that the phenomenon of consciousness is not an all-or-nothing property, but is to be regarded as gradually present. If one assumes that the mental is interwoven with the material and that both manifest themselves in different degrees, then consciousness is different in degree or complexity. This approach is central to *gradual panpsychism*, which we will return to later.

Handling large amounts of knowledge: Both models discussed above are interesting for computers. Why? Can it be difficult for computers to process large amounts of knowledge when they are perfect at handling huge databases? On the one hand, the usual internet search engines show us every day how quickly we can get any information on demand at any time. For example, the search engine operator Google uses, maintains and expands the so-called *Knowledge Graphs*, which contain knowledge in formalised form. In May 2020, 500 billion facts about five billion terms were stored in it. Despite this huge amount of data, search engines can find their way around it in a fraction of a second. On the other hand, the situation is incomparably more complex when not only something is being searched for, but when factual knowledge is to be used to find connections, explanations or justifications.

This is a central challenge in so-called everyday reasoning, an important subfield of artificial intelligence. Let’s take the following example from a test database: The task here is to explain a fact, such as the statement “My body casts a shadow on the grass”. In our example, the question is “What caused this?” and the alternative answers are “The sun was rising.” and “The grass was cut.”. The system must now select the more plausible of the two answers. It quickly becomes clear what knowledge is needed to answer such questions, which are trivial for humans: We need to know that shadows are caused by lighting and that the rising sun is a source of light, while the condition of the grass has little to do with the shadows cast. The problem for an artificial system is to identify these relevant parts of knowledge in the huge knowledge base and then to process them. The processing itself, i.e. the logical reasoning from individual facts,

has already been very well studied. There are numerous powerful systems for logic-based automatic reasoning that also play an important role in industrial software development. However, if one wants to combine these systems with extensive knowledge sources such as the aforementioned Knowledge Graphs, new approaches are necessary. In the research project CoRg¹, for example, a system was developed that has many aspects of Baars’ theatre concept. The stage is the working area of logical reasoning, the individual logical formulas are brought into the centre of the action by the spotlight and the knowledge base corresponds to the dark auditorium, which has enormous knowledge, experiences and memories. All processing is controlled behind the scenes by algorithms, strategies and heuristics.

Mind-wandering and creativity as indicators of consciousness: Of course, the system described above does not have human-like consciousness—but in the sense of the theories of Tononi and Baars presented before, it is “somewhat conscious”. This allows phenomena related to consciousness to be explored experimentally. Mind-wandering is one such phenomenon. Mind-wandering in humans occurs when a person is awake, turned away from the outside world, thinking of nothing concrete, indulging in daydreams. Neuroscientists have shown that in these states free from external stimuli, the so-called *idle* or *default mode network* (DMN) is active [5]. These mental time-outs provide an opportunity to spontaneously run through thoughts, produce new perspectives, simulate scenarios that can be helpful for processing the past or planning future actions. When a person’s attention is taken up again by a task, the energy of the DMN is shut down in favour of other areas that concentrate on the task. Such mind-wandering is also possible in our AI system: starting from the sentence “The dog hurt its paw”, the CoRg system wanders via terms such as *danger*, *wound*, *disease* to *infection* or in another chain via *wounded*, *fight*, *conflict* to *military*.

A similar mechanism can be used to approach the topic of creativity. There are AI systems that generate literary texts, compose music or produce paintings. The painting Edmond de Belamy, a portrait of a fictitious person produced by an artificial neural network, became famous. The painting was auctioned at Christie’s for more than \$400,000. The valuation of such artistic artefacts is highly controversial, especially the question of whether there is really a creative act at work here. One could answer the question in the affirmative if one considers procedures for testing the creative potential of people. For example, the *Remote Associate Test* [2] has been used as a creativity test in different variants since the 1950s. In the simplest form, test persons are presented with three unrelated terms, such as *snow*, *carrot*, *sledge*.

¹ Cognitive Reasoning, <http://corg.hs-harz.de>.

They then have to associate another term that connects all three terms in a meaningful way—in our case, this could be the term *snowman*. It is conceivable that the mechanism of meaningful association is closely related to the mind-wandering described above. And indeed, the CoRg system can process such tests and in our case also suggest the *snowman*.

Why do we need consciousness when AI systems are so successful with lots of training? The focus of this article's reflections so far has been the concept of knowledge and in particular very large amounts of knowledge and their mastery. However, there are AI systems that are not based on explicit knowledge at all and they are currently being euphorically celebrated. The world-champion board players, the language translation programmes, the controls of autonomous vehicles or even the art-producing systems mentioned above mostly work with artificial neural networks. Such systems have no a priori knowledge; they learn by being presented with a large number of training examples or by generating them themselves. The AI system MuZero [6], for example, needs about a million training games against itself to learn both the rules and the playing style of chess or Go to the point where it ranks among the best in the world. These brilliant successes are made possible by machine learning, ever-growing amounts of available data and advances in hardware development. Such systems are also very powerful in everyday reasoning.

A major disadvantage, however, is that they usually cannot explain why they arrived at their result. They have learned to solve a task well. However, they do not have the necessary declarative knowledge to point out connections and thereby communicate with us about the solution. In AI research, there has therefore been a trend towards explainable AI for some time. Researchers are trying to combine symbolic or knowledge-based techniques with statistical machine learning methods, such as neural networks, in order to obtain not only results, but also explanations. Since 2018, the European AI network *Claire* (Confederation of Laboratories for Artificial Intelligence Research in Europe)² has also been dedicated to the vision of developing trustworthy AI systems that complement human intelligence rather than replacing it.

What is it like to be an AI system? We discussed that systems that are purely focused on material processes (finding and matching data, repetitive training) are not sufficient to decisively develop explainable AI. For this purpose, it could be helpful to have an instrument that, in analogy to human association, has the ability to assign concepts to each other, to undertake mind games (mind wandering) and thus creatively find solutions because they operate similarly to humans. It is foreseeable that those systems will become more and more sophisticated.

Nevertheless, they are non-biological artefacts. They have no biography and therefore do not possess phenomenal

consciousness like humans or possibly other living beings. They have no corporeality in the sense of the French philosopher and phenomenologist Merleau-Ponty [3]. His concept of consciousness is based on the fact that a living being participates in the world with its senses in order to survive and at the same time is itself part of the world. This does not apply to AI creatures (until now)—and, may be, that is a good thing. Because this way we don't have to deal with human shortcomings in them, such as moodiness, listlessness, impatience or vanity.

Nevertheless, we can attribute a certain consciousness to AI systems if we adopt Tononi's concept of gradual consciousness. This can be used to identify an area in the system that is referred to as the area of qualia, the seat of phenomenal sensation—and not only in biological systems. In this respect, this concept is similar to that of *Gradual Panpsychism*, which is propagated, among others, by Spät [7]. It is based on the assumption that in our perceptible world there is a gradual order of the mental, which increases continuously with the complexity of physical things and living organisms. This does not mean that atoms, stones or plants have a conscious experience similar to ours, but that they have certain properties that can no longer be attributed to matter alone. Physical complexity goes hand in hand with mental complexity. This perspective allows for a kind of AI consciousness, however that it is like to be.

Funding Open Access funding enabled and organized by Projekt DEAL.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Baars BJ (1997) In the theatre of consciousness. Global workspace theory, a rigorous scientific theory of consciousness. *J Conscious Stud* 4(4):292–309
2. Mednick SA (1962) The associative basis of the creative process. *Psychol Rev* 69(3):220–232
3. Merleau-Ponty M (1966) *Phänomenologie der Wahrnehmung*. De Gruyter, Berlin
4. Nagel T (1974) What is it like to be a bat? *Philos Rev* 83(4):435–450

5. Otti A, Gündel H, Wohlschläger A, Zimmer C, Sorg C, Noll-Hussong M (2012) Default-mode-Netzwerk des Gehirns. *Der Nervenarzt* 83(1):16–24
6. Schrittwieser J, Antonoglou I, Hubert T, Simonyan K, Sifre L, Schmitt S, Guez A, Lockhart E, Hassabis D, Graepel T, Lillicrap T, Silver D (2020) Mastering atari, go, chess and shogi by planning with a learned model. *Nature* 588(7839):604–609
7. Spät P (2010) Panpsychismus - Ein Lösungsvorschlag zum Leib-Seele-Problem. PhD thesis, Univ. of Freiburg
8. Tononi G (2004) An information integration theory of consciousness. *BMC Neurosci* 5(1):42