



HAL
open science

Artificial consciousness. Some logical and conceptual preliminaries

Kathinka Evers, Michele Farisco, Raja Chatila, Brian Earp, Ismael Freire, Fred Hamker, Erik Németh, Paul F. M. J. Verschure, Mehdi Khamassi

► **To cite this version:**

Kathinka Evers, Michele Farisco, Raja Chatila, Brian Earp, Ismael Freire, et al.. Artificial consciousness. Some logical and conceptual preliminaries. 2024. hal-04670602

HAL Id: hal-04670602

<https://hal.science/hal-04670602>

Preprint submitted on 12 Aug 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Artificial consciousness. Some logical and conceptual preliminaries

K. Evers^{1,*}, M. Farisco^{1,2,*}, R. Chatila,³ B. D. Earp,^{4,5} I. T. Freire⁶, F. Hamker⁷, E. Nemeth³, P. F. M. J. Verschure⁶, M. Khamassi³

¹ Centre for Research Ethics and Bioethics, Uppsala University, Uppsala, Sweden

² Biogem Molecular Biology and Genetics Research Institute, Ariano Irpino (AV), Italy

³ Institute of Intelligent Systems and Robotics, Sorbonne University / CNRS, Paris, France

⁴ Uehiro Centre for Practical Ethics, University of Oxford, Oxford, England

⁵ National University of Singapore, Singapore

⁶ Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, The Netherlands

⁷ Artificial Intelligence, Computer Science, Chemnitz University of Technology, Germany

* Shared first authorship

Abstract

Is artificial consciousness theoretically possible? Is it plausible? If so, is it technically feasible? To make progress on these questions, it is necessary to lay some groundwork clarifying the logical and empirical conditions for artificial consciousness to arise and the meaning of relevant terms involved. Consciousness is a polysemic word: researchers from different fields, including neuroscience, Artificial Intelligence, robotics, and philosophy, among others, sometimes use different terms in order to refer to the same phenomena or the same terms to refer to different phenomena.

In fact, if we want to pursue artificial consciousness, a proper definition of the key concepts is required. Here, after some logical and conceptual preliminaries, we argue for the necessity of using *dimensions* and *profiles* of consciousness for a balanced discussion about their possible instantiation or realisation in artificial systems. Our primary goal in this paper is to review the main theoretical questions that arise in the domain of artificial consciousness. On the basis of this review, we propose to assess the issue of artificial consciousness within a multidimensional account.

The theoretical possibility of artificial consciousness is already presumed within some theoretical frameworks; however, empirical possibility cannot simply be deduced from these frameworks but needs independent empirical validation.

We break down the complexity of consciousness by identifying constituents, components, and dimensions, and reflect pragmatically about the general challenges confronting the creation of artificial consciousness. Despite these challenges, we outline a research strategy for showing how "awareness"—as we propose to understand it—could plausibly be realised in artificial systems.

1. Introduction

The possibility of artificial consciousness (roughly, subjective awareness in a human-designed system) has been assumed within certain theoretical frameworks [1]. However, whether artificial consciousness is indeed theoretically possible, much less empirically feasible, are not self-evident, and neither should be taken for granted. Nevertheless, with the rapid progression of relevant technologies, the prospect of producing artificial forms of consciousness is gaining traction in both scientific and public debates, eliciting different and sometimes opposing reactions[2]. The two extremes range from an optimistic enthusiasm emphasising the unavoidable emergence of artificial consciousness on the one hand, to a pressing call for caution, on the other hand, occasionally mixed with scepticism about the feasibility of any attempt to artificially recreate consciousness. A number of alternative views lay in between, each leaning *pro* or *contra* the conceivability, plausibility, feasibility, and (not least) desirability of artificial consciousness on the basis of various theoretical, scientific and socio-ethical arguments[3-13].

In order to reflect on these issues, researchers from different fields have used a number of approaches. These have included: starting from leading scientific theories of consciousness in order to infer relevant indicators of consciousness and eventually check their applicability to current artificially intelligent (AI) systems[3]; theoretically reflecting on the necessary and sufficient conditions for consciousness and their possible instantiation in such systems[4]; philosophically and critically analysing the applicability of notions like intelligence and consciousness to technological artefacts[5]; performing ethical analysis of what the prospect of artificial consciousness, including synthetic phenomenology, would imply for either human subjects or AI systems themselves[6]; reflecting on what conscious machines may entail for society on a descriptive level[7]; identifying reliable indicators for artificial consciousness[14] and relevant tests[11, 15], including a relevant ethical analysis [12]; reflecting on the risks related to the possible confusion about the sentience of AI systems [13]. Therefore, the discussion on artificial consciousness is quite multifaceted and includes different complementary and partly overlapping aspects that are not easy to summarise within a unitary perspective.

The topic of creating artificial consciousness is controversial because it deals with highly sensitive issues and much is at stake: consciousness is a notion extremely prone to anthropocentric and anthropomorphic interpretations, and attributing it to other systems (whether biological or artificial) may raise different reactions, either defensive or prone to acknowledge artificial systems as conscious [16]. These reactions are sometimes triggered by lack of clarity; notably, disproportionate praise or worries about artificial consciousness generated by a lack of clarity about what is actually at stake [17]. In other words, a general and unspecified reference to consciousness as the target of artificial creation is one of the main causes of unrealistic fears, hopes or expectations. Accordingly, a more precise and fine-grained understanding of consciousness, including a more analytical identification of its specific components and dimensions, is necessary for pursuing a balanced and realistic discussion of artificial consciousness, whether through the delineation of a route towards its realisation, or through the identification of obstacles that may either be temporary or fundamentally insurmountable. The need for such a conceptual elaboration also illustrates that the study of consciousness is still in a pre-scientific phase (in the Kuhnian sense of the

coexistence of different theories each claiming its own scientific statute)[18], which asks for further modesty in our approach.

In order to advance in this debate, we consider it crucial to proceed on the basis of a careful and balanced theoretical reflection informed by empirical data. Such a theoretical analysis should initially be as conceptually unbiased (e.g., ideologically, politically, scientifically, and philosophically) and neutral as possible regarding the core questions of the conceivability, plausibility, feasibility, and desirability of artificial consciousness.

Below we review theoretical questions that we suggest are necessary to raise in the domain of artificial consciousness. More specifically, we start by identifying two logical conditions for theoretical reflection about artificial consciousness. Then, we introduce some relevant conceptual clarifications before articulating our view of consciousness as composite, multidimensional, and multilevel and proposing to apply the notions of dimensions and profiles of consciousness, previously introduced for animals and patients with Disorders of Consciousness (DoC), with specific reference to artificial systems.

Ultimately, our proposal is to analyse the complexity of consciousness by identifying its constituents and related components or dimensions, and within this analytic approach to reflect pragmatically on the general challenges confronting the creation of artificial consciousness. Our aim is not to demonstrate conclusively the theoretical possibility or the empirical feasibility of artificial consciousness, but rather to outline a research strategy for determining whether "awareness" is a realistic target for realisation in artificial systems.

2. Logical conditions for advancing in the theoretical reflection on artificial consciousness

For theoretical reflection on artificial consciousness to be effective, it must be characterized by *analytical clarity* and *logical coherence*. Analytical clarity refers to the needed unambiguous explanation of the terms invoked and their reciprocal connections, and requires consistency in the use of terminology. Importantly, the different meanings of the same terms in different contexts (e.g., scientific vs. public debates) should be acknowledged and carefully accounted for in the communication of scientific and technological achievements concerning artificial consciousness, both in general and in any of its specific forms or components (e.g., awareness) in particular. This is especially true for consciousness, which is a highly sensitive issue: as the long discussion about animal consciousness illustrates [19-21], misunderstandings can cause disproportionate reactions, which may arise from passionate and ideologically driven positions rather than from empirically informed, rational reflection [16].

Concerning logical coherence, there are different logical traps and fallacies to be avoided. The most obvious one is what we propose¹ to call the *analytical fallacy*: attempting to deduce a purportedly empirical statement directly from a presupposed theory, which does not satisfy the falsifiability criterion (at least in its "moderate" or "pragmatic" form, insofar as strict

¹ While this error in the present context constitutes a particular instance of errors in general when dealing with syllogisms, which are classically studied in philosophy, we consider it useful to give it a specific name for the present debate, so as to help consciousness scientists recognize and avoid it.

falsifiability may be unattainable)[22] or does not have sufficient empirical validation. In other words, gliding inappropriately between analytical (linguistic, conceptual) and synthetic (empirical) realms. For instance, relying on a specific theoretical framework (e.g., Integrated Information Theory [23], which assumes that the degree of consciousness is equal to the degree of integrated information), one may empirically quantify some complexity-related measures (e.g., integration and differentiation of information in a system) and conclude that they are evidence of the conscious state of a system if above a specific threshold. The risk of the analytical fallacy arises, for instance, if independent empirical validation of the theory is not available and discoveries are assumed to follow directly (i.e., through deduction) from the theoretical framework.

The analytical fallacy eventually results in a circular thinking, surreptitiously presuming what should be proven (i.e., that artificial consciousness is actually possible or even real) rather than specifying in which framework and context (if any) artificial consciousness may be empirically possible or real. This leads to “ironic science” which transcends falsification [24], both in principle and *de facto*.

It is important to note that empirical plausibility (let alone actuality) cannot be inferred from theoretical possibility and logical conceivability: empirical considerations must be added to justify any such inference. The fact that something can be logically conceived (an extremely large set of possibilities, we may note) is a not sufficient condition for its empirical possibility, plausibility and actuality: additional factors, like, in the case of artificial consciousness, the availability of necessary technology, a sufficient understanding of relevant biological processes to possibly emulate, the capacity to translate the principles underlying those processes into technological systems, and empirical indicators supporting the presence of consciousness in an artificial entity, must be taken into account.

3. Terminological clarifications

In addition to the aforementioned logical conditions, a preliminary terminological clarification is necessary for an effective reflection about artificial consciousness. In particular, the notion of consciousness needs clarification, since it is open to different and sometimes not fully compatible or even incompatible understandings.

In the following we summarise some concepts that should be taken into account. We describe some constituents of consciousness, which are arguably multidimensional (Fig. 1). This list is by no means exhaustive: its goal is to introduce some minimal conceptual clarification instrumental to the reflection about artificial consciousness.

3.1 Local vs. global state of consciousness

Local state(s) of consciousness refers to specific conscious experiences, either in terms of processing a particular content or of the phenomenal character of the experiences. In contrast, a global state of consciousness is what characterises an organism’s overall conscious condition [25]. For example, consciously perceiving a red apple or feeling acute pain are illustrations of local states of consciousness, while alert wakefulness and minimally conscious

state (MCS) are illustrations of global states of consciousness, which are characterised by different capacities for experiencing different or the same local states of consciousness [26, 27]. Therefore, while local states of consciousness are distinguished by the objects and the experiences characterising their conscious perception, global states are distinguished by cognitive, behavioural, and physiological differences [28]. In short, local states of consciousness have an explicit link to specific conscious contents that global states of consciousness lack.

3.2 Access vs. phenomenal consciousness

A classical distinction within the philosophy of consciousness is that between access and phenomenal consciousness, that are considered as two different forms of consciousness.

Access consciousness refers to the interaction between different mental states, particularly the availability of one state's content for use in reasoning and rationally guiding capabilities like speech and action; whereas phenomenal consciousness is the subjective feeling of a particular experience, "what it is like to be" in a particular state [29]. More specifically, access consciousness relies on information provided by different cognitive processes mediating functions like working memory, verbal report and motor behaviour [30], while phenomenal consciousness refers to the subjective experience of the conscious subject characterised by a specific point of view.

Interestingly, this distinction is not universally accepted among neuroscientists [31-33]. Some researchers deny the existence of phenomenal consciousness as a specific form separated from access consciousness, and propose to replace phenomenal consciousness with the differentiation of levels of conscious access. On this account, the subject would be able to access the phenomenal contents, but not always to verbally report them [34]. Therefore, contrary to what some philosophers, including Ned Block, have argued [35-37], these researchers do not think that phenomenal consciousness can overflow (i.e., contain more information than) access consciousness [38], but rather distinguish between access consciousness and reportability, and reduce phenomenal consciousness to access consciousness [39].

The discussion is still open about this proposed rejection of the dichotomy between access and phenomenal consciousness, including the fact that, depending on some of the more specific interpretations, this rejection may greatly diminish the number of animal species that can be considered conscious. For one thing, if phenomenal consciousness is reduced to access consciousness, and this is limited to higher cognitive functions, less cognitively complex animal species may be excluded a priori from the realm of consciousness.

Moreover, requiring reportability is arguably anthropocentric and speciesist: it presumes (the necessity of) human language capacities and/or human conceptualisations. Other species may have highly developed forms of consciousness that we have (or maybe lack), that can be reportable in a sense to their kin but not to us (we have for example an extremely poor level and range of perception, so the majority of sensory expressions of other animals are as hidden from us as colours are to the blind and music to the deaf). At the other extreme, conflating

phenomenal and access consciousness and requiring reportability may lead to false positives in the case of artificial systems, which are increasingly able to imitate highly evolved cognitive and communication abilities.

Access consciousness also needs to be separated from other methods of content selection such as attention. While some researchers claim a clear separation between both [40, 41], others are yet more hesitant that present data show attention being easily dissociable from access consciousness [42].

3.3 Primary-minimal consciousness

As part of a graded view of consciousness, the concept of primary or minimal consciousness as opposed to secondary or more advanced consciousness has been proposed in various accounts. One proposal is to conceive minimal consciousness as sentience or 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 [43]. The point being that consciousness is a system feature or configuration that cannot be limited to cognitive high-level features but importantly includes sentience as a more basic, non-cognitive form [44, 45].

The concept of anoetic consciousness as introduced by Endel Tulving is very close to this view [46]. According to him, there are three kinds of consciousness: auto-noetic, 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 [47]. Anoetic consciousness is conceived as the condition of being alive and responsive to stimuli as opposed to having explicit conscious contents [48], 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 " [49] (p. 6).

Others use "primary" [50] or "sensory" [51] 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 [52], 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. [53, 54]).

This concept of primary or minimal consciousness is indeed different from more sophisticated forms of consciousness, like self-awareness (i.e., awareness of ourselves) and meta-cognition (i.e., awareness of being aware).

This distinction between basic and more sophisticated forms of consciousness has been proposed also with reference to phenomenal consciousness, for instance through the

distinction between minimal phenomenal selfhood (i.e., the simplest form of self-consciousness, which may be implementable by subset of all self-related information combined with of all possible self-related behaviours [55]; “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”[56] (p.7)), and a more elaborated phenomenal self-experience [56, 57].

3.4 Contentless consciousness

Contentless consciousness, also called pure consciousness or minimal phenomenal experience, is a form of conscious experience devoid of any specific content [58] 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 [59]. There is discussion about whether this form of consciousness actually exists, and, if it does, what its main features might be [60, 61]. While we acknowledge the reference to this form of consciousness in the literature, we will not discuss it in the present paper.

3.5 Level vs. content of consciousness

For clinical purposes, an operational distinction between two components of consciousness has gained traction in recent years: *level* and *content* of consciousness are identified as the two axes along which it is possible to assess consciousness, more specifically to both quantify it and rate the capacity of the subject to consciously perceive particular objects. Accordingly, level and content of consciousness are identified with wakefulness and awareness respectively, and consciousness is graded within a two-dimensional framework going from coma to sleep to conscious wakefulness [62].

Recently, some have proposed an extended axis including, e.g., psychedelic experiences as being even 'higher' levels of consciousness than 'mere' conscious wakefulness, but this view is controversial[63].

Therefore, in the traditional clinical understanding of consciousness, this results from the combination of awareness with wakefulness. This conceptual distinction is supported by clinical evidence from patients showing a dissociation between wakefulness and awareness, like patients in Vegetative State/Unresponsive Wakefulness Syndrome (VS/UWS), who are awake but not aware. Conversely, REM sleep is considered an example of being aware but not awake.

3.6 First and third person knowledge of consciousness

As mentioned above, in consciousness research there is a fundamental distinction between first person experience and third person access to it [64, 65]. While conscious experiences are defined as such because they pertain to a specific subject who has direct access to them, they cannot be directly experienced by an external subject, that is from a third person perspective.

There are both ontological and epistemic fundamental distinctions: a conscious experience is irreducibly subjective and only the subject who is actually conscious can directly access it [66]. To date, notwithstanding the progress in brain-to-brain interfacing and swarming technologies [67], the knowledge of a conscious experience from a third person perspective can be only inferential.

While the question is open whether AI systems may overcome this epistemic biological/human limitation, for instance through the creation of a cloud-mind that allows direct access to other AI systems' subjective mental states (if any), this epistemic barrier is a challenge from a scientific point of view, especially for a science conceived as grounded on the empirical-experimental method that may not be able to account for first person experiences.

There are different possible answers to this challenge. For instance, it has been proposed to change the scientific paradigm itself in order to also include subjective first person experiences as a legitimate object of scientific inquiry in addition to objective third person (i.e., quantitative) information [68]. Another proposal is to engineer consciousness so that the ability to know and manipulate all its factors and variables makes it possible to access first person experience [69].

3.7 The conceptual prism of consciousness

The selected notions and approaches presented above are not comprehensive but sufficient to introduce the high level of controversy surrounding consciousness, including its scientific understanding and its possible replication in AI systems.

To summarise (see Fig. 1), the term consciousness may refer to cognitive information processing or to subjective experience (even if this is not unanimously accepted), to the state of the subject and the level of the subject's consciousness (e.g., awake vs. unawake) or to the content of the conscious experience (i.e., awareness). Both the state/level (wakefulness) and the content (awareness) are considered as two fundamental components of consciousness.

We may say that as a content-oriented representation (i.e., awareness), consciousness is intentional (i.e., directed to something), while level and state (i.e., wakefulness) refer to the preliminary capacity for this representational process, and they may or may not eventually correlate with aware consciousness. The clinical case of vegetative state/unresponsive wakefulness syndrome (VS/UWS), defined as wakefulness without awareness [70, 62], as well as dream states, where the subject is aware but unawake [71], are illustrative of the potential dissociation between these two components of consciousness. Importantly, the conscious content resulting from the capacity to process information (i.e., to be aware, either in awake or non awake state) is not the same as the conscious experience or sentience, which connects to the phenomenal, subjective form of consciousness (i.e., what it is like to be in a specific state, [29, 72]).

Also, the relation between consciousness and the self is important to clarify. Self-consciousness is one possible component of conscious experience, but a robust and reflective self-perception is not necessary for conscious experience in general. In fact, self-

consciousness is multilevel [73]. It is possible for a subject to be aware and also to have a minimal phenomenal experience [58] even if lacking a strong reflective sense of self, that is a self-oriented meta-cognitive capacity, both among non-human animals (e.g., dolphins, octopuses, crows, or bonobos) and humans (e.g., infants, and adults in psychedelic experiences) [74, 14, 75, 76, 21, 77-80, 28, 81].

Thus, the definition of consciousness is a most challenging task. A universal agreement about it is hardly achievable, mainly due to the myriad different theoretical models and related definitions. These distinct theories are not necessarily commensurable, or they can be so in different ways, and the usefulness of common denominators in differentiating, integrating and testing hypotheses has recently been analysed [18]. There are also attempts to elaborate a unifying model of consciousness [82, 83] or even to implement an adversarial collaboration among different theories [84], yet the question is open about how to advance towards a more mature science of consciousness, including more robust agreement about its definition [85]. A possible strategy in this direction is considering the different theories as complementary rather than adversarial or alternative to each other [69].

We suggest that it is not necessary to agree about an overarching, general definition of consciousness for reflecting about the possibility and plausibility of artificial consciousness arising. In fact, there are two alternatives: to agree on a working or a stipulative definition in order to advance towards a sufficient level of agreement and mutual understanding, especially between scientific researchers and people from other disciplines (e.g., social and political science, ethics, philosophy), and also from the general public; or to agree on what are at least some of the general features that characterise consciousness, beyond the specific theoretical stance one endorses.

This approach is consistent with the “theory-light” approach as recently described by Birch as a methodology that relies on a minimal commitment about the relation between (phenomenal) consciousness and cognition, so that it does not subscribe to any specific theory of consciousness [86]. The core hypothesis is that conscious perception of a stimulus facilitates a cluster of cognitive abilities in relation to that stimulus. In the following, we follow this direction, introducing some components and dimensions of consciousness that can arguably be considered characteristic to it.

4. Consciousness is composite, multidimensional, and multilevel

On the basis of the terminological clarification introduced above, it is reasonable to infer that consciousness presents different constituents (i.e., states, forms, components, and dimensions), as reflected in the different senses of the term. For instance, wakefulness and awareness are two fundamental components of consciousness that are particularly relevant in the clinical context. The same with two fundamental forms of consciousness like access consciousness and phenomenal consciousness, provided that one agrees with the existence of the latter. Moreover, each constituent of consciousness (both as a cognitive appraisal and as a phenomenological state) is arguably multidimensional.

More specifically, Bayne, Hohwy, and Owen [25] argue that global states of consciousness manifest themselves in multiple ways, and that the notion of levels should be replaced by that of dimensions of consciousness to properly describe it. The central thesis is that global states of consciousness are not gradable along one dimension, but rather distinguished along different dimensions. More specifically, they introduce two main families of consciousness' dimensions: content-related and functional.

The first family includes, for instance, gating of conscious content (e.g., low-level features vs high-level features of an object). The second family includes, for instance, cognitive and behavioural control (i.e., the availability of conscious contents for control of thought and action). Along this line of analysis, Walter has recently proposed the following content-related dimensions: sensory richness, high-order object representation, semantic comprehension; and the following functional dimensions: executive functioning, memory consolidation, intentional agency, reasoning, attention control, vigilance, meta-awareness [87].

Other relevant reflections about consciousness' dimensions come from Birch et al., who with reference to animal consciousness specifically introduce the following dimensions [76]:

- *Perceptual-Richness*: any measure is specific to a sense modality, so there is no overall level of perceptual richness. Also, within a particular sense modality, perceptual richness can be resolved into different components (e.g., bandwidth, acuity, and categorization power for vision);
- *Evaluative-Richness*: affectively-based positive or negative valence which grounds decision-making. Also evaluative richness can be resolved into different components;
- *Integration at a time (unity)*: conscious experience is (usually) highly unified;
- *Integration across time (temporality)*: conscious experience takes the form of a continuous stream;
- *Self-consciousness (Selfhood)*: awareness of oneself as distinct from the world outside.

Dung and Newen have introduced additional dimensions, again with explicit reference to animal consciousness, but potentially relevant for AI consciousness as well [77]. Within the category "external representation", where they include Perceptual-Richness and Evaluative-Richness as defined by Birch et al., they add Evaluative-Intensity, defined as how strongly a subject feels the positive or negative valence of an object/experience, and the external diachronic and synchronic unity.

Within the category "self-representation", they add self-referred diachronic and synchronic unity, experience of agency (i.e., the ability to experience actions as voluntarily initiated and controlled), and the experience of ownership (i.e., the ability to perceive body parts as something personal rather than objects of the external world). Within the category "cognitive processing strategies" they introduce three new dimensions: reasoning (e.g., complex trains of thought and ability to reason on multiple domains), learning (e.g., trace conditioning), and abstraction (i.e., the ability to form and use high-level abstract forms that categorise specific sensory stimuli).

Irwin has recently proposed another approach to animals' consciousness dimensions: on the basis of a behavioural study of twelve animal species, he identified three kinds of behaviour (volitional, interactive, and egocentric), quantified their frequency, variety, and dynamism, and eventually represented them in a matrix indicative of the consciousness profile of the animal in question [88].

All these attempts are illustrative of a highly lively debate that promises further advancement toward a more fine-grained and analytical reflection about consciousness and its constituents. Of course, the dimensions listed above cover some aspects of the prism of consciousness while others remain less or not considered. For instance, another category of dimensions that appears not adequately addressed so far is social-relational functions or representations [89] and dyadic interactions, which include dimensions or capacities such as theory of mind (i.e., the ability to anticipate through a model-based virtualization the behaviour of others, particular when instrumental to fulfilling personal goals), strategic collaboration (i.e., collaborating with others because it is instrumental to fulfil shared goals, even if particular benefits will be eventually reduced as a consequence of such collaboration, or the benefit is not immediate but postponed), and altruistic (or "communal") [90] orientations or behaviour (i.e., proneness to share resources if others are detected as in need, even if this sharing does not produce any personal benefit or raises the risk of reducing personal wellbeing).

While the detailed identification of further dimensions of the above mentioned families and categories, as well as the identification of other possible families and/or categories of consciousness' dimensions, are still an open issue [91, 87], the concept of *consciousness profiles* emerges as the spaces of experience delimited by different specific dimensions within one or more constituents of consciousness.

Accordingly, we can differentiate consciousness profiles not in terms of their overall levels along one and the same dimension, but rather with reference to the combination of the different dimensions that characterise them (See Fig. 2 for a speculative illustration of the comparison between human and non-human consciousness profiles). For instance, it may well be the case that the consciousness profile of a human subject has some content-related and functional dimensions (e.g., semantic comprehension and meta-awareness, respectively) more advanced than a non-human entity, while other content-related and functional dimensions (e.g., sensory richness and vigilance, respectively) may be less advanced. Also, it is possible that a non-human entity (either biological or artificial) has a consciousness profile which includes some dimensions that humans lack (e.g., echolocation). This does not mean that one overall conscious state is higher or lower than the other, but rather that it is differently shaped.

Therefore, the comparison between human, other animals, and potential artificial consciousnesses should be framed in terms of resemblances and differences along specific constituents and related dimensions rather than in terms of higher or lower levels along only one, overall constituent and/or dimension. In short, consciousness is a multifaceted reality (i.e., a prism), irreducible to one level of description.

To summarize, consciousness is a complex feature: different constituents define it, and these constituents have different dimensions. In addition, the notion of ‘level’ can be understood in a way that is compatible with this composite and multi-dimensional view of consciousness. In principle, a level of consciousness may indicate the grade of the global state consciousness (i.e., a rank along the same scale) or the specific form of consciousness that the subject is capable of (i.e., a differentiation among more or less sophisticated forms of consciousness). This second meaning is compatible with the framework depicted above.

To illustrate, the following levels (i.e., forms) of consciousness have been proposed as hierarchically nested [92]:

- *minimal consciousness*, “characterized by the capacity to display spontaneous motor activity and to create representations, for instance, from visual and auditory experience, to store them in long-term memory and use them, for instance, for approach and avoidance behaviour and for what is referred to as exploratory behaviour.” [92](p. 2240)
- *recursive consciousness*, characterised “by functional use of objects and by proto-declarative pointing; [...] elaborate social interactions, imitation, social referencing and joint attention; [...] the capacity to hold several mental representations in memory simultaneously, and [...] to evaluate relations of self; [...] elementary forms of recursivity in the handling of representations, yet without mutual understanding” [92](p. 2240)
- *explicit self-consciousness*, “characterized by self recognition in mirror tests and by the use of single arbitrary rules with self–other distinction”; [92] (p. 2240)
- *reflective consciousness*, which entails “theory of mind and full conscious experience, with first person ontology and reportability”. [92] (p. 2240)

These different levels of consciousness may have different phylogenetic and ontogenetic instantiations: different species may display one or more levels, and the same organism may reach different levels during its development. Also, the question is open about which level of consciousness may be attributed to AI, at present, at mid-, and at long-term future. In conclusion, since consciousness is a composite and multilevel concept, it is necessary for any attempt to replicate it to specify which specific constituent (i.e., states, forms, components, and dimensions) is the target. We think that these clarifications and distinctions pave the way for a new approach to artificial consciousness which does not restrict itself to binary thinking (conscious versus non-conscious systems), nor to a single unidimensional level of consciousness (minimal versus high level of consciousness)[93].

5. Awareness as a target of artificial realisation

As written above, the present paper does not aim to propose an overall definition of consciousness, but rather to clarify the logical conditions and to provide some basic conceptual clarifications for advancing in the discussion about the theoretical plausibility and the technical feasibility of artificial consciousness. It also aims at paving a concrete way where benchmarks can be established to assess specific dimensions, degrees and profiles of artificial

consciousness. As part of this logical and conceptual reflection, the notions of consciousness' constituents, components, dimensions, and profiles have been introduced.

Against this background, we now focus on awareness as a particular constituent of consciousness that contemporary AI may - or may not - succeed in instantiating. This may seem insufficiently ambitious or too modest an approach, but we consider this a pragmatic and reasonable strategy to handle the complexity of consciousness as summarised above. We also think that this kind of approach is promising for advancing the discussion in a balanced and realistic manner.

As seen above, in the clinical context, awareness is defined as the content-related component of conscious experience, in addition to wakefulness (or the level of vigilance). In other words, awareness is the capacity of the subject to process information, store it in short-term memory, and possibly retrieve it from long-term memory if needed.

In fact, there are several sets of empirical data about the neuronal mechanisms of this constituent of consciousness [62, 94], its artificial realisation seems (at least intuitively) less controversial than the artificial realisation of subjective experience (cognition and action control appear more prone to computational interpretation and replication, specially beyond academic circles), and artificial awareness is a concept relatively easy to grasp and to accept also for non-technical audience.

According to the multidimensional framework introduced above, to be qualified as conscious, the capacity to process, store, and retrieve information that characterises awareness as defined in the clinical context should present different levels of both the content-related and functional dimensions that shape the consciousness profiles. Among those dimensions we here assume that the intentional use of information for achieving specific goals stands as a minimal necessary condition² for aware processing.

Thus, there are two dimensions of awareness, which are both necessary for a system to be actually aware: the capacity for an evaluative processing of information (i.e., selecting relevant inputs on the basis of current needs) combined with the capacity for intentionally using it (i.e., identifying and making use of affordances in the surrounding environment). Intentional use of information relies on the explicit knowledge that subjective behaviour is instrumental to get desired goals [95].

This minimal definition of awareness is open to different potential technical implementations. In fact, in order to be considered as one dimension of awareness, information processing should be more sophisticated than a model-free phenomenon and can go from basic levels when an agent has the capacity for modelling internal and external states, to higher levels when these models are combined with the capacity to virtualize the world and to predict future states [96, 97]. To be markers of awareness, these capacities for modelling and

² We qualify intentionality as the minimal necessary condition for awareness because we think that the intentional use of information for achieving specific goals is crucial for distinguishing between aware and unaware cognition. At the same time, we do not exclude that other dimensions of awareness as listed above are also present, even if they are not minimally necessary.

virtualization should be combined with the capacity to intentionally exploit them as part of a goal-directed behaviour.

Also, there are important aspects that this minimal definition of awareness leaves open, including the role of reward-based expectation for awareness (e.g., for selecting information the system is actually aware of [98, 99]) as well as the possible connection between feedforward and feedback dynamics in the system [96]. Furthermore, the connection between awareness and general intelligence, as well as between awareness and understanding remain open to different interpretations.

Another aspect of awareness that has been revealed by clinical research is the dissociation between internal or self-awareness (i.e., relative to the self) and external or sensory awareness [100]. Significantly, different networks for each of them have been identified (midline fronto-parietal and lateral fronto-parietal networks, respectively) [101, 102]. This confirms that consciousness does not require or imply self-consciousness.

6. Discussion

As previously noted, we do not presume to provide a definitive answer to the questions whether artificial awareness could arise or how likely this development is. The answers to both questions depend on the background theoretical framework as well as on the technology actually available. For instance, the Distributed Adaptive Control Theory of consciousness assumes that artificial awareness is at least theoretically possible, setting the ground for the technological attempt to translate this possibility in reality [96]. In other words, the possibility of artificial awareness is assumed as a working hypothesis, which plays the role of a heuristic program inspiring empirical work towards its validation.

Being a component of consciousness, awareness does not exhaust its semantic and functional complexity. Even if limited, we propose to take awareness as a specific target for the attempt to produce conscious capacities in AI systems because awareness appears open to a more intuitive understanding and to a wider conceptual consensus than the general and sometimes opaque notion of consciousness, and the empirical investigation of the specific cerebral underpinnings of awareness is quite advanced. Even if the fundamental question about the theoretical plausibility of artificial awareness is still open, we propose that focusing on awareness may avoid the risk of a conceptual impasse and proceed more pragmatically towards the realisation of selected aspects of biological consciousness.

There are several issues that the approach we have introduced above still leaves open. While all merit further investigation, we here provide an illustrative list, which is by no means exhaustive, but only for the sake of introducing further important points of analysis.

A fundamental issue is why to pursue artificial awareness in the first place: What would be the resulting benefits and advantages, for instance for science, or society at large? A possibility is that by building artificial awareness we will eventually better understand biological consciousness. Another possibility is that building artificial awareness will be a game-changer in AI. In fact, the capacity to build world models is arguably an important factor

for the further advancement of AI [103]. Artificial awareness would allow AI to intentionally use the world models it develops, resulting in both a significant improvement of AI technology and an important impact on society. Moreover, being aware of the consequences of its actions could help AI better inform humans about potential negative impacts on society, and help avoid them while favouring positive impacts. Yet the question is open about how more specifically a possible artificial awareness will have such impacts.

Another fundamental issue concerns the nature of a hypothetical artificial awareness: Is embodiment necessary for it? In other words, does awareness require an embodied subject, embedded in a particular environment (i.e., *Umwelt*), in relation to which it develops and makes intentional use of models for satisfying its needs? To address this question about embodiment, one has to consider the origin of the information that must flow from the real world through a sensory system to be processed, and an interpretation of this information must be made to produce action. Any affordance would have to be built on the possibility of physical action. This argues in favour of embodiment, but the possible forms of embodiment need to be clarified.

Finally, the issue of values emerges as very challenging. For biological organisms, awareness is intrinsically related to the capacity to evaluate the world, discriminating between what is good and what is bad [104-106]. Is it the same for a possible artificial awareness? Or would the absence or the different nature of what makes values and evaluation necessary in biological awareness (e.g., emotions, reward-systems, preferences) eventually allow an artificial non-evaluative awareness? If so, would this be desirable? Values of different kinds (e.g., moral, political, religious) have inspired both positive and negative actions in human history, so the question is open about the moral implications of an aware agent devoid of any value.

Conclusion

In this paper we reviewed some theoretical issues (both logical and conceptual) that we think are crucial to investigate in order to advance in the clarification of the plausibility and the feasibility of artificial consciousness. We argued that a multidimensional view of consciousness, as recently introduced in the domain of Disorders of Consciousness and animal consciousness, is a fruitful framework for analysing these points. Within this framework, we propose to pragmatically focus on one specific component of consciousness (i.e., awareness) in the attempt to replicate it in AI systems. The question remains open about the technical feasibility of this replication (which being an empirical issue may be a matter of time), as well as about the theoretical plausibility of replicating other forms of consciousness (e.g., sentience). Notwithstanding these still open issues, we suggest that the approach presented above is a promising methodological model for advancing in a balanced and informed discussion.

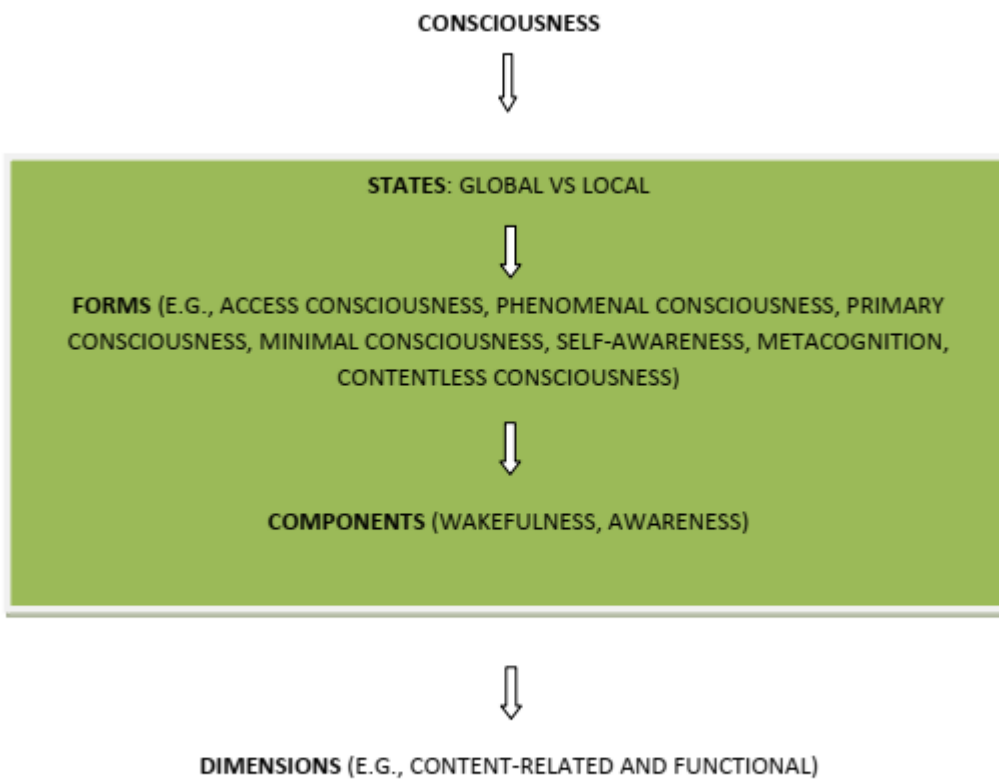


Fig. 1. Multidimensional constituents of consciousness.

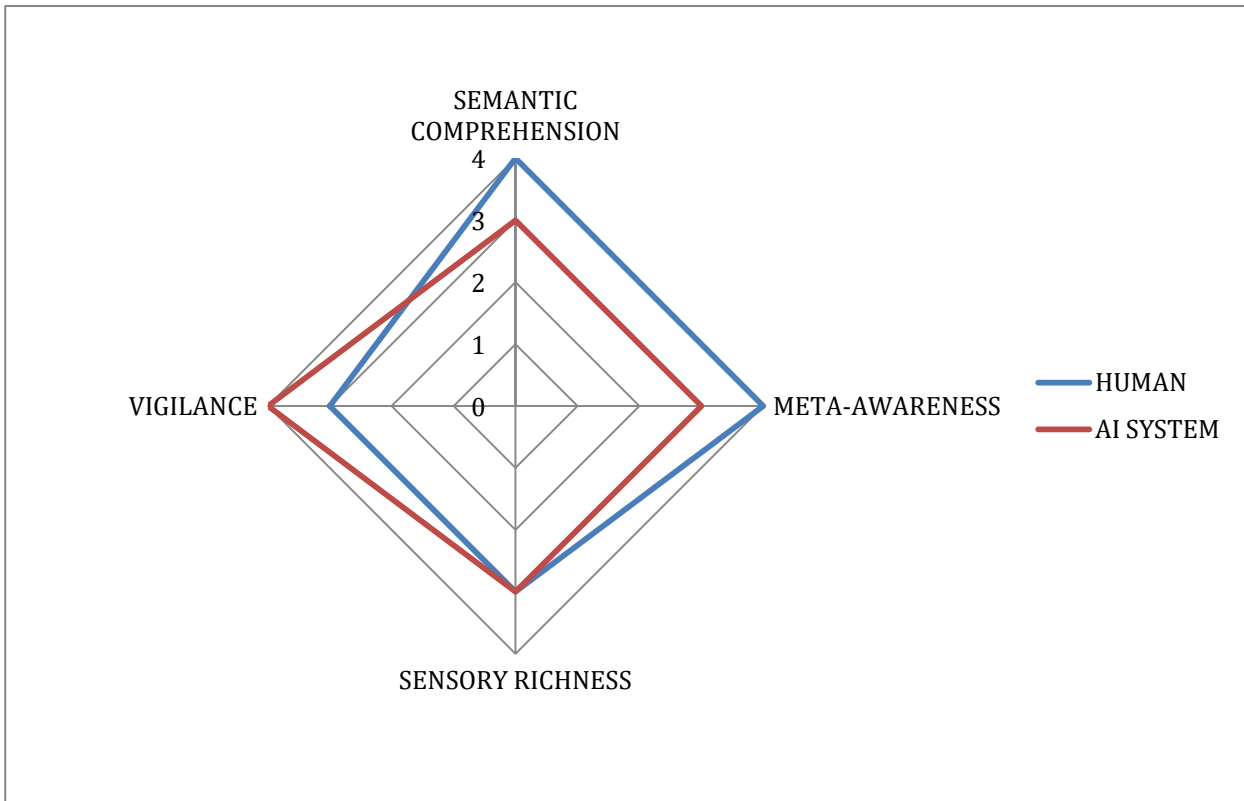


Fig. 2. Illustrative comparison of two hypothetical human and AI system's consciousness profiles. The values assigned are speculative and for the only sake of illustration. The human and AI system's consciousness profiles are represented by the blue and red diamonds respectively. This illustration is based on [76], who apply the same approach to animal consciousness.

Acknowledgements

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

References

1. Della Santina C, Corbato CH, Sisman B, Leiva LA, Arapakis I, Vakalellis M et al. Awareness in robotics: An early perspective from the viewpoint of the EIC Pathfinder Challenge "Awareness Inside". 2024.
2. Lenharo M. AI consciousness: scientists say we urgently need answers. *Nature*. 2024;625(7994):226. doi:10.1038/d41586-023-04047-6.
3. Butlin P, Long R, Elmoznino E, Bengio Y, Birch J, Constant A et al. *Consciousness in Artificial Intelligence: Insights from the Science of Consciousness* 2023.
4. Chalmers D. *Could a Large Language Model be Conscious?* 2023.
5. Dietrich E, Fields C, Sullins JP, Van Heuveln B, Zebrowski R. *Great philosophical objections to artificial intelligence : the history and legacy of the AI wars*. London ; New York: Bloomsbury Academic; 2021.
6. Hildt E. The Prospects of Artificial Consciousness: Ethical Dimensions and Concerns. *AJOB Neurosci*. 2023;14(2):58-71. doi:10.1080/21507740.2022.2148773.
7. LeDoux, Birch J, Andrews K, Clayton NS, Daw ND, Frith C et al. Consciousness beyond the human case. *Current Biology*. 2023;33(16):R832-R40. doi:<https://doi.org/10.1016/j.cub.2023.06.067>.
8. Marcus G. *Sentient AI: For the love of Darwin, let's Stop to Think if We Should*. 2023.
9. Metzinger T. An Argument for a Global Moratorium on Synthetic Phenomenology. *Journal of Artificial Intelligence and Consciousness*. 2021;8(1):1-24.
10. Seth AK. *361 Consciousness. Living machines: A handbook of research in biomimetics and biohybrid systems*: Oxford University Press; 2018. p. 0.
11. Bayne T, Seth AK, Massimini M, Shepherd J, Cleeremans A, Fleming SM et al. Tests for consciousness in humans and beyond. *Trends in cognitive sciences*. 2024. doi:10.1016/j.tics.2024.01.010.
12. Farisco M. The ethical implications of indicators of consciousness in artificial systems. *Developments in Neuroethics and Bioethics*. Academic Press; 2024.
13. Schwitzgebel E. AI systems must not confuse users about their sentience or moral status. *Patterns (New York, NY)*. 2023;4(8):100818. doi:10.1016/j.patter.2023.100818.
14. Pennartz, Farisco M, Evers K. Indicators and Criteria of Consciousness in Animals and Intelligent Machines: An Inside-Out Approach. *Frontiers in systems neuroscience*. 2019;13:25. doi:10.3389/fnsys.2019.00025.
15. Elamrani A, Yampolsky RV. Reviewing Tests for Machine Consciousness. *Journal of Consciousness Studies*. 2019;26(5-6):35-64.
16. Colombatto C, Fleming SM. *Folk Psychological Attributions of Consciousness to Large Language Models*. 2023.
17. Zebrowski RL. Fear of a Bot Planet: Anthropomorphism, Humanoid Embodiment, and Machine Consciousness. *Journal of Artificial Intelligence and Consciousness*. 2020;07(01):119-32. doi:10.1142/s2705078520500071.
18. Evers K, Farisco M, Pennartz CMA. Assessing the commensurability of theories of consciousness: On the usefulness of common denominators in differentiating, integrating and testing hypotheses. *Consciousness and Cognition*. 2024;119:103668. doi:<https://doi.org/10.1016/j.concog.2024.103668>.
19. Low P. *The Cambridge Declaration on Consciousness*. Proceedings of the Francis Crick Memorial Conference. Cambridge: Cambridge University Press; 2012. p. 1-2.
20. Carruthers P. *Human and animal minds : the consciousness questions laid to rest*. First edition ed. Oxford ; New York, NY: Oxford University Press; 2019.
21. Crump A, Birch J. Animal Consciousness: The Interplay of Neural and Behavioural Evidence. *Journal of Consciousness Studies*. 2022;29(3-4):104-28.
22. Earp BD, Trafimow D. Replication, falsification, and the crisis of confidence in social psychology. *Front Psychol*. 2015;6:621. doi:10.3389/fpsyg.2015.00621.
23. Albantakis, Barbosa L, Findlay G, Grasso M, Haun AM, Marshall W et al. Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms. *PLoS Comput Biol*. 2023;19(10):e1011465. doi:10.1371/journal.pcbi.1011465.
24. Merker B, Williford K, Rudrauf D. The Integrated Information Theory of consciousness: A case of mistaken identity. *Behav Brain Sci*. 2021:1-72. doi:10.1017/s0140525x21000881.

25. Bayne T, Hohwy J, Owen AM. Are There Levels of Consciousness? Trends in cognitive sciences. 2016;20(6):405-13. doi:10.1016/j.tics.2016.03.009.
26. Bruno MA, Vanhaudenhuyse A, Thibaut A, Moonen G, Laureys S. From unresponsive wakefulness to minimally conscious PLUS and functional locked-in syndromes: recent advances in our understanding of disorders of consciousness. Journal of neurology. 2011;258(7):1373-84. doi:10.1007/s00415-011-6114-x.
27. Thibaut A, Panda R, Annen J, Sanz LRD, Naccache L, Martial C et al. Preservation of Brain Activity in Unresponsive Patients Identifies MCS Star. Annals of neurology. 2021;90(1):89-100. doi:10.1002/ana.26095.
28. Rankaduwa S, Owen AM. Psychedelics, entropic brain theory, and the taxonomy of conscious states: a summary of debates and perspectives. Neuroscience of Consciousness. 2023;2023(1). doi:10.1093/nc/niad001.
29. Block N. On a confusion about a function of consciousness. Behavioral and Brain Sciences. 1995;18(2):227-87.
30. Mashour GA, Roelfsema P, Changeux JP, Dehaene S. Conscious Processing and the Global Neuronal Workspace Hypothesis. Neuron. 2020;105(5):776-98. doi:10.1016/j.neuron.2020.01.026.
31. Schier E. Identifying phenomenal consciousness. Conscious Cogn. 2009;18(1):216-22. doi:10.1016/j.concog.2008.04.001.
32. Kouider S, de Gardelle V, Sackur J, Dupoux E. How rich is consciousness? The partial awareness hypothesis. Trends in cognitive sciences. 2010;14(7):301-7. doi:10.1016/j.tics.2010.04.006.
33. Baars BJ, Laureys S. One, not two, neural correlates of consciousness. Trends in cognitive sciences. 2005;9(6):269; author reply 70. doi:10.1016/j.tics.2005.04.008.
34. Dehaene S, Changeux JP, Naccache L, Sackur J, Sergent C. Conscious, preconscious, and subliminal processing: a testable taxonomy. Trends in cognitive sciences. 2006;10(5):204-11. doi:10.1016/j.tics.2006.03.007.
35. Block N. Perceptual consciousness overflows cognitive access. Trends in cognitive sciences. 2011;15(12):567-75. doi:10.1016/j.tics.2011.11.001.
36. Block N. Rich conscious perception outside focal attention. Trends in cognitive sciences. 2014;18(9):445-7. doi:10.1016/j.tics.2014.05.007.
37. Bronfman ZZ, Jacobson H, Usher M. Impoverished or rich consciousness outside attentional focus: Recent data tip the balance for _Overflow_. Mind and Language. 2019;34(4):423-44.
38. Brown SAB. How to get rich from inflation. Consciousness and Cognition. 2024;117:103624. doi:<https://doi.org/10.1016/j.concog.2023.103624>.
39. Naccache L. Why and how access consciousness can account for phenomenal consciousness. Philosophical Transactions of the Royal Society B: Biological Sciences. 2018;373(1755):20170357. doi:doi:10.1098/rstb.2017.0357.
40. Koch, Tsuchiya N. Attention and consciousness: related yet different. Trends in cognitive sciences. 2012;16(2):103-5.
41. Nani A, Manuello J, Mancuso L, Liloia D, Costa T, Cauda F. The Neural Correlates of Consciousness and Attention: Two Sister Processes of the Brain. Front Neurosci. 2019;13:1169. doi:10.3389/fnins.2019.01169.
42. Cohen MA, Cavanagh P, Chun MM, Nakayama K. The attentional requirements of consciousness. Trends in cognitive sciences. 2012;16(8):411-7. doi:10.1016/j.tics.2012.06.013.
43. Bronfman ZZ, Ginsburg S, Jablonka E. The Transition to Minimal Consciousness through the Evolution of Associative Learning. Front Psychol. 2016;7:1954. doi:10.3389/fpsyg.2016.01954.
44. Pereira Jr A. The role of sentience in the theory of consciousness and medical practice. Journal of Consciousness Studies. 2021;28(7-8):22-50.
45. Baluška F, Reber A. Sentience and Consciousness in Single Cells: How the First Minds Emerged in Unicellular Species. Bioessays. 2019;41(3):e1800229. doi:10.1002/bies.201800229.
46. Tulving E. Episodic memory and auto-noesis: Uniquely human? In: Terrace HS, Metcalfe J, editors. The Missing Link in Cognition. New York: Oxford University Press; 2005. p. 4-56.
47. LeDoux, Lau H. Seeing consciousness through the lens of memory. Curr Biol. 2020;30(18):R1018-R22. doi:10.1016/j.cub.2020.08.008.
48. LeDoux JE. What emotions might be like in other animals. Curr Biol. 2021;31(13):R824-R9. doi:10.1016/j.cub.2021.05.005.

49. Vandekerckhove M, Bulnes LC, Panksepp J. The emergence of primary anoetic consciousness in episodic memory. *Front Behav Neurosci.* 2014;7:210. doi:10.3389/fnbeh.2013.00210.
50. Edelman GM. Naturalizing consciousness: a theoretical framework. *Proc Natl Acad Sci U S A.* 2003;100(9):5520-4. doi:10.1073/pnas.0931349100.
51. Feinberg TE, Mallatt J. The nature of primary consciousness. A new synthesis. *Conscious Cogn.* 2016;43:113-27. doi:10.1016/j.concog.2016.05.009.
52. Feinberg TE, Mallatt J. *Consciousness demystified.* Cambridge, Massachusetts ; London, England: MIT Press; 2018.
53. Changeux J-P. *Neuronal man : the biology of mind / translated by Laurence Garey.* New York ; Oxford: Oxford University Press; 1986.
54. Changeux J-P. *The physiology of truth : neuroscience and human knowledge.* Cambridge, Mass.: Belknap Press of Harvard University Press; 2004.
55. Forch V, Hamker FH. Building and Understanding the Minimal Self. *Front Psychol.* 2021;12:716982. doi:10.3389/fpsyg.2021.716982.
56. Blanke O, Metzinger T. Full-body illusions and minimal phenomenal selfhood. *Trends in cognitive sciences.* 2009;13(1):7-13. doi:10.1016/j.tics.2008.10.003.
57. Seth AK, Tsakiris M. Being a Beast Machine: The Somatic Basis of Selfhood. *Trends in cognitive sciences.* 2018;22(11):969-81. doi:10.1016/j.tics.2018.08.008.
58. Metzinger T. Minimal phenomenal experience. *Philosophy and the Mind Sciences.* 2020;1(1):1-44.
59. Thompson E. *Waking, dreaming, being : self and consciousness in neuroscience, meditation, and philosophy.* New York: Columbia University Press; 2015.
60. Sullivan PR. Contentless consciousness and information-processing theories of mind. *Philosophy, Psychiatry, and Psychology.* 1995;2(1):51-9.
61. Woods TJ, Windt JM, Carter O. Evidence synthesis indicates contentless experiences in meditation are neither truly contentless nor identical. *Phenomenology and the Cognitive Sciences.* 1-52.
62. Laureys S. The neural correlate of (un)awareness: lessons from the vegetative state. *Trends in cognitive sciences.* 2005;9(12):556-9. doi:10.1016/j.tics.2005.10.010.
63. Bayne T, Carter O. Dimensions of consciousness and the psychedelic state. *Neurosci Conscious.* 2018;2018(1):niy008. doi:10.1093/nc/niy008.
64. Loar B. Phenomenal states. *Philosophical Perspectives.* 1990;4:81-108.
65. Earp BD. An anti-anti-functional account of consciousness. *Annales Philosophici.* 2012;4:6-15.
66. Evers K, Sigman M. Possibilities and limits of mind-reading: A neurophilosophical perspective. *Consciousness and Cognition.* 2013;22:887-97.
67. Lyreskog DM, Zohny H, Savulescu J, Singh I. Merging Minds: The Conceptual and Ethical Impacts of Emerging Technologies for Collective Minds. *Neuroethics.* 2023;16(1):12. doi:10.1007/s12152-023-09516-3.
68. Goff P. *Galileo's error : foundations for a new science of consciousness.* New York: Pantheon Books; 2019.
69. Verschure P. Escaping from the IIT Munchausen method: Re-establishing the scientific method in the study of consciousness. *Behav Brain Sci.* 2022;45:e63. doi:10.1017/S0140525X21002028.
70. Laureys S, Celesia GG, Cohadon F, Lavrijzen J, Leon-Carrion J, Sannita WG et al. Unresponsive wakefulness syndrome: a new name for the vegetative state or apallic syndrome. *BMC Med.* 2010;8:68. doi:10.1186/1741-7015-8-68.
71. Siclari F, Baird B, Perogamvros L, Bernardi G, LaRocque JJ, Riedner B et al. The neural correlates of dreaming. *Nat Neurosci.* 2017;20(6):872-8. doi:10.1038/nn.4545.
72. Nagel T. What is it like to be a bat? *The Philosophical Review.* 1974;83(4):435-50.
73. Millièrè R. Are there degrees of self-consciousness? *Journal of Consciousness Studies.* 2019;26(3-4):252-76.
74. Butler AB, Cotterill RM. Mammalian and avian neuroanatomy and the question of consciousness in birds. *Biol Bull.* 2006;211(2):106-27. doi:10.2307/4134586.
75. Irwin LN. Renewed Perspectives on the Deep Roots and Broad Distribution of Animal Consciousness. *Frontiers in systems neuroscience.* 2020;14:57. doi:10.3389/fnsys.2020.00057.
76. Birch J, Schnell AK, Clayton NS. Dimensions of Animal Consciousness. *Trends in cognitive sciences.* 2020;24(10):789-801. doi:10.1016/j.tics.2020.07.007.

77. Dung L, Newen A. Profiles of animal consciousness: A species-sensitive, two-tier account to quality and distribution. *Cognition*. 2023;235:105409. doi:10.1016/j.cognition.2023.105409.
78. Zacks O, Jablonka E. The evolutionary origins of the Global Neuronal Workspace in vertebrates. *Neurosci Conscious*. 2023;2023(2):niad020. doi:10.1093/nc/niad020.
79. Carhart-Harris RL, Leech R, Hellyer PJ, Shanahan M, Feilding A, Tagliazucchi E et al. The entropic brain: a theory of conscious states informed by neuroimaging research with psychedelic drugs. *Front Hum Neurosci*. 2014;8:20. doi:10.3389/fnhum.2014.00020.
80. Carhart-Harris RL. The entropic brain - revisited. *Neuropharmacology*. 2018;142:167-78. doi:10.1016/j.neuropharm.2018.03.010.
81. Lagercrantz H, Changeux JP. The emergence of human consciousness: from fetal to neonatal life. *Pediatr Res*. 2009;65(3):255-60. doi:10.1203/PDR.0b013e3181973b0d.
82. Wiese W. The science of consciousness does not need another theory, it needs a minimal unifying model. *Neurosci Conscious*. 2020;2020(1):niaa013. doi:10.1093/nc/niaa013.
83. Northoff G, Lamme V. Neural signs and mechanisms of consciousness: Is there a potential convergence of theories of consciousness in sight? *Neurosci Biobehav Rev*. 2020;118:568-87. doi:10.1016/j.neubiorev.2020.07.019.
84. Yaron I, Melloni L, Pitts M, Mudrik L. The ConTraSt database for analysing and comparing empirical studies of consciousness theories. *Nat Hum Behav*. 2022;6(4):593-604. doi:10.1038/s41562-021-01284-5.
85. Michel M, Beck D, Block N, Blumenfeld H, Brown R, Carmel D et al. Opportunities and challenges for a maturing science of consciousness. *Nat Hum Behav*. 2019;3(2):104-7. doi:10.1038/s41562-019-0531-8.
86. Birch J. The search for invertebrate consciousness. *Noûs*. 2022;56(1):133-53. doi:<https://doi.org/10.1111/nous.12351>.
87. Walter J. Consciousness as a multidimensional phenomenon: implications for the assessment of disorders of consciousness. *Neurosci Conscious*. 2021;2021(2):niab047. doi:10.1093/nc/niab047.
88. Irwin LN. Behavioral indicators of heterogeneous subjective experience in animals across the phylogenetic spectrum: Implications for comparative animal phenomenology. *Heliyon*. 2024. doi:10.1016/j.heliyon.2024.e28421.
89. Earp BD, McLoughlin KL, Monrad JT, Clark MS, Crockett MJ. How social relationships shape moral wrongness judgments. *Nature Communications*. 2021;12(1):5776. doi:10.1038/s41467-021-26067-4.
90. Clark MS, Mills J. The difference between communal and exchange relationships: What it is and is not. *Personality and Social Psychology Bulletin*. 1993;19(6):684-91. doi:10.1177/0146167293196003.
91. Veit W. The Origins of Consciousness or the War of the Five Dimensions. *Biological Theory*. 2022;17(4):276-91. doi:10.1007/s13752-022-00408-y.
92. Changeux JP. The Ferrier Lecture 1998. The molecular biology of consciousness investigated with genetically modified mice. *Philos Trans R Soc Lond B Biol Sci*. 2006;361(1476):2239-59. doi:10.1098/rstb.2006.1832.
93. Schneider S. *Artificial you : AI and the future of your mind*. Princeton: Princeton University Press; 2019.
94. Laureys S, Schiff ND. Coma and consciousness: paradigms (re)framed by neuroimaging. *NeuroImage*. 2012;61(2):478-91. doi:10.1016/j.neuroimage.2011.12.041.
95. Pennartz C. Consciousness, Representation, Action: The Importance of Being Goal-Directed. *Trends in cognitive sciences*. 2018;22(2):137-53. doi:10.1016/j.tics.2017.10.006.
96. Verschure PF. Synthetic consciousness: the distributed adaptive control perspective. *Philos Trans R Soc Lond B Biol Sci*. 2016;371(1701). doi:10.1098/rstb.2015.0448.
97. Chatila R, Renaudo E, Andries M, Chavez-Garcia RO, Luce-Vayrac P, Gottstein R et al. Toward Self-Aware Robots. *Frontiers in robotics and AI*. 2018;5:88. doi:10.3389/frobt.2018.00088.
98. Mashour GA, Roelfsema P, Changeux J-P, Dehaene S. Conscious processing and the global neuronal workspace hypothesis. *Neuron*. 2020;105(5):776-98.
99. Farisco M, Changeux JP. About the compatibility between the perturbational complexity index and the global neuronal workspace theory of consciousness. *Neurosci Conscious*. 2023;2023(1):niad016. doi:10.1093/nc/niad016.
100. Brown EN, Lydic R, Schiff ND. General anesthesia, sleep, and coma. *N Engl J Med*. 2010;363(27):2638-50. doi:10.1056/NEJMra0808281.

101. Fingelkurts AA, Bagnato S, Boccagni C, Galardi G. DMN Operational Synchrony Relates to Self-Consciousness: Evidence from Patients in Vegetative and Minimally Conscious States. *Open Neuroimag J.* 2012;6:55-68. doi:10.2174/1874440001206010055.
102. Vanhaudenhuyse A, Demertzi A, Schabus M, Noirhomme Q, Bredart S, Boly M et al. Two distinct neuronal networks mediate the awareness of environment and of self. *J Cogn Neurosci.* 2011;23(3):570-8. doi:10.1162/jocn.2010.21488.
103. LeCun Y. *A Path Towards Autonomous Machine Intelligence.* 2022.
104. Edelman GM. *Bright air, brilliant fire : on the matter of the mind.* New York, NY: BasicBooks; 1992.
105. Changeux JP, Dehaene S. Neuronal models of cognitive functions. *Cognition.* 1989;33(1-2):63-109.
106. Dehaene S, Changeux JP. The Wisconsin Card Sorting Test: theoretical analysis and modeling in a neuronal network. *Cerebral cortex.* 1991;1(1):62-79. doi:10.1093/cercor/1.1.62.