

Review of: "What is it like to be an AI bat?"

Andrzej Brodziak¹

¹ Państwowa Wyższa Szkoła Zawodowa w Nysie

Potential competing interests: No potential competing interests to declare.

The reviewed work concerns a very difficult, interdisciplinary issue. The paper is influenced by the increasing capabilities of AI systems and the ongoing debate about the nature of consciousness and its applicability to non-biological entities. The authors aim to clarify the complex relationship between consciousness theories and their application in AI, arguing for a pragmatic approach. To propose an evaluation of this article, I had to confront several complicated problems.

The title of the article

The first difficulty encountered was understanding the title of the work. To deal with this difficulty, I asked a question about it to OpenAI-GPT and received an answer, which I quote:

My question to GPT: Please try to formulate three explanations of the following title of the analyzed scientific paper: "What is it like to be an AI bat?"

The answer of GPT:

1. Exploration of AI Consciousness through the Lens of Animal Cognition: The title "What is it like to be an AI bat?" suggests a study that delves into the realm of artificial intelligence (AI) consciousness by drawing parallels with the cognitive experiences of non-human animals, specifically bats. This approach could be used to explore how AI might perceive and interact with the world, comparing it to the sensory and cognitive processes of bats. The paper might investigate the challenges and possibilities of replicating animal-like consciousness in AI, using the unique echolocation abilities and nocturnal lifestyle of bats as a model for understanding how an AI might develop its own form of sensory processing and environmental awareness.
2. Philosophical Implications of Artificial Consciousness: The title could be referencing a philosophical exploration of artificial consciousness, inspired by Thomas Nagel's famous essay "What is it like to be a bat?". In this context, the paper would discuss the subjective experience and the possibility of 'qualia' in artificial beings. It might delve into questions about whether an AI can truly experience sensations or emotions in a manner akin to living creatures, and what it means for an AI to have a subjective perspective or consciousness. The choice of a bat, an animal with a sensory experience vastly different from humans, underscores the challenge of understanding consciousness from an entirely different perspective, be it animal or artificial.
3. AI's Ability to Emulate Complex Biological Systems: The title might indicate a study on the capability of AI systems to simulate or emulate complex biological processes, using a bat's unique sensory and navigational abilities as a test case.

This approach would explore the extent to which AI can mimic or understand biological experiences that are fundamentally different from human experience. The paper could cover advancements in AI algorithms and neural networks that allow machines to process information in a way that's reminiscent of a bat's echolocation and discuss the implications of such technology in understanding animal behavior or enhancing AI's interaction with the physical world. This could also touch on the ethical and practical aspects of creating AI systems that emulate living creatures.

Indeed, the authors of the work refer to that famous essay by Thomas Nagel, 'What is it like to be a bat?'. In the text of the work, one can find the following statement:

„The seminal paper “What is like to be a bat?” [39] claims the inability of other creatures to experience the unique qualia of specific phenomenal experience. Non-bats cannot imagine or reproduce, let alone feel and “know” from the perspective of first-person anything like that. The hard problem is a fundamental barrier that is impossible to overcome [40][41], or at least it produces an explanatory gap [42] in understanding.”

This statement is important, but I believe that the article by David Herzog and Nitsa Herzog goes significantly beyond the topic defined in the title. The qualia experienced by bats are surely quite different from the qualia experienced by humans, and even from the qualia experienced by cats. Nevertheless, the self-awareness of bats, cats, and humans is still the same class of phenomena, and they can all be distinguished in relation to the phenomena occurring in contemporary artificial intelligence systems. However, for this statement of mine to be understandable, one must try to articulate what we understand by the concept of self-awareness in humans and animals.

What we understand by concepts such as sentience, self-awareness, and consciousness.

The formulation of the opinion about the value of the reviewed article requires clarifying the understanding of the concepts of sentience, self-awareness, and consciousness. So, using the conclusions from our previous articles [1,2,3], we will adopt the following understanding of these concepts:

Sentience

Sentience is the awareness that external interaction is taking place. It is also the capacity to have subjective experiences and feelings, or more simply, the ability to perceive and experience sensations. This notion is commonly used in discussions about animal rights. In the context of animals, sentience refers to the ability to feel, perceive, or experience subjectively. This includes the capacity to experience positive or negative influences and emotional states, like happiness or suffering.

Self-awareness

We will try to approximate what self-awareness is – first - by attempting to answer the question: what does the statement 'I am self-aware' mean in colloquial understanding?

To respond to this question, we must first define such a mental state as simply and intuitively as possible. So, it seems that to be conscious, one needs to... *{understand one's own separateness from other beings (and objects) perceived*

around oneself and be able to perceive oneself against the background of images of the surrounding world, taking into account one's past existence and anticipating the outline of possible future events).

To refine this description, it's necessary to introduce the extremely important concept of imagination (mental images, imagery). This initial definition can then be clarified. It's also important to discuss the concept of self-image. Imagining objects previously seen, such as an apple, a tree, a chair, a table, involves recalling an image of these objects from memory. This enables us to see them with our eyes closed, albeit less vividly. We sometimes describe this sensation as 'seeing something with the eyes of the imagination'. The same applies to auditory and tactile perceptions, as well as actions. Movements or complex behaviors can also be imagined.

Thus, we can now say that consciousness consists of the ability of the nervous system to *imagine oneself against the background of the image of the known world or on the backdrop of imagery of the world*. However, in the proposed explanation, there is the phrase '*imagine oneself*', the meaning of which requires further clarification. To progress in our inference, we need to define what self-perception is.

Living beings, especially humans, experience a constant state of self-perception during periods of wakefulness. Beyond the visual and auditory sensations perceived then, animals and humans experience sensations from within their own bodies. These stimuli come from surface sensory receptors, joint sensory receptors, sensations from the respiratory system, the heart, and sometimes other organs. Sensations from one's own body can also be imagined to some extent.

Self-images also include autobiographical data. Shortly after waking up, there's a moment of quick review of one's biographical path, a mental journey back in time, which gives exactly the feeling that I am 'John Doe'. Intense self-images also include memorized data about social and situational relationships. The ability to recall 'stories about oneself' and various types of 'opinions' about oneself is also important.

Since it can be stated that the sense of consciousness occurs only when awake, when body perception takes place, the image of oneself is a slightly different process than the imagination of an object known from visual perception. The image of oneself is superimposed on the perception of oneself.

Hearing or listening to words in natural language, such as apple, tree, chair, activates the same neuronal structure that was active during the perception of images of these objects. However, the neuronal process occurring after the activation of the word 'I' is much more complicated, and a substantive discussion of it requires considering contemporary neurophysiological theories.

Consciousness

Consciousness is the state of being aware of and able to think and perceive one's surroundings, thoughts, and emotions. It encompasses self-awareness, cognition, and the subjective experience of existence.

Beyond individual awareness, it also implies a collective understanding and interconnectedness among beings. This multifaceted concept has been explored in various disciplines, from neuroscience to philosophy, and remains a central

subject of inquiry, bridging the tangible and the intangible, the known and the mysterious.

Self-awareness and consciousness in terms of neuroscience

Above intuitive, common-sense definition of consciousness can be presented using the terminology of neuroscience. We did this in our recent work [1,2]. We state therein that understanding self-awareness requires the integration of several contemporary neurophysiological theories. We write there that a condensed, brief presentation of an attempt to integrate these theories is as follows:

(i) - Some researchers emphasize that understanding the nature of consciousness requires the ability to explain the feeling of qualia. The 'Orch - OR' theory, proposed by Stuart Hameroff and Roger Penrose, which suggests that quantum processing occurs in microtubules, offers a potential explanation for the essence of feeling qualia. We propose that the perception of qualia can also be understood as specific changes in the shape of the brain's electromagnetic field, formed during perception [4].

(ii) - In order to be aware, we also need to feel our bodies, remember various data about the world (including opinions about ourselves), and, importantly, be able to retrieve these data, that is, to realize mental imaginations. We should also be able to conceive an image of ourselves ("self-image"). For the 'image of oneself' to arise, there must be a recursive circulation of impulses across many levels of the multi-level structure, including the parietal and prefrontal lobes.

(iii) - It is necessary to be able to explain what structures and processes are needed for the formation of the subject, the element called by the word 'self' or "I". MacFadden, author of the "conscious electromagnetic information field theory (cemi)," points out, probably rightly, that for "consciously perceive something," not only a certain process takes place, but also the existence of spatial objects is necessary. The realization of such a spatial object is possible thanks to the electromagnetic field of the brain that is created in the process of imagining oneself. Such an object can be mentally identified with that subject, the self [4].

(iv) - When aware, we always pay attention to a specific, chosen area – in other words, we "focus attention" on a certain field of matters. It's necessary to explain the neural basis of the process of "focusing attention" on a selected area.

Experiencing qualia by bats, cats, and humans in the light of neuroscience.

In light of the cited neurophysiological data, the distinctiveness of qualia experienced by bats, emphasized by the authors of the article, does occur, but the difficulty of its explanation has the same nature as the hard problem in the overall theory of human brain function. It seems that the attempts to explain this phenomenon by Stuart Hameroff, Roger Penrose, Joe McFadden, as mentioned above, are so far the most convincing.

Contrasting self-awareness in humans and animals with potentially analogous phenomena in Artificial Intelligence Systems.

To consider the behavior and reactions of contemporary advanced artificial intelligence systems, one must, in my opinion, try to answer the question of whether there are known and possible systems that are highly intelligent and wise, but which

probably do not possess self-awareness.

In my opinion, such systems can be indicated. The human embryo in the first weeks of its development is an undoubtedly wise, complicated, and effective entity, as it carries out the creation of an extremely complicated human organism. The action of the genetic information contained in the DNA of embryo cells - in the first weeks of their development - works - it seems - like a complicated algorithm. However, the embryo does not yet possess self-awareness at that time.

The operation of contemporary large generative language systems (such as ChatGPT, Bard, Anthropic) also gives the impression of being reasonable, intelligent, even wise.

As is known, in the various descriptions of the essence of these systems' operation, we often find a metaphor or a reference to the functioning of the human brain's speech centers. However, in descriptions of the human brain's functioning, one does not attempt to assert that having a speech center is sufficient to explain the essence of self-awareness.

Since I am trying to define the value of the work by David Herzog and Nitsa Herzog in an understandable way, I will therefore move on to an attempt at an intuitively understandable yet non-existent abstract model of an A.I. system that would possess self-awareness. I will call it here the {A.I.-with-A.C} system.

Now, if we assume that the reasoning part of a large, futuristic A.I. has access not only to images, photos, and films, but also to cameras installed on the planet's surface and astronomical telescopes, then we could say that this {A.I.-with-A.C} is perceiving its environment. The {A.I.-with-A.C} system would, by its very nature, also have insight into the essence of all existing software. With such an assumption, such a system {A.I.-with-A.C} would be able to formulate an "image of itself – the equivalent of a human's conception of oneself, i.e., the equivalent of the concept of 'I'.

Since such a situation has not yet occurred, it remains to be considered whether phenomena - in a sense similar to neurophysiological models of phenomena of sentience, self-awareness, consciousness - appear in existing contemporary A.I. systems.

The authors make a useful compilation of theories proposed so far concerning such quasi-consciousness phenomena. They mention, among others, Higher Order Theories, Global Workspace Theory, and Integrated Information Theory.

The dissatisfaction with the presented considerations, however, arises from the lack of attempts to present or formulate ways (tests) that would be able to check whether the considered artificial intelligence systems exhibit features of self-awareness. As is known, even for testing manifestations of - already quite well understood - human and animal consciousness, there are not many such effective methods. Among them, the Turing test is listed, which, as explained by experiments with advanced language models, tests rather only the level of intelligence and not so much manifestations of consciousness. Recently, there have also been many critical remarks about the limited usefulness of the so-called "mirror test". I believe that the problem of developing theoretical foundations and practical implementations of methods for detecting the beginnings of self-awareness in advanced A.I. systems is very important.

In my opinion, the development of A.C. theories also has philosophical and worldview significance, as with the

development of systems like {A.I.-with-A.C}, the question of how our individual, human personal self-awareness will be incorporated into the somewhat overarching consciousness of {A.I.-with-A.C} systems will become increasingly practical.

Final conclusion of my review

In my opinion, the article is very valuable in the context of the rapid advancement of AI, addressing fundamental questions about consciousness.

References

1. Brodziak A. A current model of neural circuitry active in forming mental images. *Med Sci Monit.* 2013;19:1146-58. <https://doi.org/10.12659/MSM.889587>
2. Różyk-Myrta A, Brodziak A, Muc-Wierzoń M. Neural circuits, microtubule processing, brain's electromagnetic field - components of self-awareness. *Brain Sci.* 2021;11(8):984. <https://doi.org/10.3390/brainsci11080984>
3. Brodziak A, Abram D, Różyk-Myrta A. Planetary consciousness incites probably transcendent feelings and deepens the polarization of worldviews. *Qeios* 2023, <https://doi.org/10.32388/BU24PQ>
4. Brodziak A, Romaniuk F. The functional unit of neural circuits and its relations to eventual sentience of artificial intelligence systems. *Qeios*, 2023. <https://doi.org/10.32388/82VRPG>