

Is it possible to identify phenomenal consciousness in artificial systems in the light of the gaming problem?¹

Michele Farisco^{a,b,*}, Kathinka Evers^a

^aCentre for Research Ethics and Bioethics, Uppsala University, Uppsala, Sweden

^bBiogem Molecular Biology and Genetics Research Institute, Ariano Irpino (AV), Italy

*Correspondence: Michele Farisco (michele.farisco@uu.se).

Abstract

We analyze the question how phenomenal consciousness (if any) might be identified in artificial systems with specific reference to the gaming problem (i.e., the fact that the artificial system is trained with human-generated data, so that possible behavioral and/or functional evidence of consciousness is not reliable). Our goal is to review selected illustrative approaches for advancing in this direction. We highlight strengths and shortcomings of each approach, finally proposing a combination of different strategies as a promising task to pursue.

¹ The manuscript has been submitted to the special issue “Evaluating Artificial Consciousness” (in *Philosophy and the Mind Sciences*).

Introduction

Among the issues arising from recent development of Artificial Intelligence (AI), the prospect of artificial forms of consciousness emerges as particularly challenging, raising significant expectations (either positive or negative) from both specialists in the field and the general public (Lenharo, 2024, Colombatto and Fleming, 2023). One of the main challenges raised by a hypothetical AI consciousness is how to identify it. In fact, the classical other minds problem, which essentially makes any attribution of consciousness inferential, is even more tricky for candidate conscious AI systems in virtue of the so-called gaming problem. The gaming problem arises in connection to the hypothesis of artificial consciousness, specifically from the hypothesis of artificial phenomenal consciousness or sentience, because artificial systems are instructed with human-generated data in order to give them the appearance of instantiating human features, wherefore functional or behavioral markers of sentience are unreliable and cannot be considered as evidence of actual phenomenal consciousness (Birch and Andrews, 2024).

Chang-Eop has recently proposed a logical argument in support of the unreliability of behavioral manifestations of conscious abilities by artificial systems, specifically by Large Language Models (LLM), concluding that in the case of LLM consciousness denial leads to contradiction while consciousness affirmation leads to indeterminacy (Chang-Eop, 2024).

The gaming problem has both theoretical and ethical relevance, since it may be argued that artificial systems should not confuse human users about their conscious state (Schwitzgebel, 2023a). Therefore, it is particularly urgent to elaborate possible strategies for managing and prospectively solving the gaming problem.

In the following we review five illustrative approaches that have been proposed by experts in the field for identifying conscious abilities in AI systems. We provide a description of these strategies for eventually highlighting that a combination among them is probably the best choice for advancing towards a more reliable attribution of conscious capacities to AI systems.

We start from a definition of phenomenal consciousness as the subjective feeling of a particular experience, or “what it is like to be” in a particular state (Block, 1995, Block, 2022), and as “sentience”, or the capacity for having valenced experiences, like pain or pleasure (Birch, 2024, Butlin and Lappas, 2025).

Then we describe five strategies that have been articulated for detecting hypothetical artificial forms of phenomenal consciousness: 1. A theory-based strategy that starts from selected theories of consciousness to infer relevant indicators (Butlin et al., 2023) 2. A life-based strategy which outlines the necessary connection of consciousness with biological life (Seth, 2024); 3. A brain-based strategy that takes the brain, its evolution, and its correlation with consciousness as a benchmark for artificial consciousness (Farisco et al., 2024b, Aru et al., 2023); 4. A consciousness-based strategy that searches for other forms of biological consciousness than the

human, in order to identify what (if anything) is really indispensable to consciousness and what is dispensable, and thus overcoming the controversy between the many theories of consciousness and proceed towards the identification of reliable evidence of artificial consciousness (Birch and Andrews, 2024); 5. A indicators-based approach that elaborates a list of indicators, conceived as features that we tend to agree characterize conscious experience and that are indicative (i.e., probabilistic rather than definitive evidence) of the presence of consciousness in artificial systems (Pennartz et al., 2019), and on their basis elaborate relevant tests for artificial consciousness (Bayne et al., 2024, Elamrani and Yampolsky, 2019).

1. Phenomenal consciousness

Phenomenal consciousness may be defined as the *subjective experience*, “what it is like to be” in a particular state (i.e., including experiences of perceptions which are not cognitively accessed) (Block, 1995, Block, 2022). Another way to conceive phenomenal consciousness is in terms of “*sentience*”, which basically refers to the capacity for having valenced (or hedonic) experiences, like pain or pleasure (Birch, 2024).

Valenced experience relates to affective states, either positive or negative. In fact, together with arousal, valence is a fundamental component of affect (Russell, 1980, Posner et al., 2005). As reviewed by Birch, there are different understandings of the nature of valence, including its description as immediate quality of the experience (i.e., raw feelings), as non-conceptual representation of value, and as imperative content (Birch, 2024). The concept of value has been recently explored in relation to hypothetical conscious AI systems (Farisco and Evers, In Press). The starting point of this analysis is the capacity for evaluation (i.e., sensitivity to reward signals and the ability to discriminate between good and bad things in the world on the basis of specific needs, motivations, and goals) as a defining feature of biological organisms, namely of the human brain. In fact, evaluation has been posited as a fundament of learning and memory (Dehaene and Changeux, 1989, Edelman, 1992), as well as of consciousness and ethical deliberation (Evers, 2009). A crucial characteristic that makes this kind of evaluation performed by biological organisms a conscious experience is the capacity for autonomously performing it on the basis of subjectively salient drives.

Importantly, evaluation thus conceived, as well as the related concept of value, include both cognitive and emotional dimensions: cognitive representations like concepts, goals, and beliefs are combined with emotional attitudes having positive or negative valence (<https://www.psychologytoday.com/intl/blog/hot-thought/201304/what-are-values>). Emotions are particularly crucial for evaluating information coming from the inside of the organism (Bennett, 2023).

The concepts of phenomenal consciousness and sentience play an important role in both theoretical and moral debates. There is a wide tendency to identify consciousness *tout court* with

phenomenal consciousness, at least in a morally relevant sense (Levy, 2014), or to affirm the necessity of phenomenal consciousness for any cognitive form of consciousness to be possible. With reference to the prospect of artificial consciousness, according to these perspectives, the central question is whether it may be possible to engineer an artificial form of subjective experience.

2. Levels of phenomenal consciousness

Since subjective experience or the phenomenal character of a conscious experience may have different levels, and these levels may depend on some specific aspects or dimensions (e.g., duration, intensity, frequency, precision, etc.), we cannot logically exclude *prima facie* that a system, whether biological or artificial, may have at least very rudimentary, low-level forms of subjective experience. Accordingly, we must avoid identifying subjective experience exclusively with high-level, sophisticated forms (e.g., those usually attributed to human subjects), and avoid an anthropocentric bias in the attempt to identify it in non-human systems (e.g., in artificial systems). Accordingly, we should avoid setting the bar too high in defining the gold standard in our search for indicators of phenomenal consciousness in AI.

Two low-level forms of consciousness are of particular interest to consider here: primary/minimal consciousness and contentless consciousness. Advocates of different models have proposed the concept of primary or minimal consciousness as opposed to secondary or more advanced consciousness within a graded view of consciousness. One proposal is to conceive minimal consciousness as subjective experience. More specifically, as the most basic (non-reflective) subjective feeling that includes exteroceptive (e.g., visual, olfactory), interoceptive (e.g., pain, hunger, thirst) and proprioceptive (bodily position) experiences (Bronfman et al., 2016, Merker, 2007). The key point of this proposal is that consciousness includes a more basic, non-cognitive form (Pereira Jr, 2021, Baluška and Reber, 2019, Baluška et al., 2025). The proponents of this view argue that the concept of consciousness as a fundamental capacity for experiencing that is distinct from high-level cognitive capacities is useful to identify what is necessary and sufficient for appropriately characterizing a (biological or artificial) system as (minimally) conscious.

The concept of anoetic consciousness as introduced by Endel Tulving is very close to this view (Tulving, 2005). According to him, there are three kinds of consciousness: autonoetic, which is related to the knowledge of the self; noetic, which is related to the knowledge of the outside world; and anoetic, which is related to the absence of explicit current knowledge (LeDoux and Lau, 2020). Anoetic consciousness is conceived as the condition of being alive and responsive to stimuli as opposed to having explicit conscious contents (LeDoux, 2021), or as "a stream of pre-reflective affective and sensorial perceptual consciousness essential for the waking state of the

organism in the absence of an explicit self-referential awareness of associated cognitive contents" (Vandekerckhove et al., 2014) (p. 6).

Others use “primary” or “sensory” (Edelman, 2003) (Feinberg and Mallatt, 2016, Edelman, 1989) consciousness to indicate a basic capacity to detect stimuli, to process their saliency and value, and to react accordingly. Importantly, primary/sensory consciousness as qualified, for instance, by Gerald Edelman can have access only to the present, actual experience.

To illustrate, for Feinberg and Mallatt (Feinberg and Mallatt, 2018), the basic form of consciousness is “value-based” as distinguished from “image-based” consciousness. The value-based consciousness does not rely on any kind of explicit, mental images of the world, but rather on an organic, and in some cases neuronal, map or schema that allows the organism to discriminate affordances (i.e., to detect and distinguish dangers and positive opportunities) in order to increase its fitness (cf. (Changeux, 1986, Changeux, 2004)).

The distinction between basic and more sophisticated forms of consciousness has been proposed also with specific reference to phenomenal consciousness, for instance through the distinction between minimal phenomenal selfhood (i.e., “the experience of being a distinct, holistic entity capable of global self-control and attention, possessing a body and a location in space and time” (Blanke and Metzinger, 2009) (p.7)), and a more elaborated phenomenal self-experience (Blanke and Metzinger, 2009, Seth and Tsakiris, 2018).

Contentless consciousness, also called pure consciousness or minimal phenomenal experience, is a form of conscious experience devoid of any specific content (Metzinger, 2020) that is related to the abovementioned distinction between more basic and more sophisticated forms of phenomenal consciousness. Notably, contentless consciousness is considered as either the grounding or the highest form of consciousness in some Eastern religions and spiritual traditions, as well as the ultimate goal in meditation (Thompson, 2015). There is discussion about whether this form of consciousness actually exists, and, if it does, what its main features might be and how to eventually operationalize it in order to elaborate criteria for identifying it (Sullivan, 1995, Woods et al., 2024). The possibility of indeterminate cases of sentience, i.e. systems for which there is no determinate fact of the matter whether they are conscious or not (see (Lee, 2020, Simon, 2017, Schwitzgebel, 2023b)), adds further complexity to the question whether artificial phenomenal consciousness/sentience is possible and how to detect it.

3. How can phenomenal consciousness be operationalized and made detectable in artificial systems?

Different approaches to the question how to detect phenomenal conscious activity in artificial systems have been suggested. Among them, we here focus on the following five:

1. Theory-based strategy: starting from selected theories of consciousness in order to infer relevant indicators (Butlin et al., 2023)
2. Life-based strategy: consciousness necessarily connects with biological life (Seth, 2024)
3. Brain-based strategy: the brain, its evolution, and its correlation with consciousness are benchmark for artificial consciousness (Farisco et al., 2024b, Aru et al., 2023)
4. Consciousness-based strategy: searching for other forms of biological consciousness than the human, in order to identify what (if anything) is really indispensable to consciousness and what is dispensable, and thus overcoming the controversy between the many theories of consciousness and proceed towards the identification of reliable evidence of artificial consciousness (Birch and Andrews, 2024) (<https://aeon.co/essays/to-understand-ai-sentience-first-understand-it-in-animals>)
5. Indicators-based strategy: starting from the features of consciousness we tend to agree about for elaborating a list of indicators of consciousness in artificial systems (Pennartz et al., 2019), and related tests for artificial consciousness (Bayne et al., 2024, Elamrani and Yampolskly, 2019).

3.1 Theory-based strategy

The first strategy we focus on consists in starting from selected theories of consciousness in order to infer indicators of consciousness from them. (Butlin et al., 2023) illustrates this kind of approach. In this paper, the authors propose a “rigorous and empirically grounded approach to AI consciousness”: they select some neuroscientific theories of consciousness and identify respective indicators of consciousness. They base their analysis on three main premises. First, they assume computational functionalism (i.e., performing computations of the right kind is necessary and sufficient for consciousness) as a working hypothesis; second, they claim that neuroscientific theories of consciousness can help us in reliably assessing AI consciousness because of the empirical support they have; and finally they argue that a theory-heavy approach (i.e., explicitly relying on empirically grounded theories of consciousness) is the best strategy for assessing AI consciousness. Basically, the theory-based strategy they argue for attributes consciousness to AI systems depending on how many indicators of consciousness inferred from the selected theories the systems in question show.

The authors acknowledge that there are many scientific theories of consciousness, not all compatible and not always commensurable either (Seth and Bayne, 2022, Evers et al., 2024, Chis-Ciure et al., 2024, Northoff and Lamme, 2020, Kuhn, 2024). Consistently, they do not endorse any specific theory, but rather derive a list of indicators from the sampled theories, and consider the evidence deriving from them cumulative, meaning that the more indicators are present, the more likely it is that the system in question is conscious.

The theories they refer to are the following, with related indicators:

- *Recurrent processing theory*, from which they derive two indicators: input modules using algorithmic recurrence and input modules generating organized, integrated perceptual representations.
- *Global workspace theory*, from which they derive four indicators; multiple specialized systems capable of operating in parallel (modules); limited capacity workspace, entailing a bottleneck in information flow and a selective attention mechanism; global broadcast: availability of information in the workspace to all modules; state-dependent attention, giving rise to the capacity to use the workspace to query modules in succession to perform complex tasks.
- *Computational higher-order theories*, from which they derive four indicators: generative, top-down or noisy perception modules; metacognitive monitoring distinguishing reliable perceptual representations from noise; agency guided by a general belief-formation and action selection system, and a strong disposition to update beliefs in accordance with the outputs of metacognitive monitoring; sparse and smooth coding generating a “quality space”.
- *Attention schema theory*, from which they derive one indicator: a predictive model representing and enabling control over the current state of attention.
- *Predictive processing*, from which they derive one indicator: input modules using predictive coding.

In addition to these theories, the authors refer also to *Agency and embodiment*, from which they derive two indicators: agency, conceived as learning from feedback and selecting outputs so as to pursue goals, especially where this involves flexible responsiveness to competing goals; embodiment, conceived as modeling output-input contingencies, including some systematic effects, and using this model in perception or control.

As mentioned above, the authors consider the evidence deriving from the identified indicators as cumulative. In fact, some AI systems already manifest some of the indicators, while there is currently no system capable of consistently implement all of them. Therefore, the authors conclude that there is no system yet that can be considered a strong candidate for consciousness according to a theory-based approach.

The second strategy we describe consists in considering consciousness as a biological phenomenon which is a feature of life and dependent on it.

Different authors propose theoretical interpretations that tend to endorse this view, including John Searle (Searle, 2007), Peter Godfrey-Smith (Godfrey-Smith, 2016, Godfrey-Smith, 2023), Evan Thompson (Thompson, 2018, Thompson, 2007), and Anil Seth (Seth, 2024, Seth, 2021), among others. Here we summarize the position of the latter, which is illustrative of the life-based strategy.

Anil Seth frames his proposal within biological naturalism, according to which consciousness is a property of only (even though not all) living systems. He acknowledges that this view has different variations, arguing that it is not necessary that life is carbon-based. In principle this leaves the door open to the possibility of a “living conscious AI”.

Importantly, Seth clarifies what is the meaning of consciousness he refers to in his analysis: conscious experience, or “to feel like something”. Accordingly, consciousness is real, not an illusion, and there is a fact-of-the-matter about whether something is conscious.

Basically, biological naturalism as defined above opposes to computational functionalism. In fact, for biological naturalism the biological basis of consciousness is not just an enabling factor for the right kind of computation, like, for instance, argued for in (Wiese, 2024), but rather a fundamental condition for consciousness. On the basis of theoretical models like predictive processing and the free energy principle, Seth argues for what he calls an “ontological continuity” between life and consciousness: living systems are autopoietic (i.e., they continually regenerate themselves), and this characteristic of the living matter is the basis for conscious activity. Interestingly, according to this view consciousness may be instantiated in very basic, rudimentary forms, a “ground state” of conscious experience characterized by “an inchoate, shapeless, formless feeling of simply ‘being alive’”.

At the theoretical level, two forms of biological naturalism are possible: a weak biological naturalism, for which life matters to consciousness “in virtue of enabling (or being necessary for) specific patterns of functional organization”, and strong biological naturalism, for which consciousness fundamentally depends on life. For both these versions, conscious AI will not come along for the ride as conventional AI gets smarter, because, as said above, consciousness is conceived as a property of living systems, therefore AI needs to be alive for being conscious. Neuromorphic and synthetic technologies could potentially be promising in this respect, if we accept the hypothesis that the more AI is similar to the brain and living systems, the more likely it is that it may be conscious.

The crucial point of the life-based strategy is that life, either carbon-based or of different nature, imposes a number of constraints and necessary conditions for consciousness to arise. For instance, Seth outlines the crucial role played by metabolic constraints, electromagnetic fields, and fine-grained timing relations (Seth, 2024). In short, there are several dependencies and contingencies accumulated in a “generative entrenchment” in the internal organization of the brain, which eventually constrain hypothetical alternative implementations of its mental features. The conclusion by Seth is that it may be impossible to separate what brains do from what they are: a failure to replicate the significant amount of functional and architectural details that appear to play an important role for consciousness would result in a change of the system’s organization, which likely affects the functions that the replicator system is able to perform.

3.2 Brain-based strategy

The third strategy we describe can be qualified as brain-based because it assumes that the brain, its evolution, and its correlation with consciousness are the benchmark for developing artificial

consciousness. Examples of this approach may be found in (Aru et al., 2023) and (Farisco et al., 2024a). The latter paper analyses artificial consciousness from an evolutionary perspective, taking the evolution of the brain and its relation with consciousness as a reference model or benchmark. This analysis reveals a number of structural and functional characteristics of the brain that arguably play a key role for human-like conscious experience. Importantly, current AI systems lack these features, or at least they fail to instantiate a sufficient level of these features. The proposal of the authors is taking inspiration from the brain to advance towards developing conscious AI.

More specifically, the following brain features are proposed to be of relevance in research on conscious AI :

- Hierarchical, nested, and multiple levels of organization
- Rewards sensitivity based on embodied multidimensional sensorimotor experience
- Spontaneous physiological activity contributing to conscious information processing
- The capacity to produce variability from the same functional organization and the long postnatal development resulting in multiple nested epigenetic synapse selections
- Degeneracy, that is different connection patterns may have the same function, leading to plasticity, individual differences, and creativity
- Capacity for interaction between individuals in a social group
- A physical and operational distinction between conscious and non-conscious brain representations, leading to the development of semantic competence and singular properties of association of conscious representations.

Starting from these features, the authors identify the following limitations of current AI that need to be ameliorated for advancing towards conscious AI:

- Computational approach with insufficient brain data
- Hardware with limited chemical elements
- Hardware implementing parallel levels of computation opposed to the nested brain organization
- Lack of an evolutionary and epigenetic development
- Lack of a multidimensional and multisensory representation based on embodiment
- Lack of social relationships and intrinsic ethical responsibility
- Lack of differentiation of the singularity of the individual and their life experience.

The hypothesis suggested is that the more an AI system implements the identified brain features correlated to consciousness, and the more an AI system improves the limitations listed above, the more likely it is that it may be or become conscious.

Importantly, (Farisco et al., 2024a) limit their analysis to human-like consciousness, but they also acknowledge that in principle it is possible to have an AI consciousness alternative to the human model. While this possibility cannot logically be excluded, assuming human consciousness and correlated brain features as reference models or benchmarks is a pragmatic choice that makes it possible to formulate realistic hypotheses.

Along a similar line of thought, Rosa Cao has argued that the functional roles of brain components involved in the generation of consciousness (e.g., neurons) can be taken to constrain physical realizations and what kind of material stuff can perform the relevant functions (Cao, 2022). In other words, functions imply a set of constraints on their realizers. In the case of neurons, which are assumed as structural units of the brain, there are many fine-grained factors that affect their functions, including location and spatial extension, history and interaction with the environment (epigenetic dynamics), sensitivity to temperature, proteins on the cellular membrane and inside the cell, ion channels, neurotransmitter receptors, etc. It is necessary to properly acknowledge these details of the brain architecture, and to take inspiration from them in order to advance towards a realistic implementation of artificial consciousness.

3.3 Consciousness-based strategy

The fourth strategy we introduce can be named consciousness-based because it aims to identify other forms of consciousness than the human to disentangle what is dispensable and what (if anything) is indispensable to consciousness. On that basis, proponents of this approach argue that it would be possible to overcome the high controversy among different theories of consciousness and also to identify reliable indicators for conscious activities in non-human systems, including AI.

This approach originates in part from the need to overcome the so-called “gaming problem”, which is defined by Birch and Andrews as “the phenomenon of non-sentient systems using human-generated training data to mimic human behaviours likely to persuade human users of their sentience” (Birch and Andrews, 2024). As the authors rightly outline, it is not a matter of deception, or not necessarily a matter of deception, but a direct consequence of the way and of the kind of data AI systems are trained with that make behavioral (either linguistic or, in the future, embodied) markers of consciousness unreliable. To overcome this crucial limitation, it is necessary to identify, if possible, what cannot be mimicked but is a strong marker of conscious state. Birch and Andrews point in the direction of architectural features that cannot be gamed, like the kinds of performed computations or the representational formats used in computations. In his latest book, Birch refers to “deep computational markers” of sentience, that is characteristics that are below the level of behavior and that for this reason AI systems cannot game (Birch, 2024). This line of reasoning appears to go in the same direction as the theory-based strategy described above. Accordingly, Birch proposes to deduce deep computational markers of sentience from computational functionalist theories. The evidence deriving from these markers increases if a system originally not programmed to display them is able to learn how to recreate them. Yet the lack of transparency of current AI systems, which literally operate as black boxes also to the developers’ eyes, poses a huge challenge to this attempt to identify deep computational markers. While Birch acknowledges this difficulty, he also rightly highlights that it is a technical rather than a logical obstacle.

Overall, the methodology that Birch and Andrews recommend to advance the identification of reliable markers of AI consciousness consists in expanding our study of consciousness in order to include also other evolved instances beyond humans. Notably, evolutionary data indicate that consciousness has originated three different times and it has subsequently evolved along three

different directions: in arthropods (including crustaceans and insects), in the cephalopods (including octopuses), and in the vertebrates. A comparative investigation of the different forms of consciousness evolved along these paths is key for first identifying what (if anything) is really indispensable for being conscious and then for checking whether artificial systems may have such features. If so, they may be regarded as strong candidates for consciousness.

3.4 Indicators-based strategy

The last strategy we identify can be defined as indicators-based approach. It basically aims at starting from some basic features of consciousness for elaborating a list of indicators of consciousness (Pennartz et al., 2019) and to apply them to artificial systems. These indicators, conceived to be indicative rather than definitive proof of conscious capacity in AI, but at the same time as operationalizable information, can serve as a basis for elaborating tests for artificial consciousness (Bayne et al., 2024, Elamrani and Yampolskly, 2019).

We may qualify this approach as heuristic because it makes explicit that the rationale behind the use of “indicator” rather than “criteria” or “markers” is that the knowledge we can get in our attempt to explore the mental features and capabilities of other subjects, including artificial systems, is always inferential and tentative. Accordingly, like in analogous positions (Ginsburg and Jablonka, 2019), the identified indicators are positive indicators or indicators of presence of consciousness: if present, they justify attribution of consciousness while their absence does not exclude the possibility that consciousness is present even if not detected or detectable². Furthermore, this approach delineates a research strategy rather than providing definitive answer to the challenge of operationalizing and detecting consciousness in AI systems.

The identified features of consciousness considered common to different theoretical frameworks are the following (Pennartz et al., 2019):

1. Qualitative richness: conscious experience is qualified by distinct sensory modalities and submodalities
2. Situatedness: conscious experience is specified by the subject’s spatiotemporal condition
3. Intentionality: consciousness is about something other than its neuronal underpinnings
4. Integration: the components of the conscious experience are perceived as a unified whole

² A complementary strategy consists in looking for negative indicators of consciousness, conceived as factors that are themselves at most slightly probability-raising, while their absence can negatively affect the probability that a system is conscious. Wanja Wiese distinguishes direct and indirect negative indicators. The direct negative indicators provide information directly relevant to the absence of consciousness. In other words, they indicate general constraints on artificial consciousness, and the absence of negative indicators is probability-lowering. The indirect negative indicators affect to what extent positive indicators are probability-raising. See <https://www.youtube.com/watch?v=XwTkCNj1L3k>. In Wiese’s formulation, there are three negations at play: negative indicators whose absence negatively affects the probability of consciousness being present. We do not find this clarifying. The additional distinction between direct and indirect negative indicators is also not helpful for increasing understanding, in our view. Instead, we would propose to stay with the simple term “indicator” and place the negation in the object in terms of absence: indicator of absence of consciousness versus indicators of presence of consciousness.

5. Dynamics and stability: conscious experiences include both dynamic changes and short-term stabilization.

Against this background, (Pennartz et al., 2019) introduce six indicators, some of which can be assessed through the analysis of behavior, while others refer to structural and organizational characteristics:

1. Goal directed behaviour (GDB) and model-based learning. In GDB the agent is driven by expected consequences of their action, and they know that their action is causal for obtaining a desirable outcome. Model-based learning depends on subjective ability to have an explicit model of ourselves and the world surrounding us.
2. Brain anatomy and physiology. Since the consciousness of mammals depends on the integrity of particular cerebral systems (i.e., thalamocortical systems), it is reasonable to think that similar structures indicate the presence of consciousness.
3. Psychometrics and meta-cognitive judgment. If a subject can detect and discriminate stimuli and can make some meta-cognitive judgments about perceived stimuli, it is probably conscious.
4. Episodic memory. If a subject can remember events (“what”) they experienced at a particular place (“where”) and time (“when”), it is probably conscious.
5. Acting out one’s subjective, situational survey: illusion and multistable perception. If a subject is susceptible to illusions and perceptual ambiguity, it is probably conscious.
6. Acting out one’s subjective, situational survey: visuospatial behaviour. If a subject is able to perceive objects as stably positioned, even when the subject is moving in their environment and scan it with their eyes, it is probably conscious.

Importantly, the evidence deriving from these indicators is cumulative: it increases with the number of indicators present at the same time.

As seen above, the indicators may be of different kinds, including anatomic, (neuro)physiological, and cognitive-behavioral. It may be that for artificial systems some are more relevant than others (e.g., the cognitive-behavioral more than physiological), or that some of these indicators should be reframed/adapted (e.g., the anatomic in terms of structure/architecture, like feedback vs feedforward systems), and even complemented with other kinds of indicators (e.g., computational analogies)

This approach can be qualified as heuristic because it does not provide definitive solutions to the issue of AI consciousness but rather aims to inspire further research through the elaboration of a theoretical framework and the identification of specific features that can be operationalized in research programs. For instance, the indicators-based approach inspires the following strategy for facing the gaming problem: observing the candidate AI system for a prolonged time in its ethological/environmental condition, to check whether it is able to adapt on the fly manifesting the ability of flexibly modulate its functions, excluding that this is just the mechanistic result of the pre-programming or training of the system.

4. Discussion

4.1 Advantages and shortcomings of each strategy

The five strategies summarized above present specific advantages and shortcomings. The theory-based approach has the advantage of an extensive theoretical base relying on empirically validated theories, as well as of the conceptual clarity about the specific notion of consciousness which is the object of the identified indicators (i.e., phenomenal consciousness or sentience). Yet this latter advantage is contrasted by the fact that this approach is necessarily selective, and while some theories may propose similar measures and indicators of consciousness (Farisco and Changeux, 2023), many differ on both points, posing also a problem in terms of commensurability (Evers et al., 2024). Other shortcomings of the theory-based approach include its controversial theoretical premise (i.e., computational functionalism), which despite its wide popularity among the AI community is deeply problematic both theoretically and ethically (Farisco-Evers 2025)(Seth, 2024); the question how to solve the indicators' necessary/sufficient issue; the fact that the theory-based strategy eventually reflects the limitations of the selected theories, including the risk of biases (e.g., anthropocentric or bio-centric), how to compare different theories in order to derive a cumulative evidence (in fact, as described in more details below, different theories often target different forms or kinds of consciousness, so that the indicators derived from them are not really cumulative), and the gold standard or generalization problem (Bayne et al., 2024): theories of consciousness are usually formulated with reference to typical adult consciousness, so that the relevance of the indicators derived from them to address artificial consciousness is questionable.

The life-based approach has the advantage of relying on the fact that all known examples of conscious systems are biological, therefore the strategy is not speculative (i.e., it has an explicit connection with actual instances of consciousness). On the other hand, even if not all of its proponents have this tendency, the life-based approach may be interpreted as excluding off-hand the possibility of forms of AI consciousness alternative to the biological ones.

The brain-based strategy has the advantage of relying on empirically grounded data about the brain bases of consciousness, therefore it has a strong scientific base. Also, this approach avoids speculations about hypothetical alternative forms of consciousness, and it is pragmatic in the sense that it is prone to be operationalized into specific roadmaps to test machine consciousness. Among the shortcomings of the brain-based approach there is the fact that it is limited to human-like forms of consciousness, so it may lead to overlooking alternative forms of machine consciousness. Moreover, since the differences between current AI and the brain are quite profound, the brain-based strategy may set the bar too high and (at least for now) be inapplicable.

The consciousness-based strategy has the advantage that it in principle avoids anthropomorphic and anthropocentric traps; in fact, it accommodates evidence coming from different instances of consciousness. Among the shortcomings of this approach there is the fact that, at least in its current form, it makes reference to selected computational theories of consciousness, thus raising the risk of reproducing the same shortcomings of the theory-based approach. In relation to this point, the search for computational markers of consciousness in AI systems is limited by the architecture of actual AI systems, which lack transparency and explicability. Finally, the

consciousness-based approach risks of replacing a big challenge (identifying AI consciousness) with another big challenge (advancing a comparative understanding of different forms of consciousness in nature).

The indicators-based strategy has the advantage of relying on what we tend to agree characterize conscious activity, trying to remain neutral with regards to specific theories of consciousness: in principle, it is compatible with different theoretical accounts. Also, it can accommodate the prismatic nature of consciousness, including its phenomenal and cognitive dimensions. Yet, resembling the generalization or gold standard problem, the indicators-based approach has the shortcoming of being conceived with reference to biological consciousness, and therefore its relevance and applicability to AI consciousness may be limited.

Against this background, no strategy stands out as decisively better than the others, and it may seem intuitively reasonable then to try to combine the strongest features of the different strategies to advance towards the identification of reliable indicators of consciousness in AI. Yet this intuition needs to be justified and informed by analytical reflection, particularly by an appropriate comparison of the strategies. To this end, a preliminary point to address is whether and possibly which of these strategies are actually commensurable (i.e., comparable because they share a common measure), and the indicators they propose eventually cumulative (i.e., they can be aggregated to get an overall “rate” of consciousness) rather than complementary (i.e., they offer indications about different forms and/or dimensions of consciousness).

4.2 Logical and empirical commensurability of the different strategies

How do the five strategies introduced above compare to each other in terms of logical and empirical structure? As summarized in Table 2, this question can be raised in a number of different perspectives, either logical or empirical, e.g., theoretical frameworks, interpretative approaches, definitions, methodologies, evidence of consciousness, principles for validation, or normative stand-points.

To illustrate, all the strategies except the theory-based share the same theoretical framework and interpretative approach (i.e., the way of assigning relevance and meaning to data). More specifically, the theory-based strategy endorses computational functionalism as a background theoretical view, assuming that consciousness is a function resulting from the right kind of computations. The life-based, the brain-based, the consciousness-based, and the indicators-based strategies rely on biological naturalism³ (i.e., they consider biological forms of consciousness as a benchmark for hypothetical AI consciousness, even if they do not in principle deny the possibility of alternative forms of consciousness) as a background theoretical view. Therefore, life-based, brain-based, consciousness-based, and indicators-based approaches are consistent in

³ To be more precise, the brain-based strategies relies on a neuro-biological naturalism, which is a more stringent form of biological naturalism. Yet this level of granularity is not necessary for this comparative analysis.

their theoretical frameworks and interpretative approaches, and the evidence deriving from them are cumulative (i.e., we may aggregate their respective evidence and get a synthetic, overall evidence). The theoretical frameworks and interpretative approaches of these four strategies differ from those of the theory-based strategy, so that the evidence deriving from the first four strategies is complementary to those deriving from the latter (i.e., we cannot aggregate their respective evidence and obtain a synthetic, overall evidence). To be more precise, if we want to know only whether the artificial system is somehow conscious (whatever its form of consciousness is), then we may aggregate the evidence deriving from all the strategies, but if we want to know more specifically which form of consciousness (i.e., biology-like vs computational) the system possibly has then we need to distinguish the evidence deriving from the different strategies.

If we consider the conceptual foundation of the strategies, all of them understand consciousness as phenomenal consciousness or sentience, either directly as the form of consciousness they are interested in or as part of a broader understanding of consciousness which include both cognitive and experiential forms. Therefore, all the strategies are consistent in terms of conceptual foundation, and we may aggregate the evidence deriving from them to eventually get an overall indicator of consciousness.

At the empirical level, we may compare the five strategies analyzed above in relation to the evidence of consciousness, that is the type(s) of data (e.g., quantitative or behavioral) considered necessary and/or sufficient for inferring consciousness. In this case, the theory-based, the consciousness-based, and the indicators-based strategies focus on the system's operations (i.e., what it does) and on its architecture (i.e., how it is internally organized). Therefore, they are empirically consistent in relation to evidence of consciousness, and the evidence deriving from them is cumulative. The life-based and the brain-based strategies focus on the system's nature (i.e., its inherent character), so that they are consistent in relation to evidence of consciousness, and the evidence deriving from them is also cumulative. Therefore, the first three strategies differ from the latter two with regards to the evidence of consciousness, and their respective evidence of consciousness is complementary. The point being that the empirical evidence for consciousness may derive from either the system's functions or its architecture: if we are interested in how empirical data may support in general the attribution of conscious capacities to AI, then we can aggregate the evidence from all the strategies, while if we want to know how much indicative of consciousness is a specific empirical data then we need to distinguish between function- and architecture-focused theories.

Still at the empirical level, we may compare the five strategies also in relation to their methodology, that is the general research strategy defining how the research process is undertaken. In this case, all the strategies are consistent, because they all propose to check the identified indicators empirically, that is looking for third-person, empirically testable evidence, even if they have different starting points for identifying the indicators.

In conclusion, the consistency among the five identified strategies, and consequently the kind of relationship between the evidence for consciousness deriving from them (i.e., complementary vs cumulative) depend on which specific dimension of logical or empirical structure we consider. For instance, all the five strategies appear logically consistent in relation to their conceptual foundations (i.e., they define consciousness as phenomenal consciousness or sentience, or are compatible with such definition of consciousness), and empirically consistent in relation to their methodology (i.e., they share the necessity of empirical testing the identified indicators).

This implies that in order to get a greater evidence of AI consciousness we should first clarify which specific logical and/or empirical dimensions we refer to and then eventually cumulate or complement the evidence deriving from the different strategies. Therefore, the best “meta-strategy” for advancing towards a better and stronger evidence of AI consciousness is combining different specific strategies, but how to concretely achieve this goal is a matter of analytical reasoning, and it needs a clear identification of which structure (i.e., logical or empirical) and respective dimension to address. What we suggest to be necessary is to combine different kinds of indicators in order to develop tests for consciousness that go beyond the classical two approaches (i.e., architectural or behavioral (Elamrani and Yampolskly, 2019, Bayne et al., 2024)) and that include an ethological/ecological and diachronic (i.e., extended in time) observation for addressing the issue of artificial consciousness.

4.3 Relevance of the different strategies to identify different levels of consciousness

As described above, consciousness, including phenomenal consciousness, may arguably exist at different levels, the main of which are primary or minimal consciousness, contentless consciousness, and reflective consciousness. In short, primary/minimal consciousness is the non reflective and non cognitive level of consciousness, or the fundamental capacity for experiencing. Contentless consciousness is the level of consciousness devoid of any specific content, like a raw feeling of existing. Reflective consciousness is the most sophisticated level of consciousness, characterizing a subject able to cognitively access their own experience.

The five strategies presented above appear to have different relevance to these three levels of consciousness (See Table 3). All the strategies potentially provide information relevant to address reflective level of consciousness, while not all appear relevant to the other two levels. The theory-based strategy stands on the strong computational functionalist premise, which excludes non cognitive and contentless levels of consciousness. Both the life-based and the consciousness-based strategies allow the possibility of both primary/minimal and contentless consciousness, for instance in the form of the feeling of being alive. Both the brain-based and the indicators-based strategies appear to be compatible, at least in principle, with primary/minimal consciousness but not with contentless consciousness, because both these strategies take a representational view of consciousness (i.e., consciousness refers to something, either cognitively or non-cognitively).

Since phenomenal consciousness is arguably multilevel, it is important to take into account which specific levels the different strategies above are relevant to, in order to avoid excluding off-hand some levels of consciousness because the chosen strategy is not compatible with them. In conclusion, the combination of different strategies is the best way to prevent this risk.

We propose that combining the different strategies described above is necessary for advancing towards a more reliable assessment of artificial consciousness for two reasons among others: first, consciousness is a multidimensional and multilevel feature, and combining different logical and empirical strategies increases the chances of covering this complexity; second, in order to address the gaming problem, it is crucial to look for as many indicators as possible, from structural to architectural, from functional to socio-ethological.

Conclusion

We have reviewed five strategies for operationalizing and identifying hypothetical phenomenal consciousness in AI systems. Each strategy has its own specific empirical and logical characterization, but they can be compared with each other in relation to different empirical and logical dimensions, resulting compatible or complementary. We propose to combine different strategies for elaborating more reliable indicators of consciousness in AI systems, either combining or complementing the evidence derived from compatible or complementary strategies, respectively. We argue that this combination is necessary for two reasons among others: the multidimensional and multilevel nature of consciousness, and the gaming problem.

Funding

This research has received funding from the project Counterfactual Assessment and Valuation for Awareness Architecture—CAVAA (European Commission, EIC 101071178).

Table 1

Approach	Main tenets	Advantages	Shortcomings
Theory-based	<ul style="list-style-type: none"> • Inferring indicators of consciousness from selected scientific theories of consciousness 	<ul style="list-style-type: none"> • Extensive theoretical base relying on empirically validated theories • Conceptual clarity about the specific notion of consciousness which is the object of the indicators 	<ul style="list-style-type: none"> • Computational functionalism regarding consciousness is deeply problematic both theoretically and ethically • <i>per force</i> selective: the theories out there are too many. Some of them may lead to propose similar measures and indicators of consciousness, but many differ on both points which poses a problem in terms of commensurability • affected by the limitations of the selected theories • risk of biases (e.g., anthropocentric or bio-centric) • how to solve the indicators' necessary/sufficient issue • how to compare different theories in order to derive a cumulative evidence (in fact different theories often target different forms or kinds of consciousness, so that the indicators derived from them are not really cumulative)

			<ul style="list-style-type: none"> • generalization problem, or the gold standard problem: theories of consciousness are usually formulated with reference to typical adult consciousness, so that the relevance of the indicators derived from them to address artificial consciousness is questionable.
Life-based	<ul style="list-style-type: none"> • Inferring indicators of consciousness from life features associated with consciousness 	<ul style="list-style-type: none"> • it relies on the fact that all known examples of conscious systems are biological, therefore it is not speculative 	<ul style="list-style-type: none"> • it may be interpreted as excluding off-hand the possibility of forms of AI consciousness alternative to the biological forms
Brain-based	<ul style="list-style-type: none"> • Inferring indicators of consciousness from brain features associated with consciousness 	<ul style="list-style-type: none"> • it relies on empirically-grounded data about the brain bases of consciousness, therefore it has a strong scientific base • it avoids speculations about hypothetical alternative forms of consciousness • it is pragmatic, in the sense that is prone to be operationalized into specific approaches to test 	<ul style="list-style-type: none"> • it is limited to human-like forms of consciousness, so it may lead to overlook alternative forms of machine consciousness • the distance between current AI and the brain is quite big, so this approach may set the bar too high and eventually be not applicable

		machine consciousness	
Consciousness-based	<ul style="list-style-type: none"> Comparing different kinds of consciousness to infer what is indispensable to it (and then check for it in AI) 	<ul style="list-style-type: none"> in principle, it avoids anthropomorphic and anthropocentric traps: it accommodates evidence coming from different instances of consciousness 	<ul style="list-style-type: none"> it eventually make reference to selected computational theories of consciousness, finally raising the risk of reproducing the same shortcomings of the theory-based approach the search for computational markers of consciousness in AI systems is limited by the architecture of actual AI systems, which lack transparency and explicability it risks of replacing a big challenge (identifying AI consciousness) with another big challenge (advancing a comparative understanding of different forms of consciousness in nature).
Indicators-based Approach	<ul style="list-style-type: none"> Inferring indicators of consciousness from agreed-upon features of consciousness 	<ul style="list-style-type: none"> it relies on what we tend to agree characterize conscious activity, trying to stay neutral with regards to specific theories of consciousness in principle, it is compatible with different theoretical 	<ul style="list-style-type: none"> it is conceived with reference to biological consciousness, and therefore its relevance and application to AI consciousness may be limited

		<p>accounts</p> <ul style="list-style-type: none"> • it can accommodate the prismatic nature of consciousness, including its phenomenal and cognitive dimensions 	
--	--	---	--

Table 2

KIND OF STRUCTURE	DIMENSION	THEORY-BASED	LIFE-BASED	BRAIN-BASED	CONSCIOUSNESS-BASED	INDICATORS-BASED
LOGICAL	CONCEPTUAL FOUNDATIONS (Understanding and definition of key concepts)	Phenomenal Consciousness	Sentience	Conscious process, compatible with Phenomenal Consciousness	Sentience	Both Access and Phenomenal Consciousness
		X	X	X	X	X
	THEORETICAL FRAMEWORK & INTERPRETATIVE APPROACH (The way of assigning relevance and meaning to data)	Computational functionalism.	Biological naturalism.	(Neuro-)biological naturalism.	Biological naturalism.	Biological naturalism.
		X	V	V	V	V
EMPIRICAL	EVIDENCE OF CONSCIOUSNESS (i.e., Type(s) of data (e.g., quantitative or behavioural) considered necessary and sufficient for inferring consciousness)	System's operations (what it does) + architecture (how it is internally organized)	System's nature (its inherent character)	System's nature + architecture	System's operations (what it does) + architecture (how it is internally organized)	System's operations (what it does) + architecture (how it is internally organized)

		X	V	V	X	X
	METHODOLOGY (i.e., General research strategy defining how the research process has to be undertaken)	From theories to indicators, to be checked empirically (i.e., looking for third-person, empirically testable evidence)	From life definition to indicators, to be checked empirically (i.e., looking for third-person, empirically testable evidence)	From brain features to indicators, to be checked empirically (i.e., looking for third-person, empirically testable evidence)	From different types of consciousness to indicators, to be checked empirically (i.e., looking for third-person, empirically testable evidence)	From working definition of consciousness and its features to indicators, to be checked empirically (i.e., looking for third-person, empirically testable evidence)
		X	X	X	X	X

Table 3

LEVEL OF CONSCIOUSNESS	THEORY-BASED	LIFE-BASED	BRAIN-BASED	CONSCIOUSNESS-BASED	INDICATORS-BASED
PRIMARY/MINIMAL	NO	YES	YES	YES	YES
CONTENTLESS	NO	YES	NO	YES	NO
REFLECTIVE	YES	YES	YES	YES	YES

References

- ARU, J., LARKUM, M. E. & SHINE, J. M. 2023. The feasibility of artificial consciousness through the lens of neuroscience. *Trends in Neurosciences*, 46, 1008-1017.
- BALUŠKA, F., MILLER, W. B., SLIJEPCEVIC, P. & REBER, A. S. 2025. Sensing, feeling and sentience in unicellular organisms and living cells. *BioSystems*, 247, 105374.
- BALUŠKA, F. & REBER, A. 2019. Sentience and Consciousness in Single Cells: How the First Minds Emerged in Unicellular Species. *Bioessays*, 41, e1800229.
- BAYNE, T., SETH, A. K., MASSIMINI, M., SHEPHERD, J., CLEEREMANS, A., FLEMING, S. M., MALACH, R., MATTINGLEY, J. B., MENON, D. K., OWEN, A. M., PETERS, M. A. K., RAZI, A. & MUDRIK, L. 2024. Tests for consciousness in humans and beyond. *Trends Cogn Sci*.
- BENNETT, M. S. 2023. *A brief history of intelligence : evolution, AI, and the five breakthroughs that made our brains*, New York, Mariner Books.
- BIRCH, J. 2024. *The Edge of Sentience : Risk and Precaution in Humans, Other Animals, and AI*.
- BIRCH, J. & ANDREWS, K. 2024. To Understand AI Sentience, First Understand it in Animals. *Intellectica*, 81, 213-226.
- BLANKE, O. & METZINGER, T. 2009. Full-body illusions and minimal phenomenal selfhood. *Trends Cogn Sci*, 13, 7-13.
- BLOCK 1995. On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18, 227-287.
- BLOCK, N. J. 2022. The border between seeing and thinking. *Philosophy of mind series*. New York, NY: Oxford University Press,.
- BRONFMAN, Z. Z., GINSBURG, S. & JABLONKA, E. 2016. The Transition to Minimal Consciousness through the Evolution of Associative Learning. *Front Psychol*, 7, 1954.
- BUTLIN, P. & LAPPAS, T. 2025. Principles for Responsible AI Consciousness Research. *arXiv:2501.07290* [Online].
- BUTLIN, P., LONG, R., ELMOZNINO, E., BENGIO, Y., BIRCH, J., CONSTANT, A., DEANE, G., FLEMING, S. M., FRITH, C., JI, X., KANAI, R., KLEIN, C., LINDSAY, G., MICHEL, M., MUDRIK, L., PETERS, M. A. K., SCHWITZGEBEL, E., SIMON, J. & VANRULLEN, R. 2023. Consciousness in Artificial Intelligence: Insights from the Science of Consciousness.
- CAO, R. 2022. Multiple realizability and the spirit of functionalism. *Synthese*, 200, 506.
- CHANG-EOP, K. 2024. The Logical Impossibility of Consciousness Denial: A Formal Analysis of AI Self-Reports. Available: *arXiv:2501.05454*.
- CHANGEUX, J.-P. 1986. *Neuronal man : the biology of mind / translated by Laurence Garey*, New York ; Oxford, Oxford University Press.
- CHANGEUX, J.-P. 2004. *The physiology of truth : neuroscience and human knowledge*, Cambridge, Mass., Belknap Press of Harvard University Press.
- CHIS-CIURE, R., MELLONI, L. & NORTHOFF, G. 2024. A measure centrality index for systematic empirical comparison of consciousness theories. *Neurosci Biobehav Rev*, 161, 105670.
- COLOMBATTO, C. & FLEMING, S. M. 2023. Folk Psychological Attributions of Consciousness to Large Language Models.

- DEHAENE, S. & CHANGEUX, J. P. 1989. A simple model of prefrontal cortex function in delayed-response tasks. *J Cogn Neurosci*, 1, 244-61.
- EDELMAN 1992. *Bright air, brilliant fire : on the matter of the mind*, New York, NY, BasicBooks.
- EDELMAN, G. M. 1989. *The remembered present : a biological theory of consciousness*, New York, Basic Books.
- EDELMAN, G. M. 2003. Naturalizing consciousness: a theoretical framework. *Proc Natl Acad Sci U S A*, 100, 5520-4.
- ELAMRANI, A. & YAMPOLSKLY, R. V. 2019. Reviewing Tests for Machine Consciousness. *Journal of Consciousness Studies*, 26, 35-64.
- EVERS 2009. *Neuroetique. Quand la matière s'éveille*, Paris, Odile Jacob.
- EVERS, K., FARISCO, M. & PENNARTZ, C. M. A. 2024. Assessing the commensurability of theories of consciousness: On the usefulness of common denominators in differentiating, integrating and testing hypotheses. *Consciousness and Cognition*, 119, 103668.
- FARISCO, M. & CHANGEUX, J. P. 2023. About the compatibility between the perturbational complexity index and the global neuronal workspace theory of consciousness. *Neurosci Conscious*, 2023, niad016.
- FARISCO, M. & EVERS, K. In Press. The capacity for evaluation of the human brain and its implications for an artificial moral subject. In: VIJAYARAGHAVAN, S. & FELSEN, G. (eds.) *Neuroscience and Society*. London: CRC Press - Taylor&Francis.
- FARISCO, M., EVERS, K. & CHANGEUX, J.-P. 2024a. Is artificial consciousness achievable? Lessons from the human brain. *Neural Networks*, 180, 106714.
- FARISCO, M., EVERS, K. & CHANGEUX, J. P. 2024b. Is artificial consciousness achievable? Lessons from the human brain. *Neural Netw*, 180, 106714.
- FEINBERG, T. E. & MALLATT, J. 2016. The nature of primary consciousness. A new synthesis. *Conscious Cogn*, 43, 113-27.
- FEINBERG, T. E. & MALLATT, J. 2018. *Consciousness demystified*, Cambridge, Massachusetts ; London, England, MIT Press.
- GINSBURG, S. & JABLONKA, E. 2019. *The evolution of the sensitive soul : learning and the origins of consciousness*, Cambridge, Massachusetts, The MIT Press.
- GODFREY-SMITH, P. 2016. Mind, Matter, and Metabolism. *Journal of Philosophy*, 113, 481-506.
- GODFREY-SMITH, P. 2023. Nervous Systems, Functionalism, and Artificial Minds.
- KUHN, R. L. 2024. A landscape of consciousness: Toward a taxonomy of explanations and implications. *Progress in Biophysics and Molecular Biology*, 190, 28-169.
- LEDOUX & LAU, H. 2020. Seeing consciousness through the lens of memory. *Curr Biol*, 30, R1018-R1022.
- LEDOUX, J. E. 2021. What emotions might be like in other animals. *Curr Biol*, 31, R824-R829.
- LEE, A. Y. 2020. Does sentience come in degrees? *Animal Sentience*, 29.
- LENHARO, M. 2024. AI consciousness: scientists say we urgently need answers. *Nature*, 625, 226.
- LEVY, N. 2014. The Value of Consciousness. *J Conscious Stud*, 21, 127-138.
- MERKER, B. 2007. Consciousness without a cerebral cortex: a challenge for neuroscience and medicine. *Behav Brain Sci*, 30, 63-81; discussion 81-134.
- METZINGER, T. 2020. Minimal phenomenal experience. *Philosophy and the Mind Sciences*, 1, 1-44.

- NORTHOFF, G. & LAMME, V. 2020. Neural signs and mechanisms of consciousness: Is there a potential convergence of theories of consciousness in sight? *Neurosci Biobehav Rev*, 118, 568-587.
- PENNARTZ, FARISCO, M. & EVERS, K. 2019. Indicators and Criteria of Consciousness in Animals and Intelligent Machines: An Inside-Out Approach. *Front Syst Neurosci*, 13, 25.
- PEREIRA JR, A. 2021. The role of sentience in the theory of consciousness and medical practice. *Journal of Consciousness Studies*, 28, 22-50.
- POSNER, J., RUSSELL, J. A. & PETERSON, B. S. 2005. The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. *Dev Psychopathol*, 17, 715-34.
- RUSSELL, J. A. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39, 1161-1178.
- SCHWITZGEBEL, E. 2023a. AI systems must not confuse users about their sentience or moral status. *Patterns (N Y)*, 4, 100818.
- SCHWITZGEBEL, E. 2023b. Borderline consciousness, when it's neither determinately true nor determinately false that experience is present. *Philosophical Studies*, 180, 3415-3439.
- SEARLE, J. R. 2007. Biological Naturalism. In: VELMANS, M. & SCHNEIDER, S. (eds.) *The Blackwell Companion to Consciousness*. Malden MA, Oxford, Victoria: Blackwell Publishing Ltd.
- SETH. 2024. Conscious artificial intelligence and biological naturalism.
- SETH, A. 2021. *Being you : a new science of consciousness*, New York, NY, Dutton.
- SETH, A. K. & BAYNE, T. 2022. Theories of consciousness. *Nat Rev Neurosci*, 23, 439-452.
- SETH, A. K. & TSAKIRIS, M. 2018. Being a Beast Machine: The Somatic Basis of Selfhood. *Trends Cogn Sci*, 22, 969-981.
- SIMON, J. A. 2017. Vagueness and zombies: why 'phenomenally conscious' has no borderline cases. *Philosophical Studies*, 174, 2105-2123.
- SULLIVAN, P. R. 1995. Contentless consciousness and information-processing theories of mind. *Philosophy, Psychiatry, and Psychology*, 2, 51-59.
- THOMPSON, E. 2007. *Mind in life : biology, phenomenology, and the sciences of mind*, Cambridge, Mass., Belknap Press of Harvard University Press.
- THOMPSON, E. 2015. *Waking, dreaming, being : self and consciousness in neuroscience, meditation, and philosophy*, New York, Columbia University Press.
- THOMPSON, E. 2018. *Biopsychism, Minimal Life, and Sentience* [Online]. Available: <https://psa2018.philsci.org/user-profile/abstract/public/352/biopsychism-minimal-life-and-sentience> [Accessed 30/07/2021].
- TULVING, E. 2005. Episodic memory and autonoesis: Uniquely human? In: TERRACE, H. S. & METCALFE, J. (eds.) *The Missing Link in Cognition*. New York: Oxford University Press.
- VANDEKERCKHOVE, M., BULNES, L. C. & PANKSEPP, J. 2014. The emergence of primary anoetic consciousness in episodic memory. *Front Behav Neurosci*, 7, 210.
- WIESE, W. 2024. Artificial consciousness: a perspective from the free energy principle. *Philosophical Studies*, 181, 1947-1970.
- WOODS, T. J., WINDT, J. M. & CARTER, O. 2024. Evidence synthesis indicates contentless experiences in meditation are neither truly contentless nor identical. *Phenomenology and the Cognitive Sciences*, 23, 253-304.