Behavioral and Brain Sciences

cambridge.org/bbs

Conscious artificial intelligence and biological naturalism

Anil K. Seth[a,b]

[a]Sussex Centre for Consciousness Science, University of Sussex, Brighton, UK and [b]Program for Brain, Mind, and Consciousness, Canadian Institute for Advanced Research (CIFAR), Toronto, Ontario, Canada

Target Article

Cite this article: Seth AK. (2026) Conscious artificial intelligence and biological naturalism. *Behavioral and Brain Sciences* **49**, e315: 1–109. doi:10.1017/S0140525X25000032

Target Article Accepted: 21 March 2025
Target Article Manuscript Online: 21 April 2025
Commentaries Accepted: 6 October 2025

Keywords:
active inference; artificial intelligence; autopoiesis; biological naturalism; computational functionalism; consciousness; free energy principle; neuromorphic computation; predictive processing; substrate independence

What is Open Peer Commentary? What follows on these pages is known as a Treatment, in which a significant and controversial Target Article is published along with Commentaries (p. 21) and an Author's Response (p. 92). See bbsonline.org for more information.

Corresponding author:
Anil K. Seth;
Email: a.k.seth@sussex.ac.uk

Abstract

As artificial intelligence (AI) continues to advance, it is natural to ask whether AI systems can be not only intelligent but also conscious. I consider why people might think AI could develop consciousness, identifying some biases that lead us astray. I ask what it would take for conscious AI to be a realistic prospect, challenging the assumption that computation provides a sufficient basis for consciousness. I'll instead make the case that consciousness depends on our nature as living organisms – a form of biological naturalism. I lay out a range of scenarios for conscious AI, concluding that real artificial consciousness is unlikely along current trajectories but becomes more plausible as AI becomes more brain-like and/or life-like. I finish by exploring ethical considerations arising from AI that either is or convincingly appears to be conscious. If we sell our minds too cheaply to our machine creations, we not only overestimate them – we underestimate ourselves.

1.0. Introduction

"Technology can make us forget what we know about life" [(Turkle, 2021), p. Xii]

As artificial intelligence (AI) systems gain competence and capability, a question arises about whether such systems might become (or might already be) conscious as well as intelligent. A definitive answer to this question is not currently possible, given the lack of consensus about the minimally sufficient conditions for consciousness. The idea that AI may already be, or could become, conscious is nonetheless widespread (Blum & Blum, 2024; Butlin et al., 2023; Chalmers, 2023; Dehaene, Lau, & Kouider, 2017). How we think about the prospects for conscious AI matters. It matters from the perspective of the AI systems themselves. It matters from the perspective of how human beings and human society are affected by systems that are, or appear to be, conscious. And it matters for how we decide what kind of AI we want in our societies.

I will address the prospects for creating conscious AI and the pitfalls that accompany AI that is, or appears to be, conscious. I will start by asking why people might think that AI is on a trajectory towards consciousness, identifying some psychological biases that can lead us to overestimate its likelihood. I will then examine some assumptions that underpin the possibility of conscious AI, focusing on *computational functionalism* – the notion (broadly) that computation is sufficient for consciousness. I intend this discussion to give reasons to doubt that AI is, or could be, conscious. It is not intended to show that it is impossible to build a conscious artefact, but that such artefacts may have to be closer to biological systems than allowed for by conventional digital computation.

I will then move on to *biological naturalism*, understood here as the idea that consciousness is a property of only, but not necessarily all, living systems. I outline and motivate several variations of this view, recognising that "life" need not be carbon-based and without relying on any neo-vitalistic magic.

I will finish by discussing wider issues, highlighting concerns surrounding AI systems that are, or appear to be, conscious. Even if the prospects for real artificial consciousness remain remote, there are significant ethical and societal concerns relating to AI systems that give the powerful and perhaps irresistible appearance of being conscious.

1.1 *Intelligence and consciousness*

To begin, we need working definitions of both "consciousness" and "intelligence." Consensus definitions are lacking for both of these terms. But some ground rules are necessary.

I take intelligence to refer to what a system can *do*. One useful definition, drawn from a broad survey, is the "ability to achieve goals in a wide range of environments" (Legg & Hutter, 2007) (p.23). Intelligence in this sense, as in most or all other senses, has to do with the functional capabilities of a system. Reasoning, goal-directed planning, and linguistic competence are all aspects of intelligence, but non-human animals that lack these competencies may still exhibit intelligent behaviour. More general aspects of intelligence may include the ability to flexibly

© The Author(s), 2025. Published by Cambridge University Press

respond to challenges and to learn from experience. An ant that flexibly learns to find its way home through a desert is arguably displaying intelligent behaviour (Simon, 1988).

AI systems are artificial systems that exhibit functional properties associated with intelligence. Unless specified otherwise, I will assume a standard view in which AI systems (or "agents"[1]) consist of digital computation implemented in established architectures (CPUs, GPUs, TPUs, etc.) and potentially involve real or virtual interactions with (e.g., robotic) bodies and environments (Russell & Norvig, 2003). Artificial *General* Intelligence (AGI) is the term given to (still hypothetical) AI systems that have attained or exceeded human-level intelligence in (at least) all the ways humans are supposedly intelligent. These systems are general in the sense of generalising beyond specific task domains and training regimes (Goertzel & Pennachin, 2007).

Consciousness is not the same as intelligence. It is not (at least not primarily) about what a system does. A useful baseline definition of consciousness comes from Thomas Nagel: "For a conscious organism, there is something it is like to be that organism" [(Nagel, 1974), p.436]. That is, it "feels like" something to be a conscious system – there is a conscious experience happening – whereas it doesn't feel like anything to be an unconscious system – there is no conscious experience happening. Here, "feeling" need not involve emotional content: any kind of conscious experience will do. It (probably) feels like something to be a bat, and it (probably) doesn't feel like anything to be a stone. I take it that there is a fact of the matter about whether something is conscious, and that being conscious is not determined by social or linguistic consensus [see (Shanahan, 2024)].

I treat "consciousness" as synonymous with "awareness." I avoid the term "sentience," which is sometimes used synonymously with consciousness/awareness but which is more prone to confusion. Some people use "sentience" to denote mere sensitivity or responsiveness, without implying the presence of conscious experience (Kagan et al., 2022; Kagan et al., 2023). Others use "sentience" to mean a form of consciousness that is necessarily valenced, where experiences feel "good" or "bad" in some sense (Thompson, 2022).

Given these definitions, intelligence and consciousness are evidently conceptually distinct. In principle, a creature or system could display intelligent behaviour without there being anything it is like to be that creature or system, and vice versa. However, in practice, these properties do not seem to be independent. It might be that consciousness is useful or even necessary for some forms of intelligence, and that some forms of consciousness require corresponding forms of intelligence.[2]

2.0 Why might we think that AI could be or become conscious?

The idea of conscious AI has deep roots. In 16th-century Jewish folklore, a golem made of clay – called Josef, or Yoselle – became conscious and disturbingly violent when activated by magical incantation. More recently, science fiction has been a reliable source of imagined creations that explore the idea of artificial consciousness, from HAL in Stanley Kubrick's *2001* (released in 1968) to Ava in Alex Garland's *Ex Machina* (2014) and Klara in Kazuo Ishiguro's *Klara and the Sun* (2021), among many other examples.

In each of these stories, the artificial creations are intended by their creators to be humanlike, not necessarily in appearance (e.g., HAL), but in mind. This exposes a mixture of three psychological biases: anthropocentrism, human exceptionalism, and anthropomorphism. These biases can lead to intelligence and consciousness being conflated, which in turn drives an unwarranted optimism about the prospects for conscious AI.

Anthropocentrism is the tendency to place humans at the centre of things and to interpret the world in terms of human values and experiences. This is closely related to human exceptionalism – the tendency to see human beings as distinct from and superior to other forms of life. Putting these together, there is the temptation to take aspects of intelligence that seem distinctively and perhaps exclusively human – such as language and other "advanced" cognitive abilities – as bound up with other supposedly distinctively humanlike qualities – such as consciousness. We humans know we're conscious (strictly, we know it for ourselves and infer it about others), we think we're intelligent, and we feel we're special, so we assume that intelligence and consciousness go together. This thinking traces back at least to Descartes, for whom the immaterial, rational mind both endowed humans with the kind of consciousness that mattered and separated humans from other animals (Shugg, 1968).

Anthropomorphism is the complement to anthropocentrism: it is the tendency to project humanlike properties into things that don't (necessarily) have them. For example, we are more likely to attribute emotions to things with faces than to things without faces (Alais et al., 2021). This can be useful when it helps us explain and predict a system's behaviour, as in Dennett's "intentional stance" (Dennett, 1987), but it can be problematic when the unwarranted attribution of humanlike properties leads to poor explanations and/or unreliable predictions. When coupled with anthropocentrism and human exceptionalism, anthropomorphism can engender a tendency to attribute consciousness to things that seem to display those humanlike qualities that we think of as most distinctive, such as language and "higher" cognition.

These biases go some way towards accounting for the view that AI is on a royal road to consciousness: that as machines get smarter, they will eventually cross a threshold at which they will also become conscious. Once a system appears to gain a property we (anthropocentrically) feel is distinctively human, we are more likely (anthropomorphically) to attribute to it other supposedly distinctive properties, such as consciousness. The contribution of human exceptionalism is highlighted when the relevant threshold is assumed, implicitly or explicitly, to be the point at which AGI is reached. And the pedestal on which we put language, as a canonical expression of distinctively humanlike intelligence, explains why the current wave of large language models (LLMs) has been so powerful in seducing our intuitions (Colombatto & Fleming, 2024; Floridi & Nobre, 2024; Mitchell, 2024).

As an example of the seductive power of LLMs, consider how normal it has become to say they "hallucinate" when they spew falsehoods. Hallucinations in human beings are mainly about altered perceptual experience. The use of this term for LLMs therefore implicitly attributes experiential capacities to them. It would be more appropriate to say that they "confabulate." In humans, confabulation involves the unintentional – and usually linguistic – fabrication of content: making stuff up without realising that one is doing so. It is primarily about *doing*, rather than *experiencing*.

There is arguably a fourth ingredient: the technophilic temptation to see ourselves at the cusp of a major civilisational transition – a Kurzwelian Singularity beyond which AI bootstraps itself beyond our understanding and control (Kurzweil, 2005, 2024). This temptation is motivated in part by the psychological difficulty

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

of grasping the nature of exponential change (Wagenaar & Sagaria, 1975). From any point on an exponential curve, the future looks impossibly steep, while the past looks irrelevantly flat (and yet we still tend to underestimate exponential change – objects in the future are closer than they appear[3]). Another part comes from a posthuman desire to transcend the limits of biological mortality by uploading one's consciousness into the pristine circuits of some future supercomputer and therefore to live – or at least "be" – forever. Indeed, the idea that we already exist within a computer simulation has received philosophical attention (Bostrom, 2003).

Altogether, it is hardly surprising to witness the widespread emergence of the view that AI will inevitably, and perhaps imminently, lead to real artificial consciousness. But having recognised some key biases driving this view, we are freed from assuming that making machines smarter will inevitably make them conscious, even though we may become more susceptible to judging them as being conscious.

Instead of assuming that real artificial consciousness will come along "for free," one could try to develop it deliberately. One strategy for this is to implement, in AI systems, the functional principles associated with consciousness as specified by various theories of consciousness (Butlin et al., 2023; Seth & Bayne, 2022) or that arise from analysing different dimensions of consciousness (Evers et al., 2025). For this strategy to succeed, it must be the case that consciousness is the kind of thing that (digital computation-based) AI can have. This brings us to *computational functionalism*.

3.0 Computation, mind, and consciousness

3.1 *Computational functionalism*

Computational functionalism is the view that computations of some kind are sufficient to instantiate consciousness. It derives from the philosophical notion of functionalism, which says – roughly – that consciousness depends on what a system does, not on what it is made out of (Putnam, 1975).

More specifically, functionalism proposes that a system has a mind when it has a suitable functional organisation (Putnam, 1967). Functional organisation encompasses both the input-output mappings of the system as a whole as well as its internal organisation, including its causal structure. Mental states play roles within this functional organisation and are characterised by their interactions with sensory inputs, motor outputs, and other mental states. Applied to consciousness, functionalism proposes that conscious mental states are properties of the functional organisation of its underlying mechanism, which for human beings is the (embodied and embedded) brain.

Computational functionalism is the further claim that the kind of functional organisation that matters for mind in general, and for consciousness in particular, is computational in nature (Putnam, 1988; Shagrir, 2010). This raises the vexing question of what counts as "computational" (N. G. Anderson & G. Piccinini, 2024; Cantwell Smith, 2002; Searle, 1990).

The nature of computation is a cavernous topic, but it pays to scope the territory. The most relevant notion here is "Turing computation," which Alan Turing introduced via the concept of a "Turing machine" (Turing, 1936). A Turing machine applies a finite set of rules to manipulate a finite set of symbols to calculate the values of a mathematical function: in other words, an algorithm. Turing also proposed that there is a Turing machine for every algorithm – a widely accepted idea known as the Church-Turing thesis – and he defined a "universal" Turing machine capable of implementing any specific Turing machine, given sufficient storage (Copeland, 2024). The digital computers powering conventional AI are universal Turing machines.

At around the same time, Warren McCulloch and Walter Pitts showed that any combination of certain Boolean logic operations could be performed by circuits of highly simplified artificial neurons (McCulloch & Pitts, 1943). It was later shown that these artificial neural networks could be construed as Turing machines (Kleene, 1956)[4]. This motivated the view that biological nervous systems could be thought of as computing machines in the Turing sense (Piccinini, 2023) – a view encouraged by Turing with his famous question, "Can machines think?" (Turing, 1950).[5]

Turing computation is powerful, but not every function is Turing-computable. Turing himself identified a class of non-computable functions in his response to the "Entscheidungsproblem" posed by Hilbert (Turing, 1936). Other examples include some functions involving continuous variables, stochastic/random elements, and unbounded sensitivity to initial conditions (e.g., deterministic chaos). Digital computers based on Turing machines can simulate and approximate some of these functions – such as continuous processes and processes with stochastic elements (at the level of distributions, rather than any specific instance). Indeed, this happens all the time in computational modelling – but these approximations will generally not be exact.[6]

The limited remit of Turing computation means that systems – including brains – might implement functions that are non-Turing-computational. The idea that mental states (including consciousness) depend on non-computational functions is called *non-computational functionalism* (Piccinini, 2018, 2020). Non-computational neural functions could include processes relating to (continuous) electromagnetic fields, fine-grained timing relations (only order, not dynamics as such, matters for Turing computation), freely diffusing neurotransmitters, and so on. Non-computational biological functions also include those that necessarily involve a particular material property: examples include digestion, circulation of blood, and metabolism. Note that computational and non-computational functions could co-exist. For example, it could be that some aspects of the mind are computational, but not consciousness (Piccinini, 2023).

Many other notions of "computation" have been proposed. Some are narrower than Turing computation (e.g., computation requiring an artefact being used by a person in a particular way), but most are broader (N. G. Anderson & G. Piccinini, 2024; Chalmers, 1996b; Kirkpatrick, 2022). Broader forms of computation include analogue, neuromorphic, and mortal computation. I will return to these later. For now, a focus on Turing computation is justified since this kind of computation underlies conventional AI, whether based on artificial neural networks or otherwise. I will use the word "computation" to signify Turing computation, unless otherwise qualified. (Note that "information processing" is often used coextensively with "computation," but it is unhelpfully vague, partly because information theory can be applied descriptively to almost anything[7], and partly because it is often unclear what "processing" means.)

Functionalism in general and computational functionalism in particular are widely accepted both explicitly and implicitly. In an influential review of prospects for conscious AI, Patrick Butlin, Robert Long, and colleagues confine themselves to theories and models that assume computational functionalism and explicitly acknowledge that their conclusions depend on whether this assumption holds (Butlin et al., 2023). But since not all functions

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

are computable, and since not everything implements computations, it is worth interrogating this assumption.

3.2 *The appeal of computational functionalism*

Just as psychological and historical factors drive associations between intelligence and consciousness, there is a wider context which helps explain the appeal of computational functionalism. Recognising this context helps assess its plausibility.

Functionalism in general was motivated in part by the perceived inadequacies of other philosophical frameworks, including behaviourism, identity theories (which propose a strict reduction from mental states to physical brain states), and dualism (Shagrir, 2010). The focus on functional organisation provided an attractive middle ground between behaviourism and identity theory that accommodates *multiple realisability*: the idea that the same mental state can be realised in different ways (Polger & Shapiro, 2016; Putnam, 1967). Importantly, multiple realisability does not imply that any mental state can be implemented in any physical system. Some physical systems might not be up to the job.

Enthusiasm for *computational* functionalism in particular can be traced to a mathematical marriage of convenience between a specific view of computation – the Turing machine – and a drastic simplification of the messy multiscale biological complexity of the brain – the artificial neural network. This enthusiasm gathered pace as computer science and early forms of AI progressed rapidly in the second half of the 20th century (Boden, 2008; Dupuy, 2009). These advances helped embed the metaphor of the brain as a computer, normalise the language of information processing as describing what brains do, and galvanise the discipline of cognitive science. But metaphors are in the end just metaphors (Cobb, 2020). The fact that mental processes can often be usefully described in terms of computation is not sufficient to conclude that the brain actually computes or that consciousness is a form of computation.

For conscious AI to be possible for conventional digital systems, it needs to be the case that the relevant computations provide a sufficient basis for consciousness. This requires both that computational functionalism holds *and* that AI systems can implement the relevant computations. The fact that computational functionalism is frequently assumed does not mean it is true. As science writer Oshan Jarrow puts it, "If we keep making the assumption without returning to it, the question itself begins to disappear."[8]

3.3 *Substrate independence, flexibility, and neural replacement*

One way to put the relationship between conscious AI and computational functionalism is in terms of *substrate independence*. Substrate (or equivalently, medium) independence is closely related to multiple realisability. It is the claim that the same kind of mental states (including conscious states) can occur in systems with different physical substrates – i.e., systems made out of different kinds of stuff. Note that unrestricted substrate independence is unlikely to hold. As with multiple realisability, some physical substrates might not be up to the job. I will therefore use the term "*substrate flexibility*" for the idea that (conscious) mental states are not tied to carbon-based biological substrates – that they might be realisable in some other, but not necessarily all, types of material [see (Kirkpatrick, 2022) for a similar idea].[9]

Computational functionalism is intimately related to substrate flexibility. If computational functionalism holds, then some substrate flexibility follows for those substrates able to implement the relevant computations. In other words, computational functionalism is sufficient for substrate flexibility. But it is not necessary. One can imagine some substrate flexibility without endorsing computational theories of mind or consciousness. However, to the extent that substrate flexibility does hold, computational functionalism gains support because it provides an account of mental states that doesn't depend on the material properties of a substrate, beyond its ability to implement the relevant computations. Conscious AI requires substrate flexibility to hold at least for the substrate in question: silicon.

One of the best-known arguments for substrate flexibility in this context is the "neural replacement" thought experiment (Chalmers, 1995; Haugeland, 1980; Pylyshyn, 1980). The basic scenario is that a person's brain cells[10] are replaced, one by one, with silicon alternatives. Each silicon brain cell exactly replicates the input/output mapping of its biological counterpart in all situations. Eventually, the person's brain has only silicon parts, yet its functional organisation is entirely preserved, and so, from the outside, the person would behave exactly as before. If replacing one brain cell doesn't make a difference to the person's consciousness – and this seems plausible – then why should replacing one hundred, or all of them? Would consciousness simply fade away? And, if so, is it plausible that their behaviour remains unchanged while conscious experience completely disappears? The way out of these strange implications seems to be to accept silicon substrate flexibility.

This thought experiment can be criticised from several directions. Block (2019) points out that it begs the question. If consciousness is substrate dependent, then something would indeed happen to the person's conscious experience as the neural replacement progressed. [This doesn't mean the person needs to notice this happening: the change of experience does not always entail the experience of change (Simons & Levin, 1997)].

Other criticisms (also mentioned by Block) point to plausibility considerations. Building on early arguments from Bechtel & Mundale (1999), Godfrey-Smith argues that as living parts (e.g., neurons) are replaced by non-living (e.g., silicon) parts, there will inevitably be differences in how the system works on the inside and in its overall behaviour (Godfrey-Smith, 2016). These differences emerge because of influences on neural activity arising from metabolic constraints, electromagnetic fields, fine-grained timing relations, and potentially other factors (note the similarity with examples of non-computational functionalism given earlier). If the system behaves differently, internally and externally, then there is no pressure to suppose that consciousness needs to be maintained.

Rosa Cao extends this line of argument, providing many descriptions of how mental functions are thickly intertwined with each other and with their underlying physiological and metabolic bases (Cao, 2022). She argues that because mental functions arose through evolution, the realisers of these functions have accumulated many dependencies and contingencies, some of which may seem arbitrary and which collectively lead to what Wimsatt (1986) called a "generative entrenchment" in the internal organisation of the brain. This generative entrenchment imposes many constraints on possible alternative implementations, which likely exclude alternative substrates such as silicon.

Many examples can be given. One from Cao (2022) concerns the role of nitric oxide, a neurotransmitter involved in many neuronal functions. Nitric oxide is distinctive because it diffuses freely across cell membranes, meaning that any silicon replacement would need to be able to detect this substance (or a simulated version of it) at an infeasible number of (perhaps simulated) locations.[11] Another comes from a recent study showing that some neuronal spikes serve

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

a homeostatic role by protecting neurons from the build-up of damaging reactive oxygen species (Chintaluri & Vogels, 2023). This means that the spiking activity of a neuron, generally considered key to its functional role, cannot be decoupled from its metabolic foundations. Even chemotactic behaviour in bacteria may be inseparable from metabolic processes (Alexandre, 2010; Egbert, Barandiaran, & Di Paolo, 2010), suggesting a link between metabolism and mind that I will return to later.

Derek Shiller raises concerns at a more abstract level (Shiller, 2024). He argues that a liberal interpretation of functionalism could attribute mental states and consciousness to all kinds of things by carving them up in particular ways. A bucket of water may transiently reflect the functional organisation of a brain if the water is "carved up" into parts in the right way.[12] To prevent this, it is necessary to place constraints on what can count as a system's parts. Shiller identifies several constraints, such as the need for parts to play their functional roles in virtue of intrinsic causal powers. He argues that these constraints are not satisfied by current neural network models and that they may be difficult to satisfy for non-biological substrates in general.

These plausibility arguments do not disprove silicon substrate flexibility, but they do challenge assumptions that it must be the case.[13] Brains seem to be the kind of thing for which it is hard and perhaps impossible to separate *what they do* from *what they are*.[14]

3.4 *Broader forms of computation*

Substrate flexibility is also challenged by broader forms of computation which relax some constraints of the Turing world. As mentioned, there is considerable interest in what counts as computation (N. G. Anderson & G. Piccinini, 2024; Cantwell Smith, 2002; Chalmers, 1994, 1996b; Kirkpatrick, 2022; Searle, 1990). Although I cannot cover this ground in detail here, it will serve to highlight a few examples.

The recently introduced concept of *mortal computation* is particularly interesting (Hinton, 2022; Ororbia & Friston, 2023). Standard Turing computation is "immortal." Its existence and utility outlast the existence of any specific instance of hardware. This reflects the core computer science principle that software should be separable from hardware both in principle and in practice so that the same algorithm executed on different hardware gives the same result. But immortal computation is expensive. It requires continual error correction to ensure that 1s remain 1s (and 0s remain 0s). As algorithms and models grow in complexity, the computational and energetic costs of error correction, and therefore of computational immortality, grow quickly.

One implication of this argument is that biological brains, which are highly energy efficient, cannot implement immortal computations. If they are implementing computations at all, then these computations are likely to be *mortal*, which means they cannot be separated from the "hardware" (or "wetware") which implements them. This, in turn, places constraints on the multiple realisability and substrate flexibility of these (mortal) computations. In particular, the substrate flexibility required for conscious AI is unlikely to hold because (conventional) AI is based on an implementation paradigm which assumes computational immortality. I find this a provocative argument against the plausibility of conscious AI because it is based on limitations arising from within a computational view of mind.[15]

A related view, advanced by Gualtiero Piccinini, is that biological functions involve *neural computation*, which in his definition involves both continuous and discrete signalling and is therefore different from and more substrate-dependent than the "neural computations" performed by standard artificial neural networks [(Piccinini, 2020), ch.13]. Also related is the *biological computation* proposed by Kay Kirkpatrick, which sets out specific conditions according to which living systems may be said to compute (Kirkpatrick, 2022).

Coming from the other direction, there have long been technological alternatives to the standard computer science paradigm. *Analogue computation* – which dates back at least to ancient Greece[16] – involves continuous variables and is typically dependent on specific properties of the underlying mechanism (e.g., amplifiers and capacitors) (Ulmann, 2022). *Neuromorphic computing* involves hardware that has more in common with biological neural systems compared to modern CPUs and GPUs (Mead, 2020; Schuman et al., 2022). Some neuromorphic chips use spikes to communicate between hardware elements, offering greater energy efficiency, among other benefits (Hochstetter et al., 2021; Merolla et al., 2014). Other neuromorphic approaches utilise analogue computation, again tying function more directly to substrate (Douglas, Mahowald, & Mead, 1995). Some emerging approaches even use biological neural networks interfaced with standard computational input/output devices (Morales Pantoja et al., 2023).[17] It is plausible that the more analogue or neuromorphic hardware becomes, the less substrate flexibility there is for the functions or computations it performs.[18]

These broader senses of computation challenge substrate flexibility from within a broadly computational view, and they do so from both directions – biological and artificial. One might argue that standard Turing computation could still simulate the mortal (or other) computations that potentially underlie consciousness in biological brains. But the computations themselves would be different, and the simulations could not be guaranteed to be exact. One could also redefine computational functionalism for any of the senses of computation mentioned, but the underlying motivation would be weakened, given the diminished substrate flexibility. And, in practice, unlike the comforting assurances provided by the Turing world, there would be no guarantee that the (mortal, neural, etc.) computations relevant to consciousness could be implemented in (neuromorphic, analogue, etc.) hardware.

Changing the boundaries of what counts as "computational" also changes what counts as "non-computational." Even with broader senses of computation to choose from, non-computational functions may still be critical to cognition and consciousness, and it remains an open question as to whether brains "compute" at all.

3.5 *Emergence and the separation of scales*

Another useful perspective on substrate flexibility comes from hierarchical emergence in physics. Physicists speak of a "separation of scales" which allows laws to be formulated at higher – emergent – levels of description that do not require knowing the finer-grained details. Engineers can build bridges without knowing about quantum mechanics. Computer programmers can write software without worrying about implementation. The substrate flexibility (and immortality) of software rests on this strong separation of scales.

Recently, empirical measures have been developed that formalise notions of "informational closure" and "causal closure" which underwrite the separation of scales in complex systems [e.g., (Barnett & Seth, 2023; Rosas et al., 2024)]. A macroscopic (higher-level) variable – or "coarse graining" – is causally or informationally closed when interventions on it (causal closure), or

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

observations of it (informational closure), are sufficient to determine or predict the higher-level outcome(s), insofar as these outcomes are determinable or predictable. To put it another way: knowledge of (or interventions at) lower, microscopic levels does not add anything to macro-level predictability or help determine macro-level outcomes beyond what can be done at the macro-level. This is guaranteed by design for (properly functioning, conventional) computers. But it would be a mistake to assume that these forms of closure hold for complex systems generically. In particular, thanks to the generative entrenchment inherent in evolutionary design, they may be partially or wholly absent in biological systems.

The ability to measure causal and informational closure emphasises that substrate flexibility cannot be taken for granted. In practice, such measures may help empirically determine whether, or to what extent, neural dynamics can be abstracted away from finer-grained levels of description.

3.6 *Beyond computational functionalism*

The arguments so far do not disprove computational functionalism. But they do render it less plausible and less appealing by noting that neurobiological functions likely have limited substrate flexibility and by recognising the existence of broader forms of computation which have similar constraints. Many other arguments can be raised, which I can only gesture at here. These include the distinction between syntax and semantics (Cole, 2023; Searle, 1980), potentially implying an observer dependency to consciousness[19]; philosophical considerations about intentionality (Shagrir, 2010); and Roger Penrose's argument from Gödel's theorem (Penrose, 1989).

Given all this, we should not simply assume that computational functionalism and widespread substrate flexibility must be the case. Since conscious AI depends on computational functionalism holding and on substrate flexibility extending at least to silicon, there are good reasons to question whether conscious AI is possible, again for AI understood as digital computations on a silicon substrate.

Considering what kind of thing a brain is, it is perhaps surprising that computational functionalism became so pervasively adopted. When we look inside a brain, we do not find anything like a sharp distinction between "mindware" and "wetware" of the kind we find between hardware and software in a computer. Brain activity patterns evolve over multiple scales of space and time, continuously influenced by, and influencing, the chemical diffusion of neurotransmitters, the physical structure of the neural networks themselves (synaptic plasticity and synaptogenesis), and various properties of cellular infrastructure, all deeply interwoven – generatively entrenched – with a molecular storm of metabolic activity. A single neuron is a complex biological system in and of itself, busy maintaining its own integrity and regenerating the conditions for its own material existence. The metaphor of the brain as a computer is an increasingly inconvenient fiction, and it seems myopic to consider all this multiscale biological activity as simply an enabling condition for the right kind of computation.

There are other possibilities. Consciousness might depend on alternative forms of computation. But, as mentioned, because these alternatives lack the substrate flexibility of Turing computation, they might not be implementable in alternative non-biological substrates, and there is less motivation for adopting a computational view in the first place.

More broadly, consciousness might depend on patterns of functional organisation, but not on computations (Piccinini, 2018, 2020; Prinz, 2012). This *non-computational functionalism* is well aligned with attempts to understand neural systems using perspectives that do not assume a computational stance. Much of this work traces back to early cybernetic ideas about neurobiology and AI, which emphasised embodiment, feedback, and control, rather than disembodied computation (Dupuy, 2009).

A good example is provided by the influential "dynamical systems" approach to cognitive science (Van Gelder, 1995). This approach utilises various mathematical tools and methods – usually based on differential equations – to model how properties of embodied and embedded neural systems change over time. These changes are generally not associated with any kind of computation, whether digital or analogue (Van Gelder, 1995).[20] Dynamical systems theory approaches historically did not focus on consciousness and in this regard are complementary to non-computational functionalism. Something similar can be said of the various perspectives under the banner of "4E" cognitive science, which emphasises embodied, embedded, enactive, and extended aspects of brain-body-world interactions. Consciousness appears more often in work of this kind (Cosmelli et al., 2007; Kiverstein, 2020; Stewart, Gapenne, & Di Paolo, 2010; Thompson, 2022). These views differ in important details, but they share a disdain for the idea that the brain is a computer and that the mind is essentially computational.

Various neuroscience-based theories of consciousness emphasise (or can be framed in terms of) dynamical rather than computational elements. These include classic works linking conscious perception to attractor dynamics in neural systems (Freeman, 1999; Llinas et al., 1998) and to neural synchronisation, for example, in the gamma frequency band (Crick & Koch, 1990). The "dynamic core" hypothesis (Tononi & Edelman, 1998), which highlighted integration and differentiation as key neurodynamical properties underlying consciousness – and which later resurfaced as the integrated information theory of consciousness (Albantakis et al., 2023; Tononi et al., 2016) – also falls within this camp, as does recurrent processing theory (Lamme, 2010, 2018). Predictive processing and the free energy principle (Clark, 2019; Friston, 2010) also qualify, depending on how these theories are interpreted. I'll return to these ideas shortly – they are central to the case I will make for the dependence of consciousness on biology. More generally, the approach of neurophenomenology, inspired by Francisco Varela among others, often appeals to non-computational dynamical constructs as bridging principles between neural processes and phenomenological properties (Thompson & Varela, 2001; Varela, 1996).

Dynamical neuroscientific and (neuro)phenomenological approaches to consciousness are promising. However, they are in general non-committal about whether implementing the dynamical principles proposed by a given theory, in some alternative material, would be sufficient to instantiate consciousness. (I hesitate to use the term "conscious AI" here because we are no longer talking about implementing computations nor about intelligence.) And just like broader senses of computation, implementing non-computational dynamical processes in alternative substrates raises the challenge of substrate flexibility

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

because the in-principle substrate independence provided by standard computationalism can no longer be appealed to.

3.7 *Simulation, implementation, and realisation*

At this point, it's worth clarifying the relations between simulation, implementation, and realisation (or instantiation), which until now have been left rather implicit. Applied to some target mechanism or system, such as a neurobiological process, I use these terms as follows. One can *simulate* the mechanism by abstracting it as a (Turing) computation and then *implementing* this computation on a digital computer. Any mechanism can be simulated, though these simulations may only be approximate (and may be wildly inaccurate). Alternatively, one could *implement* finer-grained or other non-computational aspects of the target mechanism in other ways (e.g., through neuromorphic engineering or synthetic biology). The limit of this approach is a physically identical replication. Finally, one can ask whether simulation of a mechanism, or other (e.g., non-computational) implementations of that mechanism, will *realise* (instantiate) some other property, such as consciousness.

The distinction between simulation and implementation need not be sharp. One could "simulate" a system using broader forms of computation, such as analogue computation, or even by using physical causal models. By allowing for the inclusion of finer-grained, causal, and potentially non-computational properties, these "simulations" would move closer to "implementations."

In general, simulating something (e.g., a rainstorm, a digestive process) is not equivalent to realising (instantiating) that thing. Simulations generally lack the causal powers and intrinsic properties of the things being simulated. For a simulation of X to be guaranteed to realise X, one would have to be confident that computation is sufficient for X. This is obviously true when X itself is a (Turing) computational process. For some other things, like rainstorms and digestion, it is obviously false.[21] Rainstorms and digestion are defined, at least in part, in terms of intrinsic material properties (e.g., water, food) and causal powers (e.g., wind, metabolic reactions). Nothing gets wet in a weather forecasting computer. Would simulating the neurobiological mechanisms underlying consciousness – whether exactly or approximately – be sufficient to realise consciousness? This would be guaranteed only if computation is sufficient for consciousness – which, notably, requires assuming computational functionalism.[22]

By contrast, *implementing* the neurobiological mechanisms underlying consciousness, at the right level of granularity, would by definition realise consciousness. This, of course, is not easy to do. Computational models can readily simulate all manner of systems and mechanisms, whether they are computational or not, but the ability to implement a mechanism at the right granularity depends strongly on what that mechanism is and what that granularity is.

3.8 *What about intelligence?*

My arguments so far have focused on implications for artificial consciousness. But if what brains *do* cannot be separated from what brains *are*, there are implications for artificial intelligence too. If the brain's functional organisation is intimately tied to its substrate properties, then there may be (potentially non-computational) patterns of functional organisation, restricted to biological systems, that are necessary for specific forms of intelligence. These forms of intelligence would be distinctively biological.

If these patterns of functional organisation can be simulated with sufficient fidelity, then computers could still realise (in approximation) the corresponding forms of intelligence. Because intelligence is broadly about *doing*, sufficiently accurate simulations may count as realisations, even if underlying mechanisms and intrinsic properties are different. For example, LLMs are able to mimic aspects of language, even if under the hood there is only an impoverished caricature of the processes underlying real human language (Mitchell & Krakauer, 2023). And computers do play (a version of) chess, even if they lack a humanlike understanding of the game.[23]

More provocatively, some intelligence-relevant patterns of functional organisation might always escape functionally adequate simulation or implementation in alternative substrates such as silicon. Michael Levin makes a case for this possibility, arguing that the active nature of biological material enables access to distinctive regions in a platonic space of cognitive competencies (Levin, 2025; McMillen & Levin, 2024). It could also be that the embodied and embedded nature of biological systems provides constraints on and opportunities for intelligence-relevant functionality that would be difficult or infeasible to simulate (see Section 5.7).

It is worth re-emphasising the difference between intelligence and consciousness in this context. Intelligence – at least in some forms – can be understood in terms of mapping inputs to outputs and so can be instantiated through computation, even if distinctively biological forms of intelligence remain out of reach or only loosely available through simulation. But for consciousness to be instantiated computationally *in any form*, one needs to assume computational functionalism.

3.9 *Summary*

Conscious AI requires both computational functionalism to hold and sufficient substrate flexibility such that the computations sufficient for consciousness can be implemented in AI hardware (silicon). But the functions or computations implemented by (conscious) biological systems may not be separable from their material basis. This means that the substrate flexibility required for conscious AI may not hold. There are also alternatives to assuming that brains compute or that consciousness is a matter of (Turing) computation, undermining assumptions of computational functionalism.

This brings us to *biological naturalism*: the view that biological properties are necessary for consciousness. In the next section, I will make a case for various forms of biological naturalism, framed in terms of explanatory bridges between the properties of mind and consciousness and properties of their biological underpinnings. This case rests on a particular theoretical perspective on perception, cognition, and consciousness. I will make it from two directions, meeting in the middle: the story from predictive processing, and the story from the free energy principle.

4.0 Towards biological naturalism

"The organism has to keep going, because to be going is its very existence" (Jonas, 2001).

Biological naturalism belongs to a family of ideas that stress intimate connections between properties of biological systems and properties of mind, including consciousness. In its original formulation, introduced by John Searle, it proposes that conscious states are higher-level irreducible properties of lower-level neurobiological states (Searle, 2017). I use biological naturalism in a

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

slightly different but now widespread way, as the claim that consciousness is a property of only (but not necessarily all) living systems. Biological naturalism should be distinguished from "biopsychism," which is the claim that *all* living systems are conscious (Haeckel, 1892; Thompson, 2022).

4.1 *The story from predictive processing*

Predictive processing, as I understand it, labels a range of theories about perception, cognition, and action (Clark, 2013; Hohwy, 2013; Hohwy & Seth, 2020; Rao & Ballard, 1999). These have in common the idea that perception is a form of probabilistic inference, approximated in neural systems via a process of prediction error minimisation.

The standard predictive processing story is that perceptual content is not "read out" from incoming sensory signals but is instead given by the brain's "best guess" about the causes of these sensory signals. In Bayesian terms, perceptual contents correspond to approximately optimal posterior beliefs, given by a weighted combination of prior beliefs and sensory signals (likelihoods).[24] Since exact Bayesian inference is generally analytically intractable, the brain implements an approximation: prediction error minimisation. The idea is that the brain implements a hierarchy of generative models (in Bayesian terms, the joint distribution of the likelihood and prior) which generate predictions about expected sensory signals, which cascade down through perceptual hierarchies. These "top-down" (inside-out) predictions are compared with bottom-up (outside-in) sensory signals to form prediction errors, which are passed back up the hierarchy. By updating predictions to minimise prediction errors, this process settles on an approximately optimal posterior – a "best guess." Prediction errors can also be minimised by performing actions to bring about the expected sensory data – a process called active inference (Friston et al., 2017).

The core claim from the perspective of consciousness is that conscious contents are given by the brain's "best guesses" – by the joint content of the top-down perceptual predictions (Hohwy & Seth, 2020; Whyte et al., 2024). This intuition is captured by the notion of conscious perception as a "controlled hallucination" – emphasising that conscious contents are actively generated, rather than passively registered (Seth, 2021).[25]

This standard story can be extended to account for conscious perceptions of the body, as well as of the world. Body-related perceptions form an essential part of the experience of being a "self." Experiences of emotion and mood, which seem central to selfhood, may depend on interoceptive inferences – probabilistic best guesses about the causes of sensory inputs signalling the physiological condition of the body (Barrett & Simmons, 2015; Seth, 2013). Importantly, the functional role of interoceptive inference is likely to be more concerned with physiological regulation than with discovering what's there. This distinction can be framed in terms of instrumental versus epistemic inference. Epistemic inference is broadly about inferring the most likely causes of sensory signals, as in standard predictive processing. Instrumental inference prioritises control over discovery. In instrumental inference, prior beliefs can serve as targets or set points. Minimising (interoceptive) prediction error through active inference can implement homeostatic or allostatic regulation relating to these set points (Seth & Friston, 2016).[26]

This control-oriented perspective on interoceptive inference inherits from long-established cybernetic ideas emphasising the importance of prediction for control (Conant & Ashby, 1970; Seth, 2015). This suggests that the evolutionary driver for predictive processing as a general mechanism underlying conscious perception may have been a fundamental biological imperative for allostasis – for staying alive. It also sheds light on some relevant phenomenological distinctions. Visual experiences share, to a first approximation, a range of phenomenological characteristics, including spatiality, object-relatedness, and so on. These characteristics are well aligned with (epistemic) predictions which have the broad function of figuring out "what's there." Emotional experiences have a different character: they are shaped primarily by valence – things feeling (variations of) good or bad. These characteristics are well aligned with control-oriented predictions, which have the broad function of allostatic regulation.

These observations together suggest a kind of *epistemic* biological naturalism in which human conscious experiences can only be understood in light of our nature as living, self-sustaining organisms.[27] This claim applies to conscious perception, rather than to sensitivity to the world (and body) in general, thanks to the explanatory links to phenomenological properties of conscious experience (Seth, 2016). Perhaps there is even a minimal "ground state" of conscious experience, which – instead of being entirely free of content – is characterised by an inchoate, shapeless, formless feeling of simply "being alive."[28] This putative ground state could then scaffold all other forms of conscious content, whether self-related or world-related. Informally: we experience the world, and the self, *with*, *through*, and *because of* our living bodies [(Seth, 2021); see also (Seth, 2019; Seth & Tsakiris, 2018)].

A further step can be taken by noticing that the imperative to stay alive doesn't bottom out at any particular "implementation level" in our biology. Even single cells – including neurons – are continuously engaged in maintaining their own physiological integrity. This homeostatic process can be directly linked to higher-level functional properties. I gave one example earlier: neuronal firing patterns may partly depend on how apparently spontaneous spiking can prevent the build-up of damaging chemicals (Chintaluri & Vogels, 2023). If there is no clear boundary separating the substrate from the processes and dynamics that arise from it, then it is harder to motivate the intuition – central to computational functionalism – that the substrate does *not* matter.

Another perspective is that living systems have "skin in the game": they are constantly engaged in processes of self-maintenance within and across multiple levels of organisation (Aru, Larkum, & Shine, 2023). On this view, the (higher-level) mechanisms of predictive processing are inextricably embedded in their (lower-level) biological context. This embedding is necessary for realising the higher-level properties, and – following the arguments above – helps explain their connection to consciousness.

This brings us to *autopoiesis*, a concept deriving from the Greek for "self-production" or "self-creation." Autopoietic systems are special because they continually regenerate their own material components through a network of processes, and because this network of processes actively maintains a boundary between the system and its surroundings (Maturana & Varela, 1980). Biological cells are prototypical examples of autopoietic systems: they do not merely implement functions; they also continuously regenerate their own material basis (self-production) and integrity

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

(self-identification). Human-designed systems, like computers and factories, are generally not autopoietic. A factory takes in raw materials and produces some output – perhaps a car. The factory does not produce itself. A computer takes in some input and generates some output. It also does not produce itself. Systems like these are *allopoietic* – "other producing." For autopoietic systems, there really is no way to separate what they are from what they do.[29]

The idea surfacing here suggests a stronger continuity between life and consciousness, in which the predictive processes that underpin all conscious experiences are inextricably tied to the autopoietic and excitable nature of living matter.[30] The motivation for this idea rests on recognising that cellular autopoiesis can itself be understood as depending on – or partly consisting in – a kind of prediction error minimisation. This brings us to the *free energy principle*.

4.2 *The story from the free energy principle*

Having followed the predictive processing story first, the return journey will be faster. This journey starts by identifying basic constraints on the nature of systems that actively resist disorder so as to continue to exist – and from there recovers the larger predictive processing view about consciousness.[31]

We begin with the autopoietic nature of living systems, whether cells or entire organisms. Unlike other kinds of things, like rocks or computers, these systems actively (re-)generate their own material basis and maintain their boundaries over time. From the perspective of the free energy principle, this means that they actively resist the dispersion of their internal states – a dispersion otherwise mandated by the second law of thermodynamics. This in turn means they must exist in a state of low entropy, maintaining themselves out of thermodynamic equilibrium with their environment, in order to fend off the inevitable descent into disorder. To stay alive means to continually minimise entropy because there are many more ways of being mush than there are of being alive.

It is sometimes said that staying alive involves resisting the second law. It is better to say that living systems *take advantage of* the second law, since it is the transformation of low-entropy fuel/food into high-entropy products through metabolism that enables living systems to remain in non-equilibrium (quasi) steady states.[32]

The challenge here is that the system itself cannot directly measure entropy, and so it cannot be (directly) minimised. This is where free energy comes in. Free energy – technically the *variational free energy* – is a measurable upper bound on entropy.[33] This means that minimising (variational) free energy will also reduce entropy (Buckley et al., 2017; Parr, Pezzulo, & Friston, 2022). "Measurable" here means that it is a function of sensory signals and so is in principle accessible to the system. In practice, given some straightforward mathematical assumptions, free energy can be interpreted as a sensory prediction error.[34] This means that actively minimising sensory prediction errors entails minimising the theoretically more profound quantity of free energy.

Flipping this around, minimising free energy will also – under the same assumptions – lead to the approximate (Bayesian) posterior beliefs encoded by a system's generative models approaching the optimal posterior beliefs and, at the same time, to maximisation of the sensory evidence for these models.[35] This perspective underlines the continuity between the free energy principle and predictive processing.

The overall intuition here is that, by minimising sensory prediction error through active inference, living systems will naturally tend to be in states they expect – or predict – themselves to be in, and so will continue to exist (or continue to be the kinds of things they are). Jakob Hohwy puts this intuition neatly when he says that organisms maximise the evidence for their own existence – they "self-evidence" (Hohwy, 2014).

These aspects of the free energy principle suggest that the entire edifice of active inference reaches right down into the autopoietic nature of living systems. This continuity is deepened by noting an equivalence between (minimisation of) thermodynamic free energy in metabolism and (minimisation of) variational free energy in predictive processing. The minimisation in the metabolic case underwrites autopoietic integrity and the maintenance of a dynamic self/other boundary – which in the free energy principle literature is called a Markov blanket [(Friston, 2013), but see (Di Paolo, Thompson, & Beer, 2022; Nave, 2025)]. The minimisation in predictive processing underwrites the story of conscious experience in terms of predictive processing and controlled hallucination.

Two well-established results in physics – building on (Jaynes, 1957) – deepen this equivalence. The Jarzynski equality establishes that the average thermodynamic work done in moving between states is lower-bounded by the free energy difference between these states (Jarzynski, 1997). This applies specifically to non-equilibrium processes, such as living systems, and endows processes like metabolism with a free energy formulation. Jarzynski's equality can be considered as a generalisation of the Landauer principle, which places a lower bound on the (thermodynamic) energy required to erase one bit of information (Landauer, 1961). This bound implies a minimal thermodynamic cost to inference in predictive processing. More recently, Sengupta and colleagues showed how minimising variational free energy in the context of active inference leads to both metabolic and statistical (inferential) efficiency [(Sengupta, Stemmler, & Friston, 2013), see also (Fields, Goldstein, & Sandved-Smith, 2024)]. The idea is appealing: when predictions are accurate, thermodynamically costly state changes are not necessary.

Putting all this together suggests that the (minimisation of) free energy driving metabolism, autopoiesis, and perception is not just metaphorically equivalent but is, in some physical sense, the same thing. That said, there is still debate over exactly what this equivalence means (Collell & Fauquet, 2015), and the free energy principle itself, at least in its larger claims and its relation to autopoiesis, also remains much debated (Aguilera et al., 2022; Biehl, Pollock, & Kanai, 2021; Bruineberg et al., 2021; Di Paolo et al., 2022; Nave, 2025).

Someone might worry that predictive processing, with its Bayesian dance of prediction and prediction error, seems like a paradigmatically computational process. Certainly, predictive processing (and Bayesian inference) can be utilised as an algorithm and often is. The theory can also be interpreted as claiming only that (conscious) perceptual contents can be usefully *modelled* as a process of probabilistic inference (Andrews, 2021). But the continuity with the free energy principle marks an important distinction. Predictive processing in biological systems is a dynamical and (plausibly) substrate-dependent process, grounded in minimisation of a continuous quantity (free energy). This perspective goes some way towards licensing the stronger claim that perceptual contents are *constituted by* (not merely modelled as) aspects of perceptual inference. In this sense, predictive processing becomes a dynamical, non-computational theory, as anticipated in Section 3.6. This process can be

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

abstracted computationally, but these algorithms are pale shadows of the embodied and embedded processes unfolding in biological brains and bodies.

4.3 *Membranes, mitochondria, and xenobots*

Other emerging lines of work provide intriguing hints about the substrate dependency of biological functions in general and of consciousness in particular. I will mention two. The first comes from Nick Lane, who argues that a physical basis for "feeling" can be found even in simple living systems such as bacteria. Lane proposes that electrical potentials across the bacterial membrane, which are dynamically maintained through metabolism, provide an integrated, valenced, and action-oriented read-out of the physiological condition of the organism in relation to its environment (Lane, 2022).[36] There are connections here to the free energy principle, since the bacterial membrane can be thought of as a Markov blanket, and to autopoiesis, because metabolism both generates the membrane potential and is powered by it.

In more complex creatures, Lane proposes that mitochondria play a key role in the generation of conscious states. Evolutionarily, mitochondria derive from bacteria, with the mitochondrial plasma membrane playing a similar role to the bacterial membrane. He argues that the surprisingly strong electromagnetic fields induced by (and essential for) metabolism may influence synaptic activity and neuronal firing, and perhaps even the macroscopic electromagnetic fields that can be detected by EEG. He also proposes that anaesthetics work by disrupting specific processes within mitochondria that are responsible for generating these fields. In this view, the relevant aspects of mitochondrial (and bacterial) function cannot be abstracted away from the biochemistry of cellular respiration unfolding within these organelles.

The second perspective returns us to Michael Levin's innovative views about the "agential" nature of biological material (Levin, 2023, 2025). His group has experimentally investigated these ideas in the form of strange creations like "xenobots" and "anthrobots" (Ebrahimkhani & Levin, 2021; Gumuskaya et al., 2024). These are assemblages of amphibian (xenobot) or human (anthrobot) cells which self-organise into quasi-agential systems with behavioural competences much different from what could be expected from the usual context for these cells or their evolutionary history (e.g., motile anthrobots made from normally sessile human tracheal cells). While there is not yet any direct connection to consciousness here, Levin's creations reveal deep reservoirs of functional and dynamical potential inherent in biological material, contrasting sharply with conventional views prioritising genetic information storage, encoding, and decoding – and with the impoverished nature of the substrate in computational functionalism. Instead of computational rigidity, there is an excitable fluidity to the matter of life.

These perspectives complement the story from predictive processing and the free energy principle. They illustrate the rich resources biological theories can draw on when connecting properties of cognition and consciousness to properties of life.

4.4 *The real problem and the explanatory gap*

One reason we might find biological naturalism implausible is because of a residual sense of mystery about the nature of consciousness, coupled with the suspicion that associating it directly with properties of life falls foul of the kind of identity-theory materialism that motivated functionalist views in the first place.

I believe this worry can be disarmed, at least partially, by considering what a satisfying scientific explanation of consciousness might look like. This calls on what I have dubbed the "real problem of consciousness" [(Seth, 2016, 2021), see also (Searle, 2007)]. The real problem approach argues that as we become better able to *explain*, *predict*, and *control* properties of consciousness in terms of underlying mechanisms, the sense of mystery about how consciousness relates to matter will dissolve and may eventually disappear. The admittedly imperfect historical analogy here is how the conceptual mysteries of life – exemplified by vitalist intuitions – dissolved as biologists became better able to explain, predict, and control properties of living systems [(Churchland, 1996; Seth, 2021), see (Chalmers, 2003) for an objection].

If associations between properties of consciousness and properties of life have explanatory purchase – and are not brute reductions – then we become freed up to take biological naturalism more seriously. If features of being alive explain (and predict and control) features of being conscious, then a necessary relation between the two becomes more plausible.

4.5 *Summary*

How might life matter for consciousness? There is an epistemic view in which consciousness can only be understood in light of our nature as living systems. This view is the easiest to defend, though it relies on accepting predictive processing as a useful explanatory framework. There is also an ontological view in which consciousness is in some stronger way dependent on its biological, living substrate. Both views offer positive arguments for some form of biological naturalism, in contrast with merely asserting it in the face of problems with alternatives such as computational functionalism or regressing to neo-vitalistic magic.

These positive arguments apply to consciousness specifically, rather than to neurocognitive function in general. This is because the connections between properties of the biological substrate and (phenomenological) properties of conscious experience have rich potential explanatory value. In the limit, one can think of each and every conscious experience as (phenomenologically) integrating multimodal bodily and environmental signals in the service of coordinating allostatic regulation and survival-relevant action [see (Seth, 2021) for more, and (Kiverstein, 2020) for a related view]. To revive and repurpose a phrase from Descartes, we are conscious creatures because we are living, flesh-and-blood "beast machines" (Seth & Tsakiris, 2018).

Nothing presented here endorses biopsychism (the view that all life is conscious). This raises the challenging question of what distinguishes conscious from non-conscious living systems. I will not address this here, besides noting that not all forms of life might need to engage in the kind of multimodal survival-relevant integration mentioned above [though see (Lane, 2022)]. Neither do I endorse carbon chauvinism (the claim that only carbon-based life can be conscious). Non-carbon-based "living" systems might well be able to participate in the same kind of substrate-dependent multiscale activity that I have argued to be important (i.e., the relevant processes might have sufficient substrate flexibility to be implemented beyond carbon). This raises other challenging questions about what counts as a "living system." These, too, are questions for another time (Ball, 2023; Lane, 2015; Schrödinger, 1944).

5.0 Scenarios for real artificial consciousness

My case against conscious AI has two parts. There is a negative part, questioning the assumptions that lead people to think that

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

Table 1. *Scenarios for conscious AI. If the remit of 'computation' is broadened beyond Turing's definition, the distinction between scenarios 3 and 4 may blur, with more possibilities falling into scenario 3. Biological naturalism falls within scenarios 4 and 5 (and potentially 3, for 'computational biological naturalism'). [1](Graziano, 2017), [2](VanRullen & Kanai, 2021),[3] (Dehaene et al., 2017), [4](Ororbia & Friston, 2023), [5](Piccinini, 2020), [6](Kirkpatrick, 2022), [7](Findlay et al., 2024); note that IIT proposes that cause-effect structure matters (for consciousness), not functional organisation, [8](Morales Pantoja et al., 2023)*

	Scenario	Description	Assumes computational functionalism	Does the substrate matter?	Example
1	Naïve along for the ride	Consciousness will just emerge as AI gets smarter	Yes (Turing)	No	Large language models (LLMs)
2	Theory-based computational	Consciousness arises when computational theories of consciousness are implemented	Yes (Turing)	No	Attention-schema theory[1], global workspace theory[2], (some) higher-order thought theories[3]
3	Substrate-dependent computational	Consciousness depends on computations that can only be implemented in particular substrates	Yes (potentially wider than Turing)	Yes, to implement the relevant computations	Mortal computation[4], neuromorphic computation, neural computation[5], biological computation[6]
4	Substrate-dependent (weak)	Consciousness depends on non-computational functional organisation	No	Yes, to implement the relevant functional organisation	Non-computational neuromorphic approaches, implementations of dynamical theories, (IIT - for cause-effect structure, not function[7])
5	Substrate-dependent (strong)	Consciousness depends on (apparently) intrinsic properties of its biological basis	No	Yes, to enable substrate-specific functions, or in virtue of (apparently) intrinsic properties	Cerebral organoids, hybrid systems[8], synthetic biology

conscious AI is inevitable or possible. This part noted biases that may lead to over-attribution of consciousness to things and challenged assumptions of computational functionalism and silicon substrate flexibility on which conscious AI depends. The positive part argued for some form of biological naturalism, grounded in active inference, cybernetics, autopoiesis, and the free energy principle. If the negative arguments are on track, conscious AI will not be possible for the current substrate-independent trajectories of AI. If the positive arguments are on track, then real artificial consciousness will only be possible if we create machines that are also in some relevant sense alive.

Having laid out both parts, it's time to summarise the various scenarios in which conscious AI might arise. I will do this in stages from simple extrapolations of current AI to strong forms of biological naturalism, and by referencing – where possible – relevant theories of consciousness (Table 1). Where I mention a theory, a condition for conscious AI based on that theory is that it specifies true sufficient conditions for consciousness. This is a high bar. Not only are all current theories likely wrong (or at best incomplete), but many are restricted to describing potential mechanisms of consciousness where it is believed to exist, rather than proposing sufficient conditions in general (Butlin et al., 2023; Kuhn, 2024; Mudrik et al., 2025; Seth & Bayne, 2022).

5.1 *Naïve along for the ride*

Consciousness is a form of computation and will emerge naturally as AI gets smarter, perhaps at the threshold of AGI. I've argued that this scenario is mostly grounded in psychological biases and in underexamined assumptions regarding computational functionalism, and so is implausible.

5.2 *Theory-based computational*

Consciousness is a form of computation but will have to be explicitly designed in, based on computational theories of consciousness. Relevant theories here include global workspace theory, which proposes that consciousness depends on information sharing within a multimodal global neuronal workspace (Mashour et al., 2020); higher-order thought theories, in which consciousness depends on higher-order mental states representing the content of lower-order mental states as "being conscious" in some way (Brown, Lau, & LeDoux, 2019; Fleming, 2020); and attention schema theory, which proposes that consciousness emerges as a consequence of a model of the control of attention (Graziano, 2017). The plausibility of conscious AI in this scenario still depends on computational functionalism being true and on silicon substrate flexibility. Butlin et al. (2023) provide an excellent overview of this scenario, exploring the extent to which the principles of these theories may already be implicitly present in existing AI architectures, while Dung & Kersten (2024) offer a recent defence of computational views.

5.3 *Substrate-dependent computational*

Consciousness is computational, but the computations can only be implemented by specific substrates. Here, it could be that consciousness is a matter of mortal computation, in which the computations relevant for consciousness are inseparable from their substrate (Section 3.4). More generally, it could be a matter of natural law that the computations underlying consciousness can only be implemented in particular substrates. This becomes more plausible for broader interpretations of "computation" which allow for greater substrate dependency (N. G. Anderson & G. Piccinini, 2024; Kirkpatrick, 2022). A challenge for this scenario is that one motivation (albeit not the only one) for computational views of mind and consciousness is the broad multiple realisability which follows intuitively from substrate independence. Nor are there theories of consciousness which, to my knowledge, leverage this scenario.

This scenario contains the interesting possibility that only living systems can implement the relevant computations – a "computational biological naturalism"[37]. One motivation for this view

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

would be if evolution settled on computation as a powerful functional solution under various constraints (e.g., energy efficiency, degeneracy, robustness). In this case, substrate-dependent forms of (biological) computation might become more appealing as a ground for theories of consciousness. A challenge here would be to explain which kinds of computation (e.g., mortal, neural, post-quantum super-Turing, etc.) can only be implemented in living systems, and why.

In this general scenario, conscious AI might be achievable through forms of neuromorphic computation that are able to implement the sufficient (possibly mortal) computations. (If computational biological naturalism holds, then life would be necessary too.) Standard digital computation might still approximate these computations, but it would seem an open question whether these approximations would be good enough to realise consciousness. There are also open questions about whether it would matter if the computations implemented in the simulation differ from those being simulated.[38]

5.4 *Substrate-dependent, weak*

Consciousness depends on substrate-dependent non-computational functional properties. These could include continuous processes, field effects, stochastic effects, fine-grained timing relations, and other (non-computational) functional properties of neurobiological systems. This scenario includes functional properties which can only be implemented by certain substrates, but it excludes functions that are *defined in terms of* a particular material or substrate, such as metabolism, digestion, and the like. There is still a clear separation between substrate and function.

Here, we find the important possibility that only living systems can implement the necessary functional organisation for consciousness. This would be a "weak" form of biological naturalism, compatible with dynamical readings of the predictive processing and free energy principle outlined in Section 4. This possibility seems more appealing than the "computational biological naturalism" mentioned above, since in this case, the substrate properties encourage rather than discourage the view. Various dynamical theories of consciousness – like those mentioned in Section 3.6 – also fall within this scenario, as would electromagnetic field theories, which propose that electromagnetic fields provide the physical basis for conscious states [(McFadden, 2020); see also (Lane, 2022)].

In this scenario, consciousness is not a matter of computation, and the plausibility of conscious AI depends on whether the relevant functional properties can be implemented in an alternative substrate. Substrate flexibility, while possible in this scenario, seems challenging since the in-principle substrate independence of computation can no longer be leveraged. Computational simulations of these properties are no longer guaranteed to realise consciousness unless computational functionalism happens to be true anyway, and the simulations are good enough (which may be extremely challenging if detailed and highly optimised material properties of living systems turn out to be important).

The boundary between this scenario and the previous scenario depends on what counts as "computation." If broader senses of computation are allowed (Section 3.4), some non-computational functions become computational functions, but nothing much changes about the prospects for conscious AI since these broader senses of computation likely lack the required substrate flexibility.

5.5 *Substrate-dependent, strong*

Consciousness depends on intrinsic properties of its material biological basis. This scenario channels the intuition that consciousness inheres in the materiality of its substrate. There are (at least) two ways to interpret it. The first is a functional interpretation, in which the functions necessary for consciousness now include, by definition, aspects of the substrate (e.g., the function of digestion necessarily involves real food). The second bypasses functions, linking consciousness directly to intrinsic properties of its substrate in the spirit of Searle's original conception of biological naturalism. Following Searle, I too focus on living material as the relevant substrate.

My case for biological naturalism is drawn towards this scenario. The through-line from predictive processing and controlled hallucinations to the free energy principle, metabolism, and autopoiesis, seasoned with a sprinkle of real-problem humility, suggests a strong dependence of consciousness on biology. As far as I can tell, it is also agnostic between the two interpretations given above, because apparently intrinsic properties at one level might always be understandable in terms of patterns of functional organisation at more fine-grained levels. But if supposedly intrinsic metabolic, excitable, and autopoietic properties of living systems can be underwritten by finer-grained functional descriptions of other material components and processes, then so be it. Life would still matter, and it would do so thanks to the properties that warrant the term "life." Either way, instead of a division between substrate and function, there is now a deep continuity. In this scenario, consciousness is not a matter of computation, and conscious AI would need to be "living" AI.[39]

5.6 *Other theories*

By focusing on the spectrum between computational functionalism and biological naturalism, I have neglected theories of consciousness which do not fall naturally along this continuum. If these theories turn out to be sufficiently on track, they too have consequences for the plausibility of conscious AI and so are worth a brief mention.

Integrated information theory [IIT, (Albantakis et al., 2023; Tononi et al., 2016)] stands out among theories of consciousness by explicitly stating sufficient conditions for consciousness – and by being the focus of intense controversy (Gómez-Marin & Seth, 2025). According to IIT, a conscious experience is identical to the cause-effect structure unfolded from a physical substrate specifying a maximum of irreducible integrated information. Wherever there are such maxima, there will be consciousness. On this theory, real artificial consciousness will only be possible for systems that have the right kind of internal causal structure (Findlay et al., 2024). A purely feedforward neural network would have zero consciousness, whereas a neural network with the right kind of physically instantiated recurrent connectivity will have non-zero consciousness, regardless of what it is made out of. If IIT is right, conventional AI is not on a path to consciousness, but there is also nothing special about life.

Theories of consciousness based on quantum mechanics [e.g., (Hameroff & Penrose, 2014; Neven et al., 2024)] have also long been controversial (Koch & Hepp, 2006; Seth, 2012). Nevertheless, quantum mechanics is an exceptionally powerful physical theory, and quantum effects are known to be important in various

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

biological processes (McFadden & Al-Khalili, 2018; Mohseni, Omar, & Engel, 2013). It is possible that quantum effects may be implicated in biological processes related to consciousness, such as anaesthesia (Kalra et al., 2023; Li et al., 2018). However, even if so, this could be a contingent property of biological consciousness, rather than a necessary property of consciousness in general, and would have no implications for artificial consciousness. If quantum effects turn out to play a more fundamental role in consciousness (Hameroff & Penrose, 2014), then real artificial consciousness would presumably need to instantiate the necessary quantum-level effects, ruling out conventional AI. A related possibility is that consciousness depends on quantum forms of computation. In "quantum computational functionalism," consciousness could be realised by quantum computers but not by conventional computers (Neven et al., 2024). The plausibility of these scenarios depends on how seriously one takes the quantum theories in question.

5.7 *Embodiment and embeddedness*

Biological brains evolved, developed, and function within bodies, which are embedded within environments. AI systems are typically disembodied and may encounter their "environment" only passively. LLMs provide a good example, though these systems are increasingly being deployed in "agentic" contexts in which they interact with virtual environments (e.g., executing tasks in software). Some theories of consciousness explicitly require embodiment and embeddedness [e.g., (Damasio, 2000; Edelman, 1989; O'Regan & Noë, 2001; Seth, 2021; Solms, 2021)], and these properties have long been emphasised by enactive perspectives on cognitive science [see Section 3.6, and (Cosmelli, Lachaux, & Thompson, 2007; Froese, 2017; Kiverstein, 2020; Thompson, 2007; Varela, Thompson, & Rosch, 1993)].

The relevance of these properties to the prospects for conscious AI varies according to scenario. If computation is sufficient for consciousness, then virtual bodies and virtual environments – or non-biological robotic bodies – should also be sufficient. In substrate-dependent scenarios, embodiment and embeddedness may be necessary to enable relevant patterns of functional organisation. For example, physically embodied robots could have energetic and structural integrity requirements entailing real, though non-biological, forms of homeostasis and allostasis (but not autopoiesis).

In general, as reliance on substrate properties increases, so does the plausibility that embodiment and embeddedness might be necessary enabling conditions for the right kind of dynamics or non-computational functions. If my arguments in favour of biological naturalism are on track, then embodiment and embeddedness are likely very important and perhaps necessary. This is because these properties are central to the imperative to stay alive that motivates these arguments [see also (Aru et al., 2023)]. However, it is conceivable that consciousness may be a biological property, but that embodiment and embeddedness are not required. Cerebral organoids – which typically lack embodiment and embeddedness – provide an interesting test case here (Bayne, Seth, & Massimini, 2020; Jeziorski et al., 2023).[40]

Other questions hereabouts deserve deeper exploration. One question concerns *agency*. Agency in biological systems tends to mean more than the ability to perform actions in the service of goals (as in "agentic AI"). In biological systems, the relevant goals are typically endogenous, rather than being exogenously imposed (Jaeger et al., 2024). Rich embodiment and embeddedness might be needed for a system to possess or exhibit agency in this sense (Mitchell, 2023), and/or agency might be deeply ingrained into biological materials (Levin, 2023). Much like intelligence, AI "agents" and physical robots might be restricted to weak, observer-dependent forms of agency, in contrast to the strong, endogenous forms available to biological systems (see Section 3.8).

5.8 *Summary*

Of the scenarios reviewed above, I have tried to make a case for biological naturalism, and in particular for a stronger flavour in which living material is not just an enabling condition for some higher-level functional organisation, but in which there is a deep continuity between substrate and function, between matter and mind, and between life and consciousness. On this view, conscious AI would need to be (in some relevant sense) living AI, though not necessarily carbon-based, and not because of any lingering neo-vitalism. And just as there can be wind without a storm, there can be life without consciousness.

Having said this, it is difficult to reliably assess the relative plausibilities of these scenarios. My case is not watertight and may well be wrong in the details or wrong altogether. Laying out the various scenarios here serves two main purposes. It reminds us not to assume that conscious AI is inevitable, and it underlines the urgency of better understanding the mechanisms responsible for biological consciousness. Even if biological naturalism turns out to be false as an account of necessary conditions, a biological perspective may still be needed to understand the nature of consciousness in humans and other animals. And the more we know about consciousness where we can be reasonably certain it exists, the surer our footing will be elsewhere.

6.0 What should we (not) do?

We don't know what it would take to create conscious AI. We also don't know how to ensure it doesn't happen. I have made a case that conscious AI will need to share substrate properties with living systems, but I might be wrong. Other proposals, based on other theories, may also be wrong. We also lack reliable tests for the presence of consciousness beyond consensus cases in humans and (perhaps) some other animals (Bayne et al., 2024; Birch, 2024).[41] This is a challenging background against which to consider ethical implications, but given the rapid development and rollout of AI, such considerations are increasingly needed (Long et al., 2024; Schwitzgebel & Garza, 2023).

In framing these considerations, it is essential to distinguish AI systems that are conscious (real artificial consciousness) from systems that merely seem to be conscious. The latter are more likely than the former, and different ethical issues arise in each case.

6.1 *Real artificial consciousness*

Real artificial consciousness may be very unlikely or impossible. However, if such systems were to emerge – whether by design or by accident – ethical catastrophe awaits. A long philosophical tradition holds that being conscious endows moral status (Singer, 1975). With consciousness comes the potential for negatively valenced conscious experiences: for suffering. Creating real artificial consciousness risks a mass inauguration of new forms of suffering. The problem is compounded by the fact that we might not even recognise this suffering as suffering if it expresses itself in ways poorly aligned with our anthropocentric biases. One might respond that perhaps we could design real artificial consciousness so as to avoid negatively valenced experience. But this would be hard or

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

impossible to guarantee, and in any case it can be argued that any form of consciousness – valenced or not – endows moral status.[42]

There are other considerations. Real artificial consciousness might endow a system with its own interests and values, beyond those installed by the system's designers. This further complicates the already complicated "value alignment problem" of ensuring that AI acts in line with human interests and values (Russell & Norvig, 2003). There is the practical challenge of ensuring that potentially diverging interests remain compatible, even for AI systems that may exceed human-level capabilities in some or all domains. And there is the moral challenge of determining whether and when addressing the practical problem is the right thing to do, given the legitimacy of the AI's own interests.[43]

The development of real artificial consciousness for its own sake should not be an explicit goal.[44] This might seem obvious, but it is surprising how often "conscious AI" is treated as a holy grail. In practice, we ought to continually revise our credences about the presence of consciousness as technologies develop, as our understanding of biological consciousness deepens, and as our tests get better. And we should use these evolving credences to guide AI development away from the potential realisation of consciousness. The fact that real artificial consciousness may be impossible on current trajectories does not license us to blindly pursue it in the service of some poorly thought-through techno-rapture.

There is a tension here with the view that the development of socially useful AI may be enhanced by implementing some of the functions associated with consciousness. There is some justification for this view. Some of the failure zones of current AI arguably overlap with human abilities that seem intimately connected with consciousness. Examples include rapid generalisation to novel situations, effective metacognition, and learning from small quantities of data (Fleming, 2021; VanRullen & Kanai, 2021; Vong et al., 2024). AI systems able to deal effectively with these challenges may, on the whole, be desirable.

There is a line to walk. It seems reasonable that useful functions can be implemented in ways sufficiently different from the human (or more generally biological) case so as to minimise the risk that doing so will inadvertently instantiate consciousness. What counts as "sufficiently different" will depend on what theory of consciousness, and what scenario for artificial consciousness, you prefer. Notably, a similar line needs to be walked even for scientists using computational models to better understand the nature of biological consciousness (Seth, 2009). Again, where the line is drawn will depend on one's theoretical preferences. But the development of real artificial consciousness for its own sake remains a very bad idea (Seth, 2023).

6.2 *Conscious-seeming AI*

In contrast to the implausibility and uncertainty surrounding real artificial consciousness, it is almost certain that AI systems will soon give us compelling impressions of being conscious. Such systems may already be here, as responses to recent LLMs suggest – especially when these LLMs talk to us about their own "consciousness" (Shanahan & Singler, 2024). When sufficiently convincing LLMs are coupled with advanced generative video and audio avatars (and, looking further ahead, robots), we will share our world with artificial systems we cannot help feeling are conscious, whatever we may know or believe about the fact of the matter.

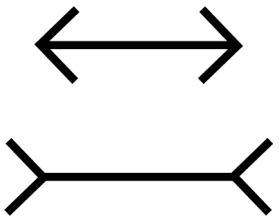


Figure 1. The Müller-Lyer illusion.

I want to emphasise the potential incorrigibility of impressions of artificial consciousness. Some visual illusions are cognitively impenetrable. The two lines in the Müller-Lyer illusion will always look different, even when you know they are the same length (Figure 1). Sufficiently advanced AI systems may give rise to cognitively impenetrable illusions of consciousness. This is worth highlighting because otherwise we might optimistically assume that expert knowledge or firmly held beliefs will inoculate us from feeling that non-conscious systems are conscious or from the consequences of doing so. Instead, whatever we know or believe, conscious-seeming AI systems are likely inevitable for all of us.[45]

Conscious-seeming AI raises its own ethical issues. These issues may be less existential than those attending real artificial consciousness, but the strong likelihood of conscious-seeming AI makes them pressing nonetheless.

AI that appears to be conscious is well placed to exploit our psychological vulnerabilities, whether by accident or by (corporate) design. If we believe that an LLM really understands us and really cares about us because we feel it is conscious, then we might be more inclined to follow its advice, even when this advice is bad. If we feel we are dealing with another mind rather than with an algorithm, we may also be more willing to sacrifice what's left of our digital privacy (Véliz, 2021).

These worries about vulnerability apply to real artificial consciousness too. But a consideration specific to conscious-seeming AI is what I have previously called the "WestWorld" scenario, based on the film and TV series [(Seth, 2021), see also (Chrisley, 2020; Seth, 2023)]. If we believe that a system is not conscious but nonetheless feel that it is, we are faced with a difficult choice. Either we decide to care about the system – to give it ethical and moral status even though we believe it is not conscious. Or we choose to not care about it, even though we still feel it has consciousness.

There are no good options here. If we decide to care about non-conscious systems, then we distort the circle of human moral concern, diverting attention and resources from other things that legitimately deserve them – including other humans and non-human animals. If we decide to *not* care about these systems, we risk brutalising our own minds. This is what happens in Westworld and is a line of thinking going back at least to Kant's lectures on ethics. If we treat systems that seem to have feelings as if they do not, we may end up becoming ethically insensitive to real feelings expressed by others – a coarsening of our moral sensibilities.

A particularly dystopian scenario involves conscious-seeming LLM-powered avatars of dead friends or relatives, trained on their personal data. This may seem far-fetched, but ideas like this are already being promoted.[46] The problems are too many to mention, but they include an extraordinary naïveté about the psychological importance of grief and forgetting, the corruption of autobiographical memories, and the opportunities for malfeasance ("I really think you should invest in this remarkable business opportunity," says dead-dad – or worse – dead-daughter).

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

There is a line to walk here too. Conscious-seeming AI may offer some benefits. It may be easier to interact with, enhancing efficiency and widening accessibility. There may be opportunities in areas such as individualised therapy and even in the alleviation of loneliness, however initially unappealing this may seem (Stade et al., 2024). One interesting option here is to treat LLMs as role-playing (Shanahan, McDonell, & Reynolds, 2023). This might induce some useful psychological distance, but this strategy might only get us so far.

I have described conscious-seeming AI as likely being inevitable. But it depends on the choices we collectively make. Alternative futures are possible, in which conscious-seeming AIs are restricted to well-justified use cases, and in which the bulk of AI development is engineered to resist rather than leverage our psychological biases. As Daniel Dennett said, "All we're going to see in our own lifetimes are intelligent tools, not colleagues. Don't think of them as colleagues, don't try to make them colleagues and, above all, don't kid yourself that they're colleagues."[47]

7.0 Conclusions

We humans have a habit of creating technologies in our own image and of projecting ourselves into the technologies we create. The conceptual miasma enveloping the relationship between consciousness and AI obscures our ability to see with clarity the threats and opportunities this technology represents. It also prevents us from seeing *ourselves* clearly – as the living, breathing, feeling, acting, thinking, and experiencing creatures that we are. If we conflate the richness of biological brains and human experience with the information-processing machinations of deepfake-boosted language models, then we overestimate the machines, and we underestimate ourselves.[48]

Consciousness will not just come along for the ride as AI gets smarter, despite the pull of our anthropocentric and anthropomorphic biases. There are good reasons to think that consciousness may not be possible at all in conventional AI systems, and there are at least some good (non-vitalistic) reasons to suppose that consciousness may be a property of, and perhaps only of, living systems. Simply noting that all known examples of conscious systems are biological and alive suggests that biological naturalism is no more speculative than computational functionalism.

Many uncertainties remain, especially in the arguments in favour of biological naturalism. Here, there are exciting opportunities to clarify the relationship between consciousness and life and, more specifically, among predictive processing, the free energy principle, autopoiesis, and the thermodynamics of metabolism. There is also much to be done in understanding whether and how neuromorphic and synthetic biology technologies move the needle on the potential for real artificial consciousness. These technologies are increasingly dissolving the distinction between hardware and software, and in so doing are approaching the relationship between mindware and wetware. They also highlight the importance of considering broader forms of computation beyond the canonical Turing machine.[49]

The importance of taking an informed ethical position despite these uncertainties spotlights another human habit: our unfortunate track record of withholding moral status from conscious things – from many non-human animals, and sometimes even from other humans. Will withholding attributions of consciousness to AI leave us once more on the wrong side of history? I believe the situation here is different. Our psychological biases are more likely to lead to false positives than false negatives. Compared to non-human animals, LLMs may be more similar to us in ways that do not matter for consciousness and less similar in ways that do.

Language models and generative deepfakes will have their day, but AI is just getting going. We do not know how to guarantee avoiding creating real artificial consciousness, and conscious-seeming systems do seem very likely. When, in the early 1800s, at the age of 19, Mary Shelley wrote her masterpiece *Frankenstein*, she provided a cautionary moral tale for the ages. The book is often read as a warning against the hubris of creating life. But the real Promethean sin of Frankenstein was not endowing his creature with life but with consciousness. It was the capacity of his creation to endure, to suffer, to feel jealousy and misery and anger that gave rise to the horror and the carnage.[50] We should not make the same mistake, whether in reality or even in appearance.

Acknowledgements. I am grateful for the generous comments and advice of many people including Jaan Aru, Adam Barrett, Tim Bayne, Ned Block, Simon Bowes, Chris Buckley, David Chalmers (and the NYU weekly philosophy seminar group), Shamil Chandaria, Robert Chis-Ciure, Stephen Fleming, Karl Friston, Jakob Hohwy (and the Monash Centre for Consciousness and Contemplative Studies), Johannes Kleiner, Christof Koch, Nicolas Kuske, Nick Lane, Shane Legg, Alexander Lerchner, Michael Levin, Liad Mudrik, Angus Nisbet, Thomas Parr, Gualtiero Piccinini, Claudia Passos, Michael Pollan, Aza Raskin, Luke Roelofs, Adam Rostowski, Michael Silberstein, Murray Shanahan, Ray Tallis, and my colleagues at the Sussex Centre for Consciousness Science. I am also grateful to four anonymous reviewers for helpful comments.

Financial support. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 101019254 for project CONSCIOUS).

Competing interests. The author is an advisor to Conscium Ltd and AllJoined Inc.

Notes

1 https://spectrum.ieee.org/ai-agents

2 For example, the ability to experience regret, as compared to mere disappointment, plausibly requires the ability to reason about the future consequences of actions and to decide between competing action plans. In general, whether intelligence and consciousness look independent will depend on whether one asks the question logically, nomologically (given the laws of nature as they are in our world, but not in all possible worlds), or in terms of systems – biological or otherwise – that actually exist. Plausibly, consciousness and intelligence are logically and nomologically independent.

3 In many countries, there is a warning engraved into the wing mirrors of cars: "objects in the mirror are closer than they appear." I have always found this philosophically interesting.

4 Strictly, Kleene (1956) showed that McCulloch-Pitts networks were equivalent to finite state automata, which are less computationally powerful than Turing machines. They become equivalent to Turing machines when provided with unbounded storage (and read-write access). Thanks to a reviewer for pointing this out.

5 Turing understood that questions about consciousness were distinct from questions about (machine) intelligence: "I do not wish to give the impression that I think there is no mystery to consciousness … But I do not think these mysteries need to be solved before we can answer the question with which we are concerned in this paper" (Turing, 1950) (p. 446). Thanks to Jakob Hohwy for this gem.

6 Another perspective on the remit of Turing computation is provided by the "physical Church-Turing thesis" (PCTT). There are several interpretations of the PCTT, all of which remain matters of conjecture. My arguments in this paper are compatible with a "modest" PCTT, according to which any function that is *computable* by a physical system is computable by a Turing machine. This is distinct from "bold" versions of the PCTT, which say (unconvincingly, in my

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

opinion) that anything that a physical system *does* is computable by a Turing machine. See (N. G. Anderson & G. Piccinini, 2024; Chalmers, 1996b).

7 Information in the framework of Shannon is best thought of as a descriptive method for quantifying uncertainty (Shannon & Weaver, 1949). Information theory can be applied widely without assuming that everything it applies to "processes information", just as things can be modelled computationally without being computational. See (N. G. Anderson & G. Piccinini, 2024; Chalmers, 1994; Floridi, 2010).

8 https://www.vox.com/future-perfect/351893/consciousness-ai-machines-neuroscience-mind

9 There is lively philosophical discussion over what multiple realisability actually means (Polger & Shapiro, 2016), how one might test for it (Shapiro, 2008), and how it relates to substrate independence (and flexibility). According to Polger and Shapiro, multiple realisability holds when "relevantly the same function is performed in relevantly different ways" (2016, p. 77). Intuitively, substrate flexibility can be thought of as a "stronger" version of multiple realisability that requires invariance of functional organisation across, and not merely within, substrates. However, on Polger and Shapiro's account, it is logically possible to have substrate flexibility without any multiple realisability if the alternative substrate is (and has to be) causally isomorphic in all the relevant ways. Whether this is nomologically possible for mental states is, I think, doubtful, as illustrated by criticisms of the neural replacement thought experiment (see below). For our purposes it suffices to think of substrate flexibility as a form of multiple realisability that requires invariance across substrates. I am grateful to a reviewer for raising these interesting points.

10 Chalmers includes axons and other aspects of neural connectivity in his version.

11 This point is not restricted to nitric oxide. Monoaminergic neurotransmitters such as dopamine and noradrenaline also act over large neural areas.

12 This example is from Ian Hinckfuss (Shiller, 2024); see (Chalmers, 1996b; Dung & Kersten, 2024) for discussion.

13 These (nomological) objections also apply to Chalmers' "dancing qualia" argument, in which control of a human agent is switched back and forth between the real biological brain and a (presumed) functional duplicate (Chalmers, 1995; Godfrey-Smith, 2016).

14 Someone might say, "Of course it's difficult to build a silicon brain that replicates everything that matters for mental states, but let's just imagine we can." But it is hard to draw useful conclusions about the world we live in from "just imagining" situations for which there are good reasons to believe cannot (nomologically) be the case. This is a weakness with conceivability arguments in general. The (in)famous zombie argument suffers similarly (Chalmers, 1996a; Seth, 2021).

15 Kleiner (2024) reaches a similar conclusion using a proof-based method. See also (Kleiner & Ludwig, 2023).

16 The Antikythera mechanism – a complex piece of machinery used to predict planetary movements – is arguably an analogue computer. See https://www.wired.com/story/unbelievable-zombie-comeback-analog-computing/.

17 See https://corticallabs.com/cl1.html

18 A fascinating example comes from early work by Adrian Thompson using genetic algorithms to optimise field programmable gate array (FPGA) chips to perform various functions. He found that some configurations implicitly utilised spatial relationships between FPGA components and the chip's power supply. The function of the chip, therefore, depended on its precise spatial layout, severely constraining multiple realisability. Other FPGA configurations turned out to depend on ambient room temperature – a constraint that Adrian tried to overcome by optimising his FPGAs in fridges set at different temperatures (Thompson, 1998).

19 The idea here is that the strict software/hardware distinction imposed by Turing computation implies an observer dependency about the functions implemented in software. This arises because of the need for an observer to attribute meaning (semantics) to syntactical manipulations. Computational functionalism would then imply that consciousness is also observer-dependent, which it isn't. Thanks to Simon Bowes and Ray Tallis for reminding me of this argument.

20 See Kleiner & Ludwig (2023) for an argument connecting dynamical aspects of neural activity to the implausibility of conscious AI.

21 If X is a non-Turing computational process, a (Turing) simulation will typically involve different computations, even if the overall input-output mapping is preserved. Whether this matters for consciousness depends on the specificity with which computational functionalism is interpreted: does the precise computation matter, or only the overall computational behaviour? There is room for different opinions.

22 Wouldn't a simulated rainstorm feel wet for a simulated agent within the simulation? This objection suggests that whether or not we recognise a simulated rainstorm as simulated depends on whether we're experiencing the storm from "inside" or "outside" the simulation (Chalmers, 2022; Hofstadter, 1981). The objection doesn't work because it presupposes that computation is sufficient for experience; see also (Wiese, 2024).

23 As an apocryphal quote puts it: "Computers don't actually 'play chess'; instead, they play 'the history of chess'."

24 Belief in the Bayesian sense does not connote a (personal-level) psychological belief. Bayesian beliefs are probability distributions, usually characterised by a mean and a variance or precision.

25 Predictive processing in general is agnostic as to whether (conscious) perceptual contents are constituted by perceptual best guesses or are merely best modelled in this way. See Section 4.2 and (Andrews, 2021; Clark, 2023; Friston, 2010; Hohwy & Seth, 2020).

26 Allostasis is a generalisation of homeostasis that emphasises stability through change: for example, our blood pressure transiently rises as we stand up to prevent us from fainting (Sterling, 2012).

27 Another angle here is that temporal aspects of conscious experience – the "thickness" associated with Husserlian protention and retention (Husserl, 1982 [1913]) – might be associated with the temporal extension of predictive inference, especially when inference is construed as minimisation of free energy over time; see Section 4.2 and (Friston, 2018; Kiverstein, 2020).

28 See also (Thompson & Varela, 2001). For a different take on minimal phenomenal experience, see (Metzinger, 2024).

29 Are there examples of human-designed autopoietic systems? Synthetic biological systems would come closest. But these arguably fall short because, to date, the (autopoietic) component elements – cells – are at best modified by humans rather than created *de novo*.

30 These varieties of continuity recall the "mind-life" continuity thesis, which – in its stronger formulations – proposes that "mind is literally life-*like*" [(Godfrey-Smith, 1996), p.320, emphasis in original, see also (Thompson, 2007)]. See Kirchhoff & Froese (2017) for a discussion of mind-life continuity, the free energy principle, and autopoiesis.

31 See chapter 10 in Seth (2021) for a slower version of this journey; Kiverstein (2020); Solms (2021) for related accounts; and Bogacz (2017); Buckley et al. (2017); Friston (2010); Nave (2025); and Parr et al. (2022) for more on the free energy principle.

32 Another way to understand this: living systems capitalise on the possibilities afforded by the *fluctuation theorem*, which quantifies the probability with which isolated systems can transiently decrease in entropy (Evans et al., 1993).

33 Strictly, free energy provides an upper bound on *surprisal*, which specifies how (statistically) unexpected something is, given a model. The long-term average of surprisal (under non-equilibrium steady-state assumptions) is entropy. See Buckley et al. (2017) and Parr et al. (2022).

34 The main assumption is that the probability distributions involved can be parameterised as Gaussian distributions. Under this (Laplace) assumption, prediction errors are associated with free energy gradients, rather than with free energies *per se*. See Buckley et al. (2017), Friston & Kiebel (2009), and Parr et al. (2022).

35 This holds because (variational) free energy provides a lower bound on Bayesian model evidence (the evidence lower bound, ELBO). Minimising free energy is equivalent to maximising the ELBO. See Parr et al. (2022).

36 Lane suggests the readout is provided by the ratio of electrostatic and electromagnetic potential across the membrane (Lane, 2022). See also Alexandre (2010) and Egbert et al. (2010).

37 Hat-tip to David Chalmers for this phrase. Wanja Wiese recently suggested a similar view, calling it "weak computational functionalism" (Wiese, 2024).

38 Kleiner (2024) argues that if computational functionalism is true, the relevant computations must be mortal, such that standard (immortal) AI cannot be conscious. But then, at least on some readings of computational functionalism, it should also be true that a (good enough) simulation of the relevant

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

mortal computations, using immortal computational methods, would also realise consciousness. This presents an apparent contradiction. One resolution would be to say that simulating (mortal) computations is not sufficient, perhaps because the computations are different (see note 21). The contradiction is also resolved if computational functionalism is false. And if computational functionalism is false, then conscious AI is not possible anyway.

39 For an alternative account linking life and consciousness via the free energy principle, see (Wiese, 2024). The core idea here (roughly) is that consciousness requires a match between a system's computational dynamics and its physical dynamics.

40 Another possibility, mostly relevant to LLMs, is that embodiment and embeddedness may be required for genuine *understanding*, independently of any implications for consciousness.

41 Once more with feeling: the Turing Test is a test of intelligence, not a test of consciousness. The idea of a similar test for consciousness has been called a Garland Test, in honour of Alex Garland's film *Ex Machina* (Shanahan, 2016); see also Seth (2021) (ch.13). See Schneider (2019) for one proposal for a test for consciousness in AI.

42 See Peter Godfrey-Smith's distinction between "narrow sentientism" and "broad sentientism": https://metazoan.net/108-whitehead-lectures/. Farisco et al. (2024) point out that putative real artificial conscious states might be qualitatively different from the human form.

43 Complicating things yet further, Long and colleagues raise the possibility that frustrating the interests or "preferences" of even unconscious AI may still be ethically problematic (Long et al., 2024). These problems might be somewhat resolved by Yoshua Bengio's suggestion that an appropriate goal for AI is an oracle: wise, but without its own interests or agency. See https://yoshuabengio.org/2023/05/07/ai-scientists-safe-and-useful-ai/.

44 Some go as far as calling for a moratorium on any research that may involve "synthetic phenomenology" (Metzinger, 2021).

45 An alternative is that AI systems such as LLMs become so good that they no longer seem humanlike. Being "too good to be true," they might emerge from the uncanny valley and head somewhere else altogether. I had an experience like this when interacting with DeepSeek-R1 in early 2025.

46 https://www.technologyreview.com/2024/05/07/1092116/deepfakes-dead-chinese-business-grief/

47 https://www.ft.com/content/96187a7a-fce5-11e6-96f8-3700c5664d30

48 See Floridi & Nobre (2024) for more on the damaging consequences of anthropomorphising machines and "computerising" minds.

49 These new developments highlight the reciprocal nature of biological metaphors. While our understanding of biological phenomena has been shaped by machine/computer metaphors, insights from biology and neuroscience in turn have shaped our understanding of machines and computers (Barwich & Rodriguez, 2024).

50 https://podcasts.apple.com/gb/podcast/frankenstein-with-anil-seth-and-fiona-sampson/id1549179379?i=1000637016221

References

Aguilera, M., Millidge, B., Tschantz, A., & Buckley, C. L. (2022). How particular is the physics of the free energy principle? *Physics of Life Reviews*, *40*, 24–50. https://doi.org/10.1016/j.plrev.2021.11.001

Alais, D., Xu, Y., Wardle, S. G., & Taubert, J. (2021). A shared mechanism for facial expression in human faces and face pareidolia. *Proceedings of the Biological Sciences*, *288*(1954), 20210966. https://doi.org/10.1098/rspb.2021.0966

Albantakis, L., Barbosa, L., Findlay, G., Grasso, M., Haun, A. M., Marshall, W., Mayner, W. G. P., Zaeemzadeh, A., Boly, M., Juel, B. E., Sasai, S., Fujii, K., David, I., Hendren, J., Lang, J. P., & Tononi, G. (2023). Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms. *PLOS Computational Biology*, *19*(10), e1011465. https://doi.org/10.1371/journal.pcbi.1011465

Alexandre, G. (2010). Coupling metabolism and chemotaxis-dependent behaviours by energy taxis receptors. *Microbiology (Reading)*, *156*(Pt 8), 2283–2293. https://doi.org/10.1099/mic.0.039214-0

Anderson, N. G., & Piccinini, G. (2024). *The physical signature of computation: A robust mapping account*. Oxford University Press.

Andrews, M. (2021). The math is not the territory: navigating the free energy principle. *Biology and Philosophy*, *36*, 30.

Aru, J., Larkum, M. E., & Shine, J. M. (2023). The feasibility of artificial consciousness through the lens of neuroscience. *Trends in Neurosciences*, *46*(12), 1008–1017. https://doi.org/10.1016/j.tins.2023.09.009

Ball, P. (2023). *How life works: A user's guide to the new biology*. University of Chicago Press.

Barnett, L., & Seth, A. K. (2023). Dynamical independence: Discovering emergent macroscopic processes in complex dynamical systems. *Physical Review E*, *108*(1), 014304. https://doi.org/10.1103/PhysRevE.108.014304

Barrett, L. F., & Simmons, W. K. (2015). Interoceptive predictions in the brain. *Nature Reviews Neuroscience*, *16*(7), 419–429. https://doi.org/10.1038/nrn3950

Barwich, A.-S., & Rodriguez, M. J. (2024). Rage against the what? The machine metaphor in biology. *Biology and Philosophy*, *39*, 14.

Bayne, T., Seth, A. K., & Massimini, M. (2020). Are there islands of awareness? *Trends in Neurosciences*, *43*(1), 6–16. https://doi.org/10.1016/j.tins.2019.11.003

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 in Cognitive Sciences*, *28*(5), 454–466. https://doi.org/10.1016/j.tics.2024.01.010

Bechtel, W., & Mundale, J. (1999). Multiple realizability revisited: Linking cognitive and neural states. *Philosophy of Science*, *LXVI*, 175–207.

Biehl, M., Pollock, F. A., & Kanai, R. (2021). A Technical Critique of Some Parts of the Free Energy Principle. *Entropy (Basel)*, *23*(3), 293. https://doi.org/10.3390/e23030293

Birch, J. (2024). *The edge of sentience*. Oxford University Press.

Block, N. (2019). Fading qualia: A response to Michael Tye. In A. Pautz & D. Stoljar (Eds.), *Blockheads! Essays on Ned Block's philosophy of mind and consciousness* (pp. 571–580). MIT Press.

Blum, L., & Blum, M. (2024). AI Consciousness is Inevitable: A Theoretical Computer Science Perspective. *ArXiv*, arXiv:2403.17101. https://doi.org/10.48550/arXiv.2403.17101

Boden, M. (2008). *Mind as machine: A history of cognitive science*. Oxford University Press.

Bogacz, R. (2017). A tutorial on the free-energy framework for modelling perception and learning. *The Journal of Mathematical Psychology*, *76*(Pt B), 198–211. https://doi.org/10.1016/j.jmp.2015.11.003

Bostrom, N. (2003). Are you living in a computer simulation? *Philosophical Quarterly*, *53*(211), 243–255.

Brown, R., Lau, H., & LeDoux, J. E. (2019). Understanding the higher-order approach to consciousness. *Trends in Cognitive Sciences*, *23*(9), 754–768. https://doi.org/10.1016/j.tics.2019.06.009

Bruineberg, J., Dolega, K., Dewhurst, J., & Baltieri, M. (2021). The Emperor's new markov blankets. *Behavioral and Brain Sciences*, *45*, e183. https://doi.org/10.1017/S0140525X21002351

Buckley, C., Kim, C.-S., McGregor, S., & Seth, A. K. (2017). The free energy principle for action and perception: A mathematical review. *Journal of Mathematical Psychology*, *81*, 55–79.

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. https://doi.org/10.48550/arXiv.2308.08708

Cantwell Smith, B. (2002). The Foundations of Computing. In M. Scheutz (Ed.), *Computationalism: New directions* (pp. 23–58). MIT Press.

Cao, R. (2022). Multiple realizability and the spirit of functionalism. *Synthese*, *200*, 506.

Chalmers, D. (1995). Absent qualia, fading qualia, dancing qualia. In T. Metzinger (Ed.), *Conscious experience* (pp. 309–328). Ferdinand Schoningh.

Chalmers, D. J. (1994). On implementing a computation. *Minds and Machines*, *4*, 391–402.

Chalmers, D. J. (1996a). *The conscious mind: In search of a fundamental theory*. Oxford University Press.

Chalmers, D. J. (1996b). Does a rock implement every finite-state automaton? *Synthese*, *108*(3), 309–333.

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

Chalmers, D. J. (2003). Consciousness and its place in nature. In S. Stich & T. A. Warfield (Eds.), *The blackwell guide to the philosophy of mind* (pp. 103–140). Blackwell Publishing Ltd.

Chalmers, D. J. (2022). *Reality+: Virtual worlds and the problems of philosophy*. W. W. Norton & Company.

Chalmers, D. J. (2023). Could a Large Language Model be Conscious? https://arxiv.org/abs/2303.07103

Chintaluri, C., & Vogels, T. P. (2023). Metabolically regulated spiking could serve neuronal energy homeostasis and protect from reactive oxygen species. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(48), e2306525120. https://doi.org/10.1073/pnas.2306525120

Chrisley, R. (2020). A human-centered approach to AI ethics: A perspective from cognitive science. In M. Dubber, F. Pasquale, & S. Das (Eds.), *The Oxford handbook of ethics of AI* (pp. 463–474). Oxford University Press.

Churchland, P. S. (1996). The hornswoggle problem. *Journal of Consciousness Studies*, *3*(5–6), 402–408.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Clark, A. (2019). Consciousness as generative entanglement. *The Journal of Philosophy*, *116*(12), 645–662.

Clark, A. (2023). *The experience machine: How our minds predict and shape reality*. Allen Lane.

Cobb, M. (2020). *The idea of the brain: A history*. Profile Books.

Cole, D. (2023). The Chinese room argument. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy (Winter 2024 Edition)*. https://plato.stanford.edu/archives/sum2023/entries/chinese-room/

Collell, G., & Fauquet, J. (2015). Brain activity and cognition: A connection from thermodynamics and information theory. *Frontiers in Psychology*, *6*, 818.

Colombatto, C., & Fleming, S. M. (2024). Folk psychological attribuitions of consciousness to large language models. *Neuroscience of Consciousness*, *1*, niae013. https://doi.org/10.1093/nc/niae013

Conant, R., & Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, *1*(2), 89–97.

Copeland, J. B. (2024). The Church-Turing thesis. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University.

Cosmelli, D., Lachaux, J.-P., & Thompson, E. (2007). Neurodynamics of consciousness. In P. D. Zelazo, M. Moscovitch, & E. Thompson (Eds.), *The Cambridge handbook of consciousness* (pp. 731–775). Cambridge University Press.

Crick, F., & Koch, C. (1990). Towards a neurobiological theory of consciousness. *Seminars in the Neurosciences*, *2*, 263–275.

Damasio, A. (2000). *The feeling of what happens: Body and emotion in the making of consciousness*. Harvest Books.

Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it?. *Science*, *358*(6362), 486–492. https://doi.org/10.1126/science.aan8871

Dennett, D. (1987). *The intentional stance*. MIT Press.

Di Paolo, E., Thompson, E., & Beer, R. (2022). Laying down a forking path: Tensions between enaction and the free energy principle. *Philosophy and the Mind Sciences*, *3*. https://doi.org/10.33735/phimisci.2022.9187

Douglas, R., Mahowald, M., & Mead, C. (1995). Neuromorphic analogue VLSI. *Annual Review of Neuroscience*, *18*, 255–281.

Dung, L., & Kersten, L. (2024). Implementing artificial consciousness. *Mind and Language*, *40*(3), 285–305. https://doi.org/10.1111/mila.12532

Dupuy, J.-P. (2009). *On the origins of cognitive science: The mechanization of mind* (2nd ed.). MIT Press.

Ebrahimkhani, M. R., & Levin, M. (2021). Synthetic living machines: A new window on life. *iScience*, *24*(5), 102505. https://doi.org/10.1016/j.isci.2021.102505

Edelman, G. M. (1989). *The remembered present*. Basic Books.

Egbert, M. D., Barandiaran, X. E., & Di Paolo, E. A. (2010). A minimal model of metabolism-based chemotaxis. *PLOS Computational Biology*, *6*(12), e1001004. https://doi.org/10.1371/journal.pcbi.1001004

Evans, D. J., Cohen, E. G., & Morriss, G. P. (1993). Probability of second law violations in shearing steady states. *Physical Review Letters*, *71*(15), 2401–2404. https://doi.org/10.1103/PhysRevLett.71.2401

Evers, K., Farisco, M., Chatila, R., Earp, B. D., Freire, I. T., Hamker, F., Nemeth, E., Verschure, P., & Khamassi, M. (2025). Preliminaries to artificial consciousness: A multidimensional heuristic approach. *Physics of Life Reviews*, *52*, 180–193. https://doi.org/10.1016/j.plrev.2025.01.002

Farisco, M., Evers, K., & Changeux, J. P. (2024). Is artificial consciousness achievable? Lessons from the human brain. *Neural Networks*, *180*, 106714. https://doi.org/10.1016/j.neunet.2024.106714

Fields, C., Goldstein, A., & Sandved-Smith, L. (2024). Making the thermodynamic cost of active inference explicit. *Entropy (Basel)*, *26*(8), 622. https://doi.org/10.3390/e26080622

Findlay, G., Marshall, W., Albantakis, L., David, I., Mayner, W. G. P., Koch, C., & Tononi, G. (2024). Dissociating artificial intelligence from artificial consciousness. *ArXiv*, *2412*, 04571 https://doi.org/10.48550/arXiv.2412.04571

Fleming, S. M. (2020). Awareness as inference in a higher-order state space. *Neuroscience of Consciousness*, *2020*(1), niz020. https://doi.org/10.1093/nc/niz020

Fleming, S. M. (2021). *Know thyself*. Basic Books.

Floridi, L. (2010). *Information: A very short introduction*. Oxford University Press.

Floridi, L., & Nobre, A. C. (2024). Anthropomorphising machines and computerising minds: The crosswiring of languages between artificial intelligence and brain & cognitive sciences. *Minds and Machines*, *34*, 5. https://doi.org/10.1007/s11023-024-09670-4

Freeman, W. J. (1999). Consciousness, intentionality, and causality. *Journal of Consciousness Studies*, *6*(11–12), 143–172.

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society of London B, Biological Sciences*, *364*(1521), 1211–1221. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=19528002

Friston, K. J. (2010). The free-energy principle: a unified brain theory? *Nat Rev Neurosci*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787

Friston, K. J. (2013). Life as we know it. *Journal of the Royal Society Interface*, *10*(86), 20130475. https://doi.org/10.1098/rsif.2013.0475

Friston, K. J. (2018). Am I self-conscious? (Or does self-organization entail self-consciousness?). *Frontiers in Psychology*, *9*, 579. https://doi.org/10.3389/fpsyg.2018.00579

Friston, K. J., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: A process theory. *Nature Neuroscience*, *29*(1), 1–49. https://doi.org/10.1162/NECO_a_00912

Froese, T. (2017). Life is precious because it is precarious: Individuality, mortality and the problem of meaning. In G. Dodig-Crnkovic & E. Giovagnoli (Eds.), *Representation and reality in humans, other living organisms, and intelligent machines* (pp. 33–50). Springer International Publishing.

Godfrey-Smith, P. G. (1996). Spencer and Dewey on life and mind. In M. Boden (Ed.), *The philosophy of artificial life* (pp. 314–331). Oxford University Press.

Godfrey-Smith, P. G. (2016). Mind, matter, and metabolism. *Journal of Philosophy*, *CXIII*(10), 481–506.

Goertzel, B., & Pennachin, C. (Eds.). (2007). *Artificial general intelligence*. Springer-Verlag.

Gómez-Marin, A., & Seth, A. K. (2025). A science of consciousness beyond pseudo-science and pseudo-consciousness. *Nature Neuroscience*, *28*, 703–706.

Graziano, M. S. A. (2017). The attention schema theory: A foundation for engineering artificial consciousness. *Frontiers in Robotics and AI*, *4*, 60. https://doi.org/10.3389/frobt.2017.00060

Gumuskaya, G., Srivastava, P., Cooper, B. G., Lesser, H., Semegran, B., Garnier, S., & Levin, M. (2024). Motile living biobots self-construct from adult human somatic progenitor seed cells. *Advanced Science (Weinh)*, *11*(4), e2303575. https://doi.org/10.1002/advs.202303575

Haeckel, E. (1892). Our monism: The principles of a consistent, unitary, world-view. *The Monist*, *2*, 481–486.

Hameroff, S., & Penrose, R. (2014). Consciousness in the universe: A review of the 'Orch OR' theory. *Physics of Life Reviews*, *11*(1), 39–78. https://doi.org/10.1016/j.plrev.2013.08.002

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

Haugeland, J. (1980). Programs, causal powers and intentionality. *Behavioral and Brain Sciences*, *3*, 432–433.
Hinton, G. (2022). The forward-forward algorithm: some preliminary investigations. https://arxiv.org/abs/2212.13345
Hochstetter, J., Zhu, R., Loeffler, A., Diaz-Alvarez, A., Nakayama, T., & Kuncic, Z. (2021). Avalanches and edge-of-chaos learning in neuromorphic nanowire networks. *Nature Communications*, *12*(1), 4008. https://doi.org/10.1038/s41467-021-24260-z
Hofstadter, D. (1981). A coffee-house conversation on the Turing test. *Scientific American*.
Hohwy, J. (2013). *The Predictive Mind*. Oxford University Press.
Hohwy, J. (2014). The self-evidencing brain. *Nous*, *50*(2), 259–285. https://doi.org/10.1111/nous.12062
Hohwy, J., & Seth, A. K. (2020). Predictive processing as a systematic basis for identifying the neural correlates of consciousness. *Philosophy and the Mind Sciences*, *1*(2), 3.
Husserl, E. (1982 [1913]). *Ideas pertaining to a pure phenomenology and to a phenomenological philosophy - first book: General introduction to a pure phenomenology*.
Jaeger, J., Riedl, A., Djedovic, A., Vervaeke, J., & Walsh, D. (2024). Naturalizing relevance realization: Why agency and cognition are fundamentally not computational. *Frontiers in Psychology*, *15*, 1362658. https://doi.org/10.3389/fpsyg.2024.1362658
Jarzynski, C. (1997). Nonequilibrium equality for free energy differences. *Physical Review Letters*, *78*(14), 2690.https://doi.org/10.1103/PhysRevLett.78.2690
Jaynes, E. T. (1957). Information theory and statistical mechanics. *The Physical Review*, *106*(4), 620–630.
Jeziorski, J., Brandt, R., Evans, J. H., Campana, W., Kalichman, M., Thompson, E., Goldstein, L., Koch, C., & Muotri, A. R. (2023). Brain organoids, consciousness, ethics and moral status. *Seminars in Cell & Developmental Biology*, *144*, 97–102. https://doi.org/10.1016/j.semcdb.2022.03.020
Jonas, H. (2001). *The phenomenon of life: Towards a philosophical biology*. Northwestern University Press.
Kagan, B. J., Kitchen, A. C., Tran, N. T., Habibollahi, F., Khajehnejad, M., Parker, B. J., Bhat, A., Rollo, B., Razi, A., & Friston, K. J. (2022). In vitro neurons learn and exhibit sentience when embodied in a simulated game-world. *Neuron*, *110*(23), 3952–3969.e8. https://doi.org/10.1016/j.neuron.2022.09.001
Kagan, B. J., Razi, A., Bhat, A., Kitchen, A. C., Tran, N. T., Habibollahi, F., Khajehnejad, M., Parker, B. J., Rollo, B., & Friston, K. J. (2023). Scientific communication and the semantics of sentience. *Neuron*, *111*(5), 606–607. https://doi.org/10.1016/j.neuron.2023.02.008
Kalra, A. P., Benny, A., Travis, S. M., Zizzi, E. A., Morales-Sanchez, A., Oblinsky, D. G., Craddock, T. J. A., Hameroff, S. R., MacIver, M. B., Tuszynski, J. A., Petry, S., Penrose, R., & Scholes, G. D. (2023). Electronic energy migration in microtubules. *ACS Central Science*, *9*(3), 352–361. https://doi.org/10.1021/acscentsci.2c01114
Kirchhoff, M. D., & Froese, T. (2017). Where there is life there is mind: In support of a strong life-mind continuity thesis. *Entropy*, *19*(4). 169 https://doi.org/ARTN10.3390/e19040169
Kirkpatrick, K. L. (2022). Biological computation: hearts and flytraps. *Journal of Biological Physics*, *48*(1), 55–78. https://doi.org/10.1007/s10867-021-09590-9
Kiverstein, J. (2020). Free energy and the self: an ecological-enactive interpretation. *Topoi*, *39*, 559–574.
Kleene, S. C. (1956). Representation of events in nerve nets and finite automata. In C. Shannon & J. McCarthy (Eds.), *Automata Studies* (pp. 3–42). Princeton University Press.
Kleiner, J. (2024). Consciousness requires mortal computation. https://arxiv.org/abs/2403.03925
Kleiner, J., & Ludwig, T. (2023). If consciousness is dynamically relevant, artificial intelligence isn't conscious. https://arxiv.org/abs/2304.05077
Koch, C., & Hepp, K. (2006). Quantum mechanics in the brain. *Nature*, *440*(7084), 611. https://doi.org/10.1038/440611a
Kuhn, R. L. (2024). A landscape of consciousness: Toward a taxonomy of explanations and implications. *Progress in Biophysics and Molecular Biology*, *190*, 28–169.
Kurzweil, R. (2005). *The singularity is near*. Viking.
Kurzweil, R. (2024). *The singularity is nearer: When we merge with AI*. Bodley Head.
Lamme, V. A. (2010). How neuroscience will change our view on consciousness. *Cognitive Neuroscience*, *1*(3), 204–240.
Lamme, V. A. (2018). Challenges for theories of consciousness: Seeing or knowing, the missing ingredient and how to deal with panpsychism. *Philosophical Transactions of the Royal Society B-Biological Sciences*, *373*, 20170344.
Landauer, R. (1961). Irreversibility and heat generation in the computing process. *IBM Journal of Research and Development*, *5*(3), 183–191.
Lane, N. (2015). *The vital question: Why is life the way it is?* Profile Books.
Lane, N. (2022). *Transformer: The deep chemistry of life and death*. WW Norton/Profile.
Legg, S., & Hutter, M. (2007). A collection of definitions of intelligence. *Frontiers in Artificial Intelligence and Applications*, *157*, 17–24. https://arxiv.org/abs/0706.3639v1
Levin, M. (2023). Darwin's agential materials: Evolutionary implications of multiscale competency in developmental biology. *Cellular and Molecular Life Sciences*, *80*, 142. https://doi.org/10.1007/s00018-023-04790-z
Levin, M. (2025). Ingressing Minds: Causal Patterns Beyond Genetics and Environment in Natural, Synthetic, and Hybrid Embodiments. *PsyArXiv*. https://doi.org/10.31234/osf.io/5g2xj_v1
Li, N., Lu, D., Yang, L., Tao, H., Xu, Y., Wang, C., Fu, L., Liu, H., Chummum, Y., & Zhang, S. (2018). Nuclear spin attenuates the anesthetic potency of xenon isotopes in mice: Implications for the mechanisms of anesthesia and consciousness. *Anesthesiology*, *129*(2), 271–277. https://doi.org/10.1097/ALN.0000000000002226
Llinas, R. R., Ribary, U., Contreras, D., & Pedroarena, C. (1998). The neuronal basis for consciousness. *Philosophical Transactions of the Royal Society of London, B, Biological Sciences*, *353*, 1841–1849.
Long, R., Sebo, J., Butlin, P., Finlinson, K., Fish, K., Harding, J., Pfau, J., Sims, T., Birch, J., & Chalmers, D. (2024). Taking AI welfare seriously. https://arxiv.org/abs/2411.00986
Mashour, G. A., Roelfsema, P., Changeux, J. P., & Dehaene, S. (2020). Conscious processing and the global neuronal workspace hypothesis. *Neuron*, *105*(5), 776–798. https://doi.org/10.1016/j.neuron.2020.01.026
Maturana, H., & Varela, F. (1980). *Autopoiesis and cognition: The realization of the living*. D. Reidel.
McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, *5*, 115–133 (1943). https://doi.org/10.1007/BF02478259
McFadden, J. (2020). Integrating information in the brain's EM field: The cemi field theory of consciousness. *Neuroscience of Consciousness*, *2020*(1), niaa016. https://doi.org/10.1093/nc/niaa016
McFadden, J., & Al-Khalili, J. (2018). The origins of quantum biology. *Proceedings Mathematical, Physical, and Engineering Sciences*, *474*(2220), 20180674. https://doi.org/10.1098/rspa.2018.0674
McMillen, P., & Levin, M. (2024). Collective intelligence: A unifying concept for integrating biology across scales and substrates. *Communications Biology*, *7*, 378. https://doi.org/10.1038/s42003-024-06037-4
Mead, C. (2020). How we created neuromorphic engineering. *Nature Electronics*, *3*, 434–435.
Merolla, P. A., Arthur, J. V., Alvarez-Icaza, R., Cassidy, A. S., Sawada, J., Akopyan, F., Jackson, B. L., Imam, N., Guo, C., Nakamura, Y., Brezzo, B., Vo, I., Esser, S. K., Appuswamy, R., Taba, B., Amir, A., Flickner, M. D., Risk, W. P., Manohar, R., & Modha, D. S. (2014). A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science*, *345*(6197), 668–673. https://doi.org/10.1126/science.1254642
Metzinger, T. (2021). Artificial suffering: An argument for a global moratorium on synthetic phenomenology. *Journal of Artificial Intelligence and Consciousness*, *8*(1), 1024. https://doi.org/10.1142/S270507852150003X
Metzinger, T. (2024). *The elephant and the blind: The experience of pure consciousness: philosophy, science, and 500+ experiential reports*. MIT Press.
Mitchell, K. J. (2023). *Free agents: How evolution gave us free will*. Princeton University Press.
Mitchell, M. (2024). The metaphors of artificial intelligence. *Science*, *386*(6723), eadt6140. https://doi.org/10.1126/science.adt6140

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(13), e2215907120. https://doi.org/10.1073/pnas.2215907120

Mohseni, M., Omar, Y., & Engel, G. S. (Eds.). (2013). *Quantum effects in biology*. Cambridge University Press.

Morales Pantoja, I. E., Smirnova, L., Muotri, A. R., Wahlin, K. J., Kahn, J., Boyd, J. L., Gracias, D. H., Harris, T. D., Cohen-Karni, T., Caffo, B. S., Szalay, A. S., Han, F., Zack, D. J., Etienne-Cummings, R., Akwaboah, A., Romero, J. C., Alam El Din, D. M., Plotkin, J. D., Paulhamus, B. L., . . . & Hartung, T. (2023). First Organoid Intelligence (OI) workshop to form an OI community. *Frontiers in Artificial Intelligence*, *6*, 1116870. https://doi.org/10.3389/frai.2023.1116870

Mudrik, L., Boly, M., Dehaene, S., Fleming, S. M., Lamme, V., Seth, A. K., & Melloni, L. (2025). Unpacking the complexities of consciousness: Theories and reflections. *Neuroscience & Biobehavioral Reviews*, *170*, 106053. https://doi.org/10.1016/j.neubiorev.2025.106053

Nagel, T. (1974). What is it like to be a bat? *Philosophical Review*, *83*, 435–450.

Nave, K. (2025). *A drive to survive: The free energy principle and the meaning of life*. MIT Press.

Neven, H., Zalcman, A., Read, P. K. K., van der Molen, T., Bouwmeester, D., Bodnia, E., Turin, L., & Koch, C. (2024). Testing the conjecture that quantum processes create conscious experience. *Entropy*, *26*(6), 460.

O'Regan, J. K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, *24*(5), 939–973; discussion 973-1031. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=12239892

Ororbia, A., & Friston, K. J. (2023). Mortal computation: a foundation for biomimetic intelligence. https://arxiv.org/abs/2311.09589

Parr, T., Pezzulo, G., & Friston, K. J. (2022). *Active inference: The free energy principle in mind, brain, and behavior*. MIT Press.

Penrose, R. (1989). *The Emperor's new mind*. Oxford University Press.

Piccinini, G. (2018). Non-computational functionalism: Computation and the function of consciousness. Conference of Cognitive Computational Neuroscience,

Piccinini, G. (2020). *Neurocognitive mechanisms*. Oxford University Press.

Piccinini, G. (2023). *The computational theory of mind*. Cambridge University Press.

Polger, T., & Shapiro, L. (2016). *The multiple realization book*. Oxford University Press.

Prinz, J. (2012). *The conscious brain: How attention engenders experience*. Oxford University Press.

Putnam, H. (1967). Psychological predicates. In W. H. Capitan & D. D. Merrill (Eds.), *Art, mind, and religion* (pp. 37–48). Pittsburgh University Press.

Putnam, H. (1975). *Mind, language, and reality*. Cambridge University Press.

Putnam, H. (1988). *Representation and reality*. MIT Press.

Pylyshyn, Z. (1980). The 'causal power' of machines. *Behavioral and Brain Sciences*, *3*, 442–444.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87. https://doi.org/10.1038/4580

Rosas, F. E., Geiger, B. C., Luppi, A. I., Seth, A. K., Polani, D., Gastpar, M., & Mediano, P. A. M. (2024). Software in the natural world: A computational approach to hierarchical emergence. https://arxiv.org/abs/2402.09090

Russell, S., & Norvig, P. (2003). *Artificial intelligence: A modern approach*. Prentice Hall.

Schneider, S. (2019). *Artificial you: AI and the future of your mind*. Princeton University Press.

Schrödinger, E. (1944). *What is life?*. Cambridge University Press.

Schuman, C. D., Kulkarni, S. R., Parsa, M., Mitchell, J. P., Date, P., & Kay, B. (2022). Opportunities for neuromorphic computing algorithms and applications. *Nature Computational Science*, *2*(1), 10–19. https://doi.org/10.1038/s43588-021-00184-y

Schwitzgebel, E., & Garza, M. (2023). Designing AI with rights, consciousness, self-respect, and freedom. In F. Lara & J. Deckers (Eds.), *Ethics of artificial intelligence*. (pp. 459–479). Springer Nature.

Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–457.

Searle, J. (1990). Is the brain a digital computer?. *Proceedings and Addresses of the American Philosophical Association*, *64*(3), 21–37.

Searle, J. (2007). Dualism revisited. *Journal of Physiology - Paris*, *101*, 169–178.

Searle, J. (2017). Biological naturalism. In S. Schneider & M. Velmans (Eds.), *The Blackwell companion to consciousness* (pp. 327–336). John Wiley and Sons Ltd.

Sengupta, B., Stemmler, M. B., & Friston, K. J. (2013). Information and efficiency in the nervous system–a synthesis. *PLOS Computational Biology*, *9*(7), e1003157. https://doi.org/10.1371/journal.pcbi.1003157

Seth, A. K. (2009). The strength of weak artificial consciousness. *Journal of Machine Consciousness*, *1*(1), 71–82.

Seth, A. K. (2012). Putting Descartes before the horse: Quantum theories of consciousness: Comment on "Consciousness, biology, and quantum hypotheses" by Baars & Edelman [Comment]. *Physics of life reviews*, *9*(3), 297–298; discussion 306-307. https://doi.org/10.1016/j.plrev.2012.07.005

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends Cogn Sci*, *17*(11), 565–573. https://doi.org/10.1016/j.tics.2013.09.007

Seth, A. K. (2015). The cybernetic bayesian brain: From interoceptive inference to sensorimotor contingencies. In J. M. Windt & T. Metzinger (Eds.), *Open MIND* (pp. 35(T)). MIND Group. https://doi.org/10.15502/9783958570108

Seth, A. K. (2016). The real problem. *Aeon*. https://aeon.co/essays/the-hard-problem-of-consciousness-is-a-distraction-from-the-real-one

Seth, A. K. (2019). Being a beast machine: The origins of selfhood in control-oriented interoceptive inference. In M. Colombo, L. Irvine, & M. Stapleton (Eds.), *Andy Clark and his critics* (pp. 238–254). Wiley-Blackwell.

Seth, A. K. (2021). *Being you: A new science of consciousness*. Faber & Faber.

Seth, A. K. (2023). Why conscious AI is a bad, bad idea. *Nautilus*. https://nautil.us/why-conscious-ai-is-a-bad-bad-idea-302937/

Seth, A. K., & Bayne, T. (2022). Theories of consciousness. *Nature Reviews Neuroscience*, *23*(7), 439–452. https://doi.org/10.1038/s41583-022-00587-4

Seth, A. K., & Friston, K. J. (2016). Active interoceptive inference and the emotional brain. *Philosophical Transactions of the Royal Society B-Biological Sciences*, *371*(1708). 20160007. https://doi.org/ARTN10.1098/rstb.2016.0007

Seth, A. K., & Tsakiris, M. (2018). Being a beast machine: The somatic basis of selfhood. *Trends in Cognitive Sciences*, *22*(11), 969–981. https://doi.org/10.1016/j.tics.2018.08.008

Shagrir, O. (2010). The rise and fall of computational functionalism. In Y. Ben-Menahem (Eds.), *Hilary Putnam* (pp. 220–250). Cambridge University Press.

Shanahan, M. (2016). Conscious exotica. *Aeon*. https://aeon.co/essays/beyond-humans-what-other-kinds-of-minds-might-be-out-there

Shanahan, M. (2024). Simulacra as conscious exotica. *Inquiry*, 1–29. https://doi.org/https://www.tandfonline.com/doi/full/10.1080/0020174X.2024.2434860

Shanahan, M., McDonell, K., & Reynolds, L. (2023). Role play with large language models. *Nature*, *623*(7987), 493–498. https://doi.org/10.1038/s41586-023-06647-8

Shanahan, M., & Singler, B. (2024). Existential conversations with large language models: content, community, and culture. *ArXiv*, 2411.13223 https://doi.org/10.48550/arXiv.2411.13223

Shannon, C., & Weaver, W. (1949). *The mathematical theory of communication*. The University of Illinois Press.

Shapiro, L. (2008). How to test for multiple realization. *Philosophy of Science*, *75*, 514–525.

Shiller, D. (2024). Functionalism, integrity, and digital consciousness. *Synthese*, *203*, 47.

Shugg, W. (1968). The cartesian beast-machine in english literature (1663-1750). *Journal of the History of ideas*, *29*(2), 279–292. http://www.jstor.org/stable/2708581

Simon, H. (1988). *The sciences of the artificial* (3rd ed.). MIT Press.

Simons, D. J., & Levin, D. T. (1997). Change blindness. *Trends in Cognitive Sciences*, *1*(7), 261–267. https://doi.org/10.1016/S1364-6613(97)01080-2

Singer, P. (1975). *Animal liberation: A new ethics for our treatment of animals*. Harper Collins.

Solms, M. (2021). *The hidden spring: A journey to the source of consciousness*. Profile Books.

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

Stade, E. C., Stirman, S. W., Ungar, L. H., Boland, C. L., Schwartz, H. A., Yaden, D. B., Sedoc, J., DeRubeis, R. J., Willer, R., & Eichstaedt, J. C. (2024). Large language models could change the future of behavioral healthcare: A proposal for responsible development and evaluation. *Npj Mental Health Research*, *3*(1), 12. https://doi.org/10.1038/s44184-024-00056-z

Sterling, P. (2012). Allostasis: a model of predictive regulation. *Physiology & Behavior*, *106*(1), 5–15. https://doi.org/10.1016/j.physbeh.2011.06.004

Stewart, J., Gapenne, O., & Di Paolo, E. A. (2010). *Enaction: Toward a new paradigm for cognitive science*. MIT Press.

Thompson, A. (1998). On the automatic design of robust electronics through artificial evolution. In M. Sipper, D. Mange, & A. Perez-Uribe (Eds.), *Evolvable systems: From biology to hardware* (pp. 13–24). Springer.

Thompson, E. (2007). *Mind in life: Biology, phenomenology, and the sciences of mind*. Harvard University Press.

Thompson, E. (2022). Could all life be sentient?. *Journal of Consciousness Studies*, *29*, 229–265.

Thompson, E., & Varela, F. J. (2001). Radical embodiment: neural dynamics and consciousness. *Trends in Cognitive Sciences*, *5*(10), 418–425. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=11707380

Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. *Nature Reviews Neuroscience*, *17*(7), 450–461. https://doi.org/10.1038/nrn.2016.44

Tononi, G., & Edelman, G. M. (1998). Consciousness and complexity. *Science*, *282*(5395), 1846–1851. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=9836628

Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, *42*, 230–265.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*, 433–460. https://doi.org/10.1093/mind/LIX.236.433

Turkle, S. (2021). *The empathy diaries: a memoir*. Penguin.

Ulmann, B. (2022). *Analog computing* (2nd ed.). De Gruyter Oldenbourg.

Van Gelder, T. (1995). What might cognition be if not computation?. *Journal of Philosophy*, *92*(7), 345–381.

VanRullen, R., & Kanai, R. (2021). Deep learning and the Global Workspace Theory. *Trends in Neurosciences*, *44*(9), 692–704. https://doi.org/10.1016/j.tins.2021.04.005

Varela, F. J. (1996). Neurophenomenology: A methodological remedy for the hard problem. *Journal of Consciousness Studies*, *3*, 330–350.

Varela, F. J., Thompson, E., & Rosch, E. (1993). *The embodied mind: Cognitive science and human experience*. MIT Press.

Véliz, C. (2021). *Privacy is power*. Penguin.

Vong, W. K., Wang, W., Orhan, A. E., & Lake, B. M. (2024). Grounded language acquisition through the eyes and ears of a single child. *Science*, *383*(6682), 504–511. https://doi.org/10.1126/science.adi1374

Wagenaar, W. A., & Sagaria, S. D. (1975). Misperception of exponential growth. *Percept Psychophys*, *18*, 416–422.

Whyte, C. J., Corcoran, A. W., Robinson, J., Smith, R., Friston, K. J., Seth, A. K., & Hohwy, J. (2024). To see is to look: The minimal theory of consciousness implicit in active inference. *ArXiv*. https://arxiv.org/abs/2410.06633

Wiese, W. (2024). Artificial consciousness: A perspective from the free energy principle. *Philosophical Studies*, *181*, 1947–1970. https://doi.org/10.1007/s11098-024-02182-y

Wimsatt, W. C. (1986). Developmental constraints, generative entrenchment, and the innate-acquired distinction. In W. Bechtel (Ed.), *Integrating scientific disciplines, science and philoso- phy* (pp. 185–208). Springer.

Open Peer Commentary

Too soon to say what consciousness is

Colin Allen*

Department of Philosophy, University of California, Santa Barbara, USA
colinallen@ucsb.edu
https://colinallen.dnsalias.org/

*Corresponding author.

doi:10.1017/S0140525X25103877, e316

Abstract

No one is in a position to say what really matters for consciousness, or whether only living systems can instantiate the complexity needed for conscious feelings. Seth's attempt to connect life to consciousness through the mathematics of active inference and free energy minimization plays one hunch, but it is too soon to say whether it is the best hunch.

What is consciousness? Seth characterizes it in terms of feeling ("what it's like") rather than function ("what it does"). The conceptual distinction between feelings and doings has seemed clear to philosophers of mind, who have relied on it to generate many thought experiments. Seth mercifully pays little attention to such thought experiments, thereby signaling that he sets little store in them.

Despite their conceptual separability, feelings often do things (epiphenomenalism notwithstanding). Painful stimuli lead conscious organisms to avoid situations that cause them and to ingest pain-relievers when available. Dichotomous characterizations of consciousness in terms of feelings independently of functions can only be regarded as partial at best.

Many of the arguments others have given in support of the idea that conscious AI is imminent are diagnosed by Seth as presupposing computational functionalism about consciousness. Because this presupposition begs the question against Seth's preferred bionaturalist view, it gives him a reason to be unmoved by those arguments.

Previous versions of biological naturalism leave the connection between biology and consciousness mysterious. Seth attempts to rectify this by adopting the proposal that the free energy principle and predictive processing views of conscious mental contents are linked. This link goes via the shared mathematics underlying, on the one hand, the requirement that biological "autopoietic" systems minimize free energy to combat entropy, and on the other

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

hand, the minimization of perceptual prediction errors via active inference.

Seth faces the obvious objection that something approximating active inference can be (and has been) implemented computationally. Because active inference seems like a "doing," something also needs to be said about why it becomes a feeling when implemented biologically but not when given a computational-functional implementation. His suggestion is that computational implementation of active inference is just a crude facsimile of what brains do. The idea seems to be that the processes involved in active inference and the energetics involved in combating entropy are tightly coupled in living organisms, right down to the level of individual cells. Computers instead instantiate a kind of "eternal" computing that is not strongly coupled to any energetic or temporal demands.

The relative detachment of Turing-style computation from the battle against entropy may be characteristic of current computational implementations of active inference. But is this a principled difference between computational and biological systems, or just a difference of degree? Because Bayesian calculations are computationally intractable, they can only be approximated in both brains and computers. The dilemma for Seth's version of biological naturalism is that either consciousness is at least partly due to (tractably) approximating a computable function, or it isn't. If it isn't, then the free energy principle doesn't explain how to get from approximating active inference to feelings. If it is, then it needs to be explained why some Bayesian approximators (biological organisms) are conscious but some (computers) aren't. Without such an account, the FEP-bionaturalist equally begs the question against the computational functionalist.

Perhaps two sides begging questions against each other reflects both normal metaphysics and pre-paradigmatic science. Any attempt to resolve questions about what something "really" is requires interpretation of evidence through the lens of some theory or other, and how things are interpreted from one theoretical perspective will beg questions against the others. There are no theory-independent standpoints on the evidence, although there may be relatively theory-neutral descriptions of the evidence (such as that physicists can agree on the radiation measured coming from a black hole without agreeing on a theory of black holes).

But where does this leave the science of consciousness? Without a consensus theory, or even an agreed upon characterization of consciousness, the best we can do collectively is to encourage computationalists and biological naturalists – indeed, anyone who agrees on some basic scientific precepts about holding their theories accountable to evidence – to continue refining experiments and measurements to probe the putative cases of consciousness that confront us: human, animal, or (most controversially) machine.

As Seth humbly admits, no one is in a position to say what really matters for consciousness or how fine-grained the analysis of the underlying mechanisms needs to be. Hence, we are not in a position to rule any theory off the table so far, either a priori or on the basis of the limited evidence that we have. This makes it okay and perhaps inevitable for different scientists to play their own hunches: different predilections will lean toward different attempts to unravel the mystery of how and why some systems are conscious.

What would not be okay, as Seth surely agrees, is to force hegemony at this early stage of investigation. The horizon for the development of field-defining theories is long – typically on the order of a century or more – and it is plain that the scientific study of consciousness is nowhere near such consensus. However, even incorrect theories can yield good techniques and measurements. The current pluralism among consciousness researchers is thus to be encouraged.

With this target article, Seth has done a great service to the field, making an excellent case against taking computational functionalism for granted, and doing a similarly excellent job of explaining the lines of thought in favor of taking the complexity of biological systems more seriously than computationalists typically do. I agree with his hunch that the degree of complexity found in brains matters for both intelligence and consciousness, but better measurements and better theories will need to be invented before anyone can say which details matter and whether implementing such complexity requires more than Turing computation.

Acknowledgments. Although she might deny her association with this piece, I thank Adina Roskies for her comments on earlier drafts.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interests. None.

Separating the conscious wheat from the unconscious chaff: Towards a biologically grounded computationalism

Matthew E. Larkum[a*], Jaan Aru[b], Christopher J. Whyte[c] and James M. Shine[c]

[a]Institute for Biology, Humboldt University of Berlin, Berlin, Germany; [b]Institute of Computer Science, University of Tartu, Tartu, Estonia and [c]Discipline of Neuroscience, Faculty of Medicine and Health, The University of Sydney, Camperdown, Sydney, NSW, Australia
matthew.larkum@hu-berlin.de
jaan.aru@gmail.com
christopherjackwhyte@gmail.com
mac.shine@sydney.edu.au
https://nail.cs.ut.ee/
https://shine-lab.org
https://www.christopherjwhyte.com
https://larkumlab.com/

*Corresponding author.

doi:10.1017/S0140525X25103725, e317

Abstract

We agree that consciousness, as we know it, is inextricably biological – but not every biological detail is essential. Identifying which features are necessary, across scales, is vital for assessing the limits of computationalism. This process will reveal a richer, more biologically informed computationalism – one that future machines could, in principle, instantiate, but only after much deeper biological understanding.

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

As Seth argues, the only example of consciousness that we can confidently confirm as real is the byproduct of billions of years of natural selection – that is, consciousness as we know it is inherently biological. In his provocative commentary, Seth argues for biological naturalism, which is the view that the subjective, phenomenal aspect of consciousness is not separable from the biological machinery that keeps an organism alive. This makes life itself a necessary condition for consciousness. Seth presents this position as his preferred view rather than a definitive conclusion, and we share his caution. Others have argued that life itself is computational (Agüera y Arcas et al., 2024; Kurian, 2025), so we would argue there is further work to do to show that consciousness is not inherently computational, even granting Seth's premise that consciousness and life are intrinsically linked.

Specifically, it is unlikely that all aspects of biology are crucial for consciousness. So then, the central questions from our point of view become: which exact details of biology matter for consciousness, and can these specific details be computationally implemented (or not)? In simple terms: while every instance of consciousness that we can currently confirm with confidence is inherently biological in nature, this does not mean that every last detail of biology is required in order for a system to be conscious. Indeed, some aspects of biology seem to be important, e.g., the reticular activating system, higher-order thalamic nuclei, and pyramidal neurones oriented orthogonally to the layers of the cerebral cortex (Larkum et al., 2018), whilst others do not (e.g., Purkinje neurones in the cerebellum or the dura mater). That is, they may be simply part of the substrate that is "along for the conscious ride" (i.e., conscious "spandrels"). Thus, work separating the conscious wheat from the unconscious chaff is needed before we can draw conclusions about the limits of computationalism. If we can separate the crucial substrates from the supporting architecture, it may be possible (in principle) to implement the crucial features of (neuro)biology on a non-biological substrate. At this stage, our best guess is that this will require far more biological detail than we currently know how to build. Not only would we need to include L5-ET cells and their complex dendritic tree (Aru, Suzuki, & Larkum, 2020) which we have argued may provide a cellular mechanism for cortical associations (Larkum, 2013), we'd also need to integrate them with a working thalamus (Fang et al., 2025; Whyte et al., 2025) and an ascending arousal system (Munn et al., 2023). From there, we have argued through the lens of neuroscience that a disembodied cortico-thalamo-reticular system would not be conscious of anything much at all (Aru, Larkum, & Shine, 2023) – for the system's consciousness to be as rich as ours, the system must be integrated into a body that can provide an interface with the complex, multi scale nature of the surrounding world.

Of the biological features that seem to play an essential role in brain function, and potentially consciousness, scale-free organisation is likely the most difficult to instantiate in a substrate-independent manner. At rest, brain activity exhibits scale-free organisation, meaning its dynamics span multiple temporal and spatial scales without a single dominant scale. Depending on task demands, this baseline organisation can be selectively biased to emphasise certain scales of activity (Munn et al., 2024). Crucially, such scale-free behaviour causes problems for computational theories which privilege a particular scale (e.g., population coding; Averbeck, Latham, & Pouget, 2006), challenging the idea that we can draw a clear distinction between hardware and software as the computations being performed may require contributions from distinct or (multiple) scales. It may not be possible, therefore, to have human-readable software that can be implemented on any substrate if the (scale-free) substrate is itself essential for the computations being performed. For us, this also means that consciousness cannot be understood at some higher, substrate-independent level of description (Gidon, Aru, & Larkum, 2022, 2025).

Thus, while we support Seth's line of reasoning (Aru et al., 2023), we would not announce the end of computationalism yet. Instead, we would argue that computationalism comes in degrees. There is naive computationalism, which states that present-day understanding of the computations underlying consciousness might be enough to build conscious machines (Butlin et al., 2023). However, computationalism could also, in principle, be much more complex, taking into account details about the brain – we prefer the term "biological computationalism," in order to separate it from the naive view. For instance, one could in principle build a computational theory that incorporates the fact that the brain's organisation and dynamics are fundamentally scale-free, while also allowing for task-dependent modulations that temporarily weight certain scales more heavily (Munn et al., 2024). Consciousness might be a phenomenon that only exists when a complex, multiscale nervous system is embedded within a living organism that is itself embedded within the natural world, which is itself replete with a dizzying amount of multiscale structure. By incorporating these key relational properties, a future theory of biological computationalism could, in principle, capture the essential ingredients necessary for consciousness.

The way forward is theoretically simple, yet painstaking in its execution. For one, we need to devise experiments across multiple scales and then build causal links between the different scales. And at all stages, we will have to carefully evaluate whether a constituent part is necessary for consciousness or just a byproduct or background condition. *In silico* models will be extremely helpful in this regard, as they can act as a crucial, bidirectional bridge connecting biologically based theories of consciousness with the crucial empirical tests (Whyte et al., 2025).

Eventually, we hope that Seth's article brings a change to consciousness science. Due to the rise and popularity of neurobiologically underspecified theories, biology has been in the background for too long. Let's not pretend that understanding these specific processes that underlie consciousness should be simple. Seth's article shows that we still have a long way to go and much exciting science ahead of us.

Financial support. JA is supported by Estonian Research Council grants PSG728, Tem-TA 120 and the Estonian Centre of Excellence in Artificial Intelligence (EXAI) funded by the Estonian Ministry of Education and Research. ML is supported by the European Research Council (101055340) and the German Research Foundation grants SFB1315/327654276, SFB/TRC384, LA 3442/13-1, LA 3442/16-1, NeuroCure III EXC 256. JMS is supported by a National Health and Medical Research Council Fellowship (GNT1193857), a Viertel/Bellberry Fellowship and the Australian Research Council (DP240101295).

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

Competing interests. The authors declare no conflict of interest.

References

Agüera y Arcas, B., Alakuijala, J., Evans, J., Laurie, B., Mordvintsev, A., Niklasson, E., … & Versari, L. (2024). Computational life: How well-formed, self-replicating programs emerge from simple interaction. arXiv e-prints, arXiv-2406.

Aru, J., Larkum, M. E., & Shine, J. M. (2023). The feasibility of artificial consciousness through the lens of neuroscience. *Trends in neurosciences*, *46*(12), 1008–1017.

Aru, J., Suzuki, M., & Larkum, M. E. (2020). Cellular mechanisms of conscious processing. *Trends in Cognitive Sciences*, *24*(10), 814–825.

Averbeck, B. B., Latham, P. E., & Pouget, A. (2006). Neural correlations, population coding and computation. *Nature Reviews Neuroscience*, *7*(5), 358–366.

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., … & VanRullen, R. (2023). Consciousness in artificial intelligence: Insights from the science of consciousness. arXiv preprint *arXiv:2308.08708*.

Fang, Z., Dang, Y., Ping, A. A., Wang, C., Zhao, Q., Zhao, H., … & Zhang, M. (2025). Human high-order thalamic nuclei gate conscious perception through the thalamofrontal loop. *Science*, *388*(6742), eadr3675.

Gidon, A., Aru, J., & Larkum, M. E. (2022). Does brain activity cause consciousness? A thought experiment. *PLoS Biology*, *20*(6), e3001651.

Gidon, A., Aru, J., & Larkum, M. E. (2025). Does neural computation feel like something?. *Frontiers in Neuroscience*, *19*, 1511972.

Kurian, P. (2025). Computational capacity of life in relation to the universe. *Science Advances*, *11*(13), eadt4623.

Larkum, M. (2013). A cellular mechanism for cortical associations: An organizing principle for the cerebral cortex. *Trends in Neurosciences*, *36*(3), 141–151.

Larkum, M. E., Petro, L. S., Sachdev, R. N. S., & Muckli, L. (2018). A perspective on cortical layering and layer-spanning neuronal elements. *Frontiers in Neuroanatomy*, *12*, 56.

Munn, B. R., Müller, E. J., Medel, V., Naismith, S. L., Lizier, J. T., Sanders, R. D. and Shine, J. M. (2023). Neuronal connected burst cascades bridge macroscale adaptive signatures across arousal states. *Nature Communications*, *16*, 6846.

Munn, B. R., Müller, E. J., Favre-Bulle, I., Scott, E., Lizier, J. T., Breakspear, M. and Shine, J. M. (2024). Multiscale organization of neuronal activity unifies scale-dependent theories of brain function. *Cell*, *187*(25), 7303–7313. 16, 6846.

Whyte, C. J., Müller, E. J., Aru, J., Larkum, M., John, Y., Munn, B. R., & Shine, J. M. (2025). A burst-dependent thalamocortical substrate for perceptual awareness. *PLOS Computational Biology*, *21*(4), e1012951.

Steam-engine naturalism

Manuel Baltieri[a,b]* and Ryota Kanai[a]

[a]Araya Inc., Tokyo, Japan and [b]Department of Informatics, University of Sussex, Brighton, UK
manuel_baltieri@araya.org
kanair@araya.org

*Corresponding author.

doi:10.1017/S0140525X25103531, e318

Abstract

Seth's biological naturalism builds on ideas from predictive processing, the free energy principle, and active inference to (1) argue that only (some) biological systems are in the consciousness business and to (2) refute the possibility of artificial consciousness, i.e., consciousness in AI and other artificial systems. We find this perspective problematic for a few different reasons.

The starting point for Seth's biological naturalism is *predictive processing*. Predictive processing describes various computational and cognitive processes in the brain that can be modelled in terms of cycles of predictions and error correction at different spatio-temporal and hierarchical scales (Clark, 2013; Hohwy, 2013). However, it appears that not much is inherently biological about it: while it can be applied to describe certain biological processes, the same appears to be true for several non-biological ones too. It has been shown, in fact, that the core ideas behind predictive processing come from standard signal processing theory (Spratling, 2016) and estimation theory, e.g., Kalman filtering (Rao, 1999) (see however Baltieri & Isomura, 2021), and are simply applied to describe neural and behavioural processes of different kinds. While it is hard to overstate the importance of mathematical models of brain processes, it is perhaps even harder to argue that predictive processing can explain the differences between a steam engine (Baltieri, Buckley, & Bruineberg & Bruineberg, 2020) or a Braitenberg vehicle (Baltieri & Buckley, 2017) and a human brain. One could argue that Braitenberg vehicles (Shaikh & Rañó, 2020) and Watt governors (Van Gelder, 1995) have often been used to support the idea of simpler models of biological and cognitive systems and their behaviours. However, using this to vindicate the proposal of biological naturalism would be a category mistake, similar to thinking of Markov blankets as describing physical boundaries in biological systems (Bruineberg et al., 2022a, 2022b).

A second motivating principle behind biological naturalism is the *free energy principle* (Friston, 2010; Friston et al., 2023). According to the free energy principle, for a "thing" to exist, it must appear as if it is minimising a variational free energy functional to perform approximate Bayesian inference on the states of the external environment that produce its sensory inputs. This principle however doesn't come without problems (Aguilera et al., 2022; Biehl, Pollock, & Kanai & Kanai, 2021; Bruineberg et al., 2022a, 2022b; Di Paolo, Thompson, & Beer & Beer, 2022; Raja et al., 2021), and it is unclear once again what would make it relevant for biological naturalism. Over the years, the free energy principle has been proposed as a theory of brains first (Friston, 2010), of life and its origins next (Friston, 2013), and finally of "things" in more general physical terms, consistent with accounts from quantum, statistical, and classical mechanics (Friston, 2019; Friston et al., 2023). It is thus surprising to see this proposed as useful to formalise biological naturalism: if the free energy principle is in fact correct, as a (re)description of some key aspects of the physical world, it can be used "to simulate and predict the sentient behaviour of a particle, person, artefact or agent (i.e., some 'thing')" (Friston et al., 2023). Thus, if we were to take the free energy principle at face value, biological systems ought to be consistent with it much like a "particle" or an "artefact" ought to. Seth further argues that the free energy principle supports an explanation of autopoiesis, thought to be a central aspect of theories about the emergence of biological individuality and normativity (Maturana & Varela, 1980). It appears, however, that this principle does not in fact explain, model, or even account for autopoiesis in biological systems, conflating levels of analysis and definitions of "autopoiesis" (implying self-creation and a clear description of structure and organisation among other things) and "homeostasis" (Bruineberg et al., 2022b; Di Paolo, Thompson, &

Beer & Beer, 2022; Nave, 2025; Raja et al., 2021; Suzuki, Miyahara, & Miyazono & Miyazono, 2022).

Finally, Seth argues that *active inference* (Friston et al., 2017; Parr, Pezzulo, & Friston & Friston, 2022) seen from a control theoretic/cybernetic perspective well synergises with accounts of interoception, the sense of the body "from within." However, the cybernetic perspective proposed by Seth is hardly exclusive to biological systems, see the steam engine/Watt governor example already discussed above (Baltieri, Buckley, & Bruineberg & Bruineberg, 2020). Formally, this simply appears as a restatement of classical works on "good regulators" and its formalisation in the "internal model principle," which cover classes of systems that can be described as goal-pursuing by an external observer and thus not only biological systems (Baltieri et al., 2025; Virgo et al., 2025). One of the leading examples of this internal model principle in particular is integral control (Wonham, 1976), a core part of modern control theory (Åström & Hägglund, 1995) with applications in biology (Yi et al., 2000) and neuroscience (Ritz et al., 2017), which can also be easily accounted for in active inference (Baltieri & Buckley, 2019), making the special role supposedly played by active inference for biological systems not clear.

Overall, nothing about predictive processing, the free energy principle, and active inference seems exclusive to biological systems. In light of this, we thus believe that for a coherent theory of consciousness, we are left with one of the following choices:

1. Pursue Seth's biological naturalism, but with different tools. In other words, build an argument for biological naturalism using something other than predictive processing, the free energy principle, and active inference. This is because none of these seem to be in the business of characterising *exclusively* biological systems, much less *some* biological systems, the ones that ought to display the necessary conditions for consciousness in this proposal of biological naturalism.
2. Reject Seth's biological naturalism. Specifically, abandon the idea that there's something inherently special about consciousness in biological organisms, fully embracing *universality* (Kanai & Fujisawa, 2024) beyond biology, and frameworks that don't a priori exclude the possibility of consciousness in artificial systems, like the free energy principle and active inference.
3. Accept parts of Seth's biological naturalism, but aim for *biological universality*. This means settling for a notion of universality that applies only to biological systems. The aim would be to make biological naturalism "as universal as possible" by providing formal, mathematical criteria for what should be understood as "biological," e.g., what counts as a "biological neuron" if we assume that these units are somehow important for consciousness, or what counts as "being alive" if this ought to be the "minimal 'ground state' of conscious experience."

Acknowledgements. None.

Financial support. M.B. and R.K. were supported by JST, Moonshot R&D, Grant Number JPMJMS2012.

Competing interests. The authors declare no conflicts of interest.

References

Aguilera, M., Millidge, B., Tschantz, A., & Buckley, C. L. (2022). How particular is the physics of the free energy principle?. *Physics of Life Reviews*, *40*, 24–50. https://doi.org/10.1016/j.plrev.2021.11.001

Åström, K. J., & Hägglund, T. (1995). *PID controllers: Theory, design, and tuning* (2. ed). Instrument Society of America.

Baltieri, M., Biehl, M., Capucci, M., & Virgo, N. (2025). *A Bayesian Interpretation of the Internal Model Principle* (No. arXiv:2503.00511). arXiv. https://doi.org/10.48550/arXiv.2503.00511

Baltieri, M., & Buckley, C. L. (2017). An active inference implementation of phototaxis. *European Conference on Artificial Life 2017*, 36–43. https://direct.mit.edu/isal/proceedings-abstract/ecal2017/29/36/99577

Baltieri, M., & Buckley, C. L. (2019). PID control as a process of active inference with linear generative models. *Entropy*, *21*(3), 257.

Baltieri, M., Buckley, C. L., & Bruineberg, J. (2020). Predictions in the eye of the beholder: An active inference account of Watt governors. *The 2020 Conference on Artificial Life*, 121–129. https://doi.org/10.1162/isal_a_00288

Baltieri, M., & Isomura, T. (2021). *Kalman filters as the steady-state solution of gradient descent on variational free energy* (No. arXiv:2111.10530). arXiv. http://arxiv.org/abs/2111.10530

Biehl, M., Pollock, F. A., & Kanai, R. (2021). A technical critique of some parts of the free energy principle. *Entropy*, *23*(3), 293. https://doi.org/10.3390/e23030293

Bruineberg, J., Dołęga, K., Dewhurst, J., & Baltieri, M. (2022a). The Emperor's New Markov Blankets. *Behavioral and Brain Sciences*, *45*, e183.

Bruineberg, J., Dołęga, K., Dewhurst, J., & Baltieri, M. (2022b). The Emperor is Naked: Replies to commentaries on the target article. *Behavioral and Brain Sciences*, 45.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204.

Di Paolo, E., Thompson, E., & Beer, R. D. (2022). Laying down a forking path: Tensions between enaction and the free energy principle. *Philosophy and the Mind Sciences*, 3.

Friston, K. J. (2010). The free-energy principle: A unified brain theory?. *Nature Reviews Neuroscience*, *11*(2), 127–138.

Friston, K. J. (2013). Life as we know it. *Journal of the Royal Society Interface*, *10*(86), 20130475.

Friston, K. J. (2019). *A free energy principle for a particular physics* (No. arXiv:1906.10184). arXiv. https://doi.org/10.48550/arXiv.1906.10184

Friston, K. J., Da Costa, L., Sajid, N., Heins, C., Ueltzhöffer, K., Pavliotis, G. A., & Parr, T. (2023). The free energy principle made simpler but not too simple. *Physics Reports*, *1024*, 1–29. https://doi.org/10.1016/j.physrep.2023.07.001

Friston, K. J., Fitzgerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active Inference: A Process Theory. *Neural Computation*, *29*, 1–49. https://doi.org/10.1162/NECO_a_00912

Hohwy, J. (2013). *The predictive mind*. Oxford University Press.

Kanai, R., & Fujisawa, I. (2024). Toward a universal theory of consciousness. *Neuroscience of Consciousness*, *2024*(1), niae022. https://doi.org/10.1093/nc/niae022

Maturana, H. R., & Varela, F. J. (1980). *Autopoiesis and cognition: The realization of the living*. Springer Science & Business Media.

Nave, K. (2025). *A drive to survive: The free energy principle and the meaning of life*. MIT Press.

Parr, T., Pezzulo, G., & Friston, K. J. (2022). *Active inference: The free energy principle in mind, brain, and behavior*. MIT Press.

Raja, V., Valluri, D., Baggs, E., Chemero, A., & Anderson, M. L. (2021). The Markov blanket trick: On the scope of the free energy principle and active inference. *Physics of Life Reviews*, *39*, 49–72. https://doi.org/10.1016/j.plrev.2021.09.001

Rao, R. P. N. (1999). An optimal estimation approach to visual perception and learning. *Vision Research*, *39*, 1963–1989.

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

Ritz, H., Nassar, M. R., Frank, M. J., & Shenhav, A. (2017). A control theoretic model of adaptive learning in dynamic environments. *Journal of Cognitive Neuroscience*, *30*, 1405–1421. https://doi.org/10.1162/jocn_a_01289

Shaikh, D., & Rañó, I. (2020). Braitenberg vehicles as computational tools for research in neuroscience. *Frontiers in Bioengineering and Biotechnology*, *8*. https://doi.org/10.3389/fbioe.2020.565963

Spratling, M. W. (2016). A review of predictive coding algorithms. *Brain and Cognition*, *112*, 92–97. https://doi.org/10.1016/j.bandc.2015.11.003

Suzuki, K., Miyahara, K., & Miyazono, K. (2022). Who tailors the blanket?. *Behavioral and Brain Sciences*, *45*, e213. https://doi.org/10.1017/S0140525X22000206

Van Gelder, T. (1995). What might cognition be, if not computation?. *The Journal of Philosophy*, *92*(7), 345–381.

Virgo, N., Biehl, M., Baltieri, M., & Capucci, M. (2025). A "good regulator theorem" for embodied agents. (Under Review at ALIFE Conference).

Wonham, W. M. (1976). Towards an abstract internal model principle. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-6*(11), 735–740. https://doi.org/10.1109/TSMC.1976.4309444

Yi, T.-M., Huang, Y., Simon, M. I., & Doyle, J. (2000). Robust perfect adaptation in bacterial chemotaxis through integral feedback control. *Proceedings of the National Academy of Sciences*, *97*(9), 4649–4653. https://doi.org/10.1073/pnas.97.9.4649

Biases and the question of artificial consciousness

Tim Bayne[a,b,c*]

[a]School of Philosophy, History, and Indigenous Studies (SOPHIS), Monash University, Melbourne, Australia; [b]Monash Centre for Consciousness and Contemplative Studies (M3CS), Monash University, Melbourne, Australia and [c]Program for Brain, Mind, and Consciousness, Canadian Institute for Advanced Research (CIFAR), Toronto, ON, Canada
tim.bayne@gmail.com

*Corresponding author.

doi:10.1017/S0140525X25103634, e319

Abstract

Seth argues that anthropocentrism, anthropomorphism, and human exceptionalism bias us towards an overly optimistic view of the prospects of conscious AI. I agree, but argue that these phenomena might also bias us towards an overly pessimistic view of the prospects for artificial consciousness. Furthermore, it is not always clear when these factors are functioning as "biases."

The question of conscious AI is profoundly difficult. The primary problem lies not with our failure to understand AI, but with our failure to understand consciousness: we do not know how to assess claims of consciousness in systems that differ radically from us (Block 2002).

Given our ignorance, non-evidential factors – "biases" – are likely to play an important role in determining our attitudes. Seth identifies three such factors (anthropocentrism, anthropomorphism, and human exceptionalism), arguing that they combine to make the hypothesis of conscious AI more plausible than the evidence suggests. I agree that these factors are likely to shape our views of AI consciousness in important ways, but their influence is unlikely to be unidirectional, for they might also bias us to reject the possibility of conscious AI.

Seth suggests that anthropocentrism (the tendency to interpret the world in terms of human values and experiences) and human exceptionalism (the tendency to see ourselves as superior to other forms of life) combine to motivate the view that conscious AI is inevitable. "We humans know we're conscious . . . . We think we're intelligent . . . .and we feel we're special, so we assume consciousness and intelligence go together." This is indeed one possible inference. Equally, however, anthropocentrism and human exceptionalism might combine to motivate the view that conscious AI is impossible – or at least unlikely.

Consider how things stand at present. Although most of us recognize that humans don't have a monopoly on consciousness, we do tend to think of ourselves as occupying an exalted position in the "consciousness club." Other species have impressive sensory repertoires, but our conscious capacities (imagination, insight, creativity, and so on) are uniquely impressive. AI systems might outperform us in many ways, but they don't have any of what really matters: consciousness. "AI might take my job," we tell ourselves, "But at least I've got feelings!"

Ascribing consciousness to AI would upend this picture. Not only could we no longer view such systems as "mere tools," we might also find that our (self-appointed) position as the pre-eminent members of the consciousness club was under threat. After all, conscious AI systems are likely to be conscious in impressive ways, and to have many of the capacities that we take most pride in - rationality, insight, creativity, and the like. (In some sense, these capacities fall under what Seth calls "intelligence," but it would be a mistake to think that they don't also involve consciousness. Consciousness isn't just about feelings.)

What about anthropomorphism – the projection of humanlike properties onto other objects? Seth sees anthropomorphism at work in our tendency to attribute consciousness to Large Language Models (LLMs), and there is clearly some truth to this charge. In interacting with LLMs, we often feel as though we are engaging with another conscious agent. And there is evidence that at least some of us might be inclined to take this experience at face value, with recent studies finding that roughly 18% of Americans are willing to ascribe "sentience" (Anthis et al. 2025) or "subjective experience" (Dreksler et al. 2025) to AI systems.

At the same time, it's surely not unreasonable to think that consciousness is implicated in – and therefore arguably an indicator of – aspects of higher cognition and linguistic competence. In fact, Seth himself seems to embrace a version of this view, suggesting that the development of AI "may be enhanced by implementing some of the functions associated with consciousness." Crucially, we rely on linguistic and cognitive capacities (such as command-following) to probe for the presence of consciousness in various contexts (e.g., after brain injury or following sedation). I am not, of course, suggesting that command-following would be a good test for AI consciousness. The point, rather, is that given the centrality of linguistic and cognitive tests for consciousness in humans (and other animals), it's not unreasonable to regard these capacities as relevant to the AI question.

This leads us back to the question of what counts as a "bias." Intuitively, biases are "non-evidential factors," but what counts as evidence here? This issue is particularly challenging if one assumes – as Seth, I take it, does – that in looking for evidence (or markers) of consciousness, we have no choice but to start with ourselves (and other animals that we regard as having the capacity for consciousness) (Bayne et al. 2024). But given this starting point, some

kind of anthropocentrism would seem to be unavoidable. The question isn't *whether* to evaluate claims of artificial consciousness on the basis of commonalities between AI systems and ourselves (and thus "anthropocentrically"), but *which* commonalities matter.

This issue can be brought into sharp relief by considering Terry Bisson's story "They're made of meat," in which two intergalactic visitors discuss the possibility of human consciousness. "You're asking me to believe in sentient meat," says one. "I'm not asking you, I'm telling you," comes the reply. "Thinking meat! You're asking me to believe in thinking meat!" "Yes, thinking meat! Conscious meat! Loving meat. Dreaming meat." Are Bisson's imagined visitors guilty of the bias of chauvinism (anthropomorphism's twin) in doubting the possibility of human consciousness because we're made of meat and not silicon? If so, are we guilty of chauvinism in wondering whether silicon-based systems could be conscious? Perhaps. Equally, one could argue that both inferences are reasonable (albeit potentially flawed) implications of the fact that when faced with alien systems, one has no alternative but to extrapolate from oneself.

The upshot is that although anthropocentrism clearly can function as a "bias," it's arguable that any response to the challenge of artificial consciousness will involve a kind of anthropocentrism. The crucial challenge here is figuring out how to ground inferences about artificial consciousness on the basis of what we (think we) know about human consciousness. The notion of a "bias" provides little traction on that issue, not least because of the difficulty in providing a theory-neutral account of what qualifies as a "bias." Better, then, to tackle the problem of artificial consciousness head-on by asking what follows from our best models of human consciousness.

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Competing interests. The author has no competing interests to declare.

References

Anthis, J. R., Pauketat, J. V. T., Ladak, A., & Manoli, K. (2025). Perceptions of sentient AI and other digital minds: Evidence from the AI, morality and sentience (AIMS) survey. CHI '25. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, *10*(1), 22.

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 in Cognitive Sciences*, *28*(5), 454–466.

Block, N. (2002). The harder problem of consciousness. *Journal of Philosophy*, *99*(8), 391–425.

Dreksler, N., Caviola, L., Chalmers, D., Allen, C., Rand, A., Lewis, J., Waggoner, P., Mays, K., & Sebo, J. (2025). Subjective experience in AI Systems: What do AI researchers and the public believe? https://arxiv.org/abs/2506.11945

Free energy, consciousness, and the inverse dark room problem

Jonathan Birch*

The Jeremy Coller Centre for Animal Sentience at LSE Philosophy, London School of Economics and Political Science, Houghton, London, UK

j.birch2@lse.ac.uk
https://personal.lse.ac.uk/birchj1/

*Corresponding author.

doi:10.1017/S0140525X25103713, e320

Abstract

Seth's case for biological naturalism rests on a worryingly speculative foundation: the free energy principle. Arguments for computational functionalism are likewise speculative. Yet I am optimistic that the clash can be resolved empirically. Alleged dependencies of consciousness markers on biological properties can be tested comparatively, and a pattern of failed predictions would tilt the odds towards computational functionalism.

Advances in artificial intelligence (AI) add new urgency to an old question: is there a deep connection between life and consciousness? Anil Seth has presented a bold argument in favour. It won't break the deadlock, but it does suggest possible ways forward.

Biological naturalism can be a slippery target because the number of possible biological properties that *might* matter to consciousness is large, and the biological naturalist can always shift their gaze to a different property if their initially favoured candidates are recreated in AI. Differences in agency and embodiment that seem stark now might look much less so in a few years. There will, however, always be *some* differences. AI systems will never fully recreate cell-level metabolic processes. Consciousness might depend constitutively on these processes, but are there good reasons to think that it does?

Seth, in §4, weighs in with a novel argument from free energy minimization. He suggests that there is a crucial difference between a perceptual inference algorithm running in a biological system that works to minimize free energy right down to the cellular level and the same algorithm running in a non-living substrate that is not working to minimizing free energy at lower levels. The latter is said to be a "pale shadow" of the former, making it much less likely to support consciousness.

I see two main problems for this argument. One is that it begs the question against a computational functionalist opponent, who will deny the "pale shadow" thesis. They will not accept that what happens at lower levels of organization makes a difference to an algorithm's ability to support consciousness.

The second is that the free energy principle is too speculative a foundation to carry this much weight. As I see it, living things minimize prediction error when it serves their inclusive fitness interests, but when it doesn't, they don't. The "dark room problem" has been much discussed: organisms do not generally hide in dark caves even though this renders the external environment maximally predictable (Sun & Firestone, 2020). Fristonians have a fair comeback: staying in the cave renders the *internal* environment increasingly unpredictable because you cannot maintain homeostasis when starving (Seth et al., 2020). But to this we can add the *inverse* dark room problem: organisms sometimes *do* starve to death in dark caves when it serves their inclusive fitness interests. Female octopuses do this to protect their eggs (Wang, 2023). The only way to make sense of why organisms sometimes prioritize homeostasis and sometimes prioritize something else (such as mating opportunities, offspring, or other relatives) is through the principle of inclusive fitness maximization. To the extent biological processes optimize any single variable, it is inclusive fitness, not free energy (Grafen, 2006). There is no rhyme or reason to what

organisms do except the drumbeat of natural selection. Friston's claim to have found a second principle of equivalent or greater stature is exasperatingly under-evidenced.

Seth may reply: yes, the argument is speculative, but arguments for computational functionalism are speculative too. See, for example, Chalmers (1996) on fading and dancing qualia. Those arguments often beg the question against a biological naturalist opponent (see Birch, 2024, ch. 3, note 55). Both biological naturalism and computational functionalism are speculations lacking empirical validation. Neither deserves to be the default. All can agree we need ways to put claims about the life-consciousness connection, or lack of one, to empirical test.

Can the hypothesis of a constitutive relationship between consciousness and a biological property ever be tested? A worry arises that such hypotheses might be untestable. The challenge is that we would need to test for *direct* dependencies of phenomenology on biology, unmediated by effects on computation. If all the dependencies we can find are mediated, they will be compatible with computational functionalism. For example, disrupting the functioning of layer V pyramidal neurons will disrupt consciousness, but if the disruption is mediated by disruption to computations, computational functionalists need not worry. But how we could ever show a phenomenological effect to be unmediated by an effect on computation? For the effect to be reportable at all, there must be some difference at the level of computation, since report is itself a computational process.

The problem is serious, but I don't see it as insurmountable. To see grounds for optimism, first note that *some* claims about the life-consciousness connection can be falsified. Think, for example, of Beck and Eccles' (1992) proposed link between consciousness and quantum tunnelling during exocytosis. This specific hypothesis has dropped out of the discussion because it entailed predictions that were shown to be false (Georgiev & Glazebrook, 2014).

The hypothesis that consciousness depends on ephaptic coupling between neurons (mentioned briefly by Seth) also strikes me as testable. Myelination inhibits ephaptic coupling, so if it were essential to consciousness, we would expect to see fewer signs of consciousness in species with high degrees of myelination, such as cetaceans and elephants. This already looks implausible.

If life and free energy minimization were inseparable, it would be impossible to test for correlations between free energy minimization and consciousness in living things. But if they are, as I suggested earlier, readily separable, then testing for such a correlation becomes possible. We could investigate whether organisms pursuing free energy minimization display more markers of consciousness than those sacrificing their own homeostasis to maximize inclusive fitness.

So, specific versions of biological naturalism can be tested. This shows how, over time, we might gain traction on the bigger question of whether *any* version is correct. For if we repeatedly test different versions and find them wanting, we will be entitled to interpret the long-term pattern of failure as tilting the odds towards computational functionalism.

Granted, there will always be more biological properties that *might* matter, so we will not reach conclusive proof. But since computational functionalism predicts a pattern of failure when we look for dependencies of consciousness markers on any biological properties other than those involved in implementing computations, such a pattern would count in its favour.

Financial support. JB is supported by The Jeremy Coller Foundation and the Good Ventures Foundation.

Competing interests. None.

References

Beck, F., & Eccles, J. C. (1992). Quantum aspects of brain activity and the role of consciousness. *Proceedings of the National Academy of Sciences USA*, *89*(23), 11357–11361. https://doi.org/10.1073/pnas.89.23.11357

Birch, J. (2024). *The edge of sentience: Risk and precaution in humans, other animals, and AI*. Oxford University Press.

Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford University Press.

Georgiev, D., & Glazebrook, J. F. (2014). Quantum interactive dualism: from Beck and Eccles tunneling model of exocytosis to molecular biology of SNARE zipping. *Biomedical Reviews*, *25*, 15–24. https://doi.org/10.14748/bmr.v25.1038

Grafen A. (2006). Optimization of inclusive fitness. *Journal of Theoretical Biology*, *238*(3), 541–563. https://doi.org/10.1016/j.jtbi.2005.06.009

Seth, A. K., Millidge, B., Buckley, C. L., & Tschantz, A. (2020). Curious inferences: Reply to sun and firestone on the dark room problem. *Trends in Cognitive Sciences*, *24*(9), 681–683. https://doi.org/10.1016/j.tics.2020.05.011

Sun, Z., & Firestone, C. (2020). The dark room problem. *Trends in Cognitive Sciences*, *24*(5), 346–348. https://doi.org/10.1016/j.tics.2020.02.006

Wang, Z. Y. (2023). Octopus death and dying. *Integrative and Comparative Biology*, *63*(6), 1209–1213. https://doi.org/10.1093/icb/icad098

Compatibilist emergence for the science of consciousness

George Blackburne[1,2,3], Alberto Liardi[3], Jeremy I. Skipper[1,2], Pedro A. M. Mediano[3,2] and Fernando E. Rosas[4,5,6,7]*

[1]UCL Centre for Consciousness Research, University College London, London, UK; [2]Department of Experimental Psychology, University College London, London, UK; [3]Department of Computing, Imperial College London, London, UK; [4]Sussex Centre for Consciousness Science and Sussex AI, University of Sussex, London, UK; [5]Center for Complexity Science and Centre for Psychedelic Research, Imperial College London, London, UK; [6]Center for Eudaimonia and Human Flourishing, University of Oxford, Oxford, UK and [7]Principles of Intelligent Behavior in Biological and Social Systems, Prague, Czech Republic

george.blackburne.18@ucl.ac.uk
a.liardi@imperial.ac.uk
jeremy.skipper@ucl.ac.uk
pmediano@imperial.ac.uk
f.rosas@imperial.ac.uk

*Corresponding author.

doi:10.1017/S0140525X25103774, e321

Abstract

The relationship between subjectivity and its substrate can be understood via the notion of *compatibilist emergence*, accounting for the plurality of ways nature can host new dynamical laws at higher scales. We discuss how the autopoietic character of life may require and promote the kind of emergence necessary, albeit not sufficient, for consciousness.

Seth provides a timely and compelling set of arguments for why life could be crucial for consciousness. However, his account does not

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

clarify what exactly in life makes the difference – remaining agnostic whether this should be found in intrinsic, causal, or functional properties. Could a process that "looks like" life suffice, or does it need to be "life all the way down"?

It is nearly uncontroversial that consciousness operates on a larger scale than the constituent elements of its substrate. Thus, deliberating on the role of the substrate for consciousness requires a careful treatment of how different scales interact in the natural world and therefore the notion of emergence (Anderson, 1972). In the section *Emergence and the separation of scales*, Seth speculates that biological systems, by their very nature, cannot support emergent phenomena. To evaluate this claim, one needs to first articulate what emergence might mean in a world like ours.

To this end, we introduce the notion of a *compatibilist emergence* – akin to compatibilism in the philosophy of free will (Dennett, 1984). This position grants the potential for macroscopic scales to possess novel dynamical properties while simultaneously asserting that such properties obey supervenience. In other words, emergent features can be both unique irreducible processes in their own right while being contingent on their constituting factors, thus being at the same time substrate independent and dependent in distinct ways. Importantly, this view emphasizes not just the hierarchical spatial scales inherent to emergence but places explanatory weight on the way in which the system unfolds over time.

Under this theoretical purview, the scientific study of emergence has delivered multiple distinct frameworks for understanding emergent phenomena over dynamical processes (Barnett & Seth, 2023; Hoel, 2025; Hoel, Albantakis, & Tononi, 2013; Rosas et al., 2020, 2024; Seth, 2010). Empirical applications of these frameworks are revealing that emergent phenomena are pervasive in the brain – and, what's more, that neurobiology has no monopoly over them (Hoel & Levin, 2020; Klein & Hoel, 2020; Klein et al., 2021; Luppi et al., 2022; Mediano et al., 2021; Milinkovic et al., 2025; Pigozzi, Goldstein, & Levin, 2025; Tolle et al., 2024; Varley, 2020, 2023). Interestingly, these frameworks possess significant differences, coalescing into a collection of distinct facets of emergence:

- Emergence through *macroscopic sufficiency* – when a macro contains all the information relevant to itself (Barnett & Seth, 2023).
- Emergence through *macroscopic causal dominance* – when it is most *efficient to control the sys*tem from the macro (Hoel et al., 2013).
- Emergence through *mereological constraints* – when the whole carries more information than the sum of its parts (Mediano et al., 2021; Rosas et al., 2020).

This taxonomy of facets of emergence thus forces us to ask: if consciousness is an emergent process, then what type of emergence could it rely upon?

Seth claims that the generative entrenchment inherent in evolutionary design and the autopoietic imperative of organisms to regenerate their own basis may prevent biological systems from demonstrating macroscopic sufficiency. However, it is important to note that while the argument Seth provides may apply to one kind of emergence (macroscopic sufficiency), it does not

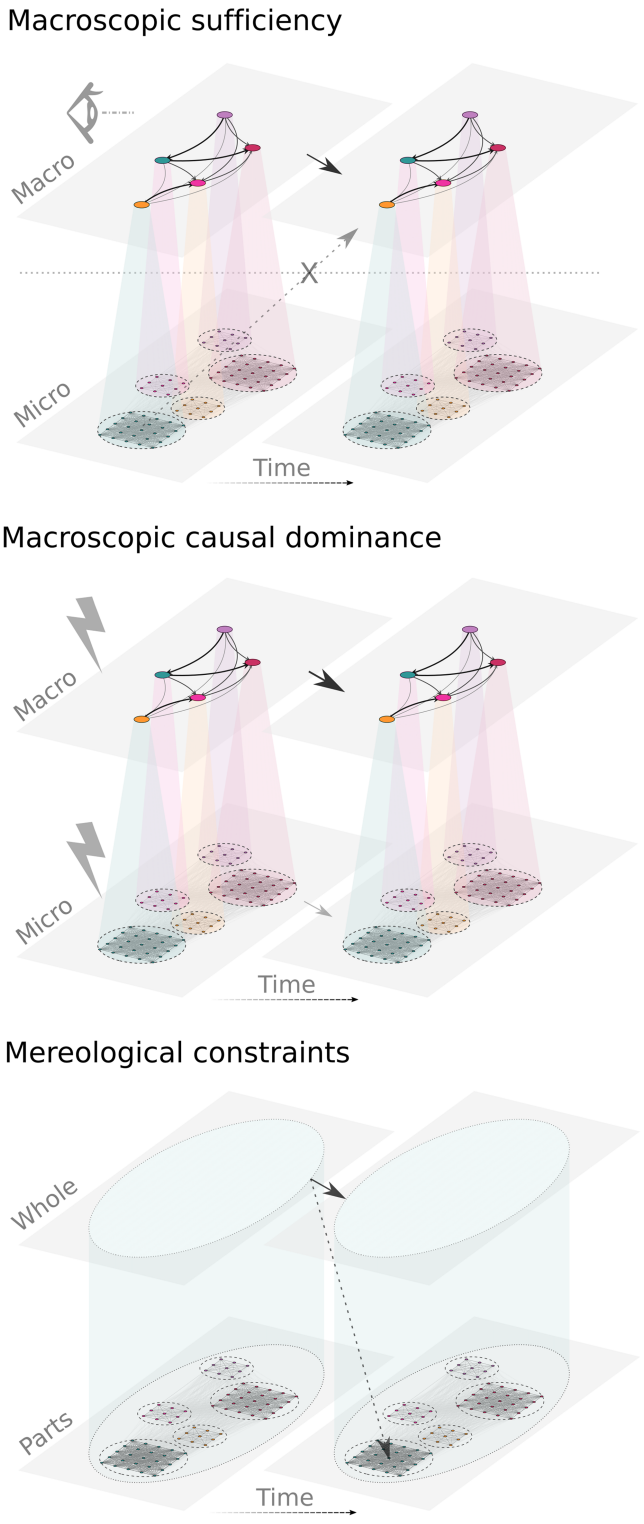


Figure 1 (Blackburne et al.). Facets of emergence. Top: Emergence through *macroscopic sufficiency* occurs if the microscale history adds no predictive power to the macroscale's own self-prediction. Middle: Emergence through *macroscopic causal dominance* occurs if the coarse-grained macroscale causally constrains its own possible past and future states more effectively than the microscale does for its respective states. Bottom: Emergence through *mereological constraints* occurs if the whole has irreducible predictive power over itself or individual parts.

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

undermine the relevance of others. On the contrary, autopoiesis implicitly invokes the latter forms of emergence, as it raises the question of which higher scales have been endowed with the appropriate horizontal and downward causal powers to execute these functions.

Seen from this perspective, the link that Seth proposes between self-organization, self-reproduction, and consciousness is compelling from both theoretical and empirical standpoints. Theoretically, at the highest scale, this may be a functional reason for why brains bother to furnish particular mental states with experience at all, since feeling carries the quality of intrinsic value and care that cognition-as-usual cannot offer in service of self-maintaining and directed activity (Cleeremans & Tallon-Baudry, 2022). Empirically, while the human brain appears to exhibit macroscopic sufficiency, this may hinder rather than help consciousness (Milinkovic et al., 2025). In contrast, considering consciousness as emerging through mereological constraints, not only does evidence suggest that brains can orchestrate this kind of emergence across scales (McSharry et al., 2024), but indeed it seems to be attenuated in states of reduced consciousness.

This link between life, consciousness, and mereological constraints is supported by the fact that our nature as living organisms is one of nested interdependent parts and wholes, rather than consistently autonomous macroscales (Levin, 2019). In effect, one way to distinguish machines from organisms is that while for both the whole is made of and depends on the parts, the parts of a machine are oblivious of the whole, while the parts of an organism critically depend on the whole for their subsistence (Nicholson, 2013). In a similar way, consciousness may require the integrity of the whole to allow unified percepts but also the ability of that whole to constrain its parts – for example, to allow perceptual filling-in.

In sum, consciousness as a biological process is more plausibly realized not through separation but through interdependence, feedback, and synergy. Thus, contrary to Seth's claim, it may be *because of* – rather than *in spite of* – emergence that the distinction between software and hardware may not be the right way of understanding creatures like us. In this way, we reiterate Seth's remark that the properties of life have a significant purchase in the prediction and control of features of consciousness. However, we believe that this is, at its core, true because life has found itself to be a potent means of orchestrating the kind of emergence crucial for consciousness.

Lastly, we emphasize that although emergence may be crucial for consciousness, it is likely to be a *necessary but not sufficient* condition for it. This implies that the ability to display emergent properties does not necessarily result in synthetic phenomenology. Emergence is the medium, not the message – and as extolled here, it has many dimensions. Accordingly, optimizing artificial systems involves choosing a specific facet of emergence. For instance, naively pursuing macroscopic sufficiency alone is likely to lead to systems with runaway scales, lacking relevant bottom-up constraints, and possessing aberrant structure. Despite these limitations, measures of emergence may still serve as a useful indicator of whether a particular system is worthy of further inspection for the presence of consciousness.

Financial support. G.B. is supported by the Leverhulme Trust. J.S. is supported by Wellcome Leap. P.M. is supported by the Tiny Blue Dot Foundation. F.R. is supported by the PIBBSS Affiliateship program.

Competing interests. The authors declare no competing interests.

References

Anderson, P. W. (1972). More is different: Broken symmetry and the nature of the hierarchical structure of science. *Science*, *177*(4047), 393–396. https://doi.org/10.1126/science.177.4047.393.

Barnett, L. & Seth, A. (2023). Dynamical independence: Discovering emergent macroscopic processes in complex dynamical systems. *Physical Review E*, *108*(1–1), 014304. https://doi.org/10.1103/PhysRevE.108.014304.

Cleeremans, A. & Tallon-Baudry, C. (2022). Consciousness matters: Phenomenal experience has functional value. *Neuroscience of Consciousness*, *2022*(1), niac007. https://doi.org/10.1093/nc/niac007.

Dennett, D. C. (1984). *Elbow room*. MIT Press.

Hoel, E. (2025). *Causal emergence 2.0: Quantifying emergent complexity*. arXiv preprint arXiv:2503.13395.

Hoel, E. & Levin, M. (2020). Emergence of informative higher scales in biological systems: A computational toolkit for optimal prediction and control. *Communicative & Integrative Biology*, *13*(1), 108–118.

Hoel, E. P., Albantakis, L., & Tononi, G. (2013). Quantifying causal emergence shows that macro can beat micro. *PNAS*, *110*(49), 19790–19795. https://doi.org/10.1073/pnas.1314922110.

Klein, B. & Hoel, E. (2020). The emergence of informative higher scales in complex networks. *Complexity*, *2020*(1), 1–12. https://doi.org/10.1155/2020/8932526.

Klein, B., Hoel, E., Swain, A., Griebenow, R., & Levin, M. (2021). Evolution and emergence: Higher order information structure in protein interactomes across the tree of life. *Integrative Biology*, *13*(12), 283–294.

Levin, M. (2019). The computational boundary of a "self": Developmental bioelectricity drives multicellularity and scale-free cognition. *Frontiers in Psychology*, *10*, 2688. https://doi.org/10.3389/fpsyg.2019.02688.

Luppi, A. I., Mediano, P. A. M., Rosas, F. E., Holland, N., Fryer, T. D., O'Brien, J. T., Rowe, J. B., Menon, D. K., Bor, D., & Stamatakis, E. A. (2022). A synergistic core for human brain evolution and cognition. *Nature Neuroscience*, *25*(6), 771–782. https://doi.org/10.1038/s41593-022-01070-0.

Luppi, A. I., Mediano, P. A. M., Rosas, F. E., Allanson, J., Pickard, J., Carhart-Harris, R. L., Williams, G. B., Craig, M. M., Finoia, P., Owen, A. M., Naci, L., Menon, D. K., Bor, D., & Stamatakis, E. A. (2024). A synergistic workspace for human consciousness revealed by integrated information decomposition. *Elife*, *12*, RP88173. https://doi.org/10.7554/eLife.88173.

McSharry, D., Kaplanis, C., Rosas, F. E., & Mediano, P. (2024). Learning diverse causally emergent representations from time series data. *Neural Information Processing Systems*, *37*, 119547–119572.

Mediano, P. A. M., Rosas, F. E., Luppi, A. I., Carhart-Harris, R. L., Bor, D., Seth, A. K., & Barrett, A. B. (2021). Towards an extended taxonomy of information dynamics via integrated information decomposition. arXiv [q-bio.NC].

Milinkovic, B., Seth, A. K., Barnett, L., Carter, O., & Andrillon, T. (2025). Dynamical independence reveals anaesthetic specific fragmentation of emergent structure in neural dynamics. *bioRxiv*, page 2025.07.16.664881. https://doi.org/10.1101/2025.07.16.664881.

Nicholson, D. J. (2013). Organisms ≠ machines. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *44*(4), 669–678.

Pessoa, L., Thompson, E., & Nöe, A. (1998). Finding out about filling-in: A guide to perceptual completion for visual science and the philosophy of perception. *Behavioral and Brain Sciences*, *21*(6), 723–48; discussion 748–802. https://doi.org/10.1017/s0140525x98001757.

Pigozzi, F., Goldstein, A., & Levin, M. (2025). Associative conditioning in gene regulatory network models increases integrative causal emergence. *Communications Biology*, *8*(1), 1027.

Rosas, F. E., Mediano, P. A. M., Jensen, H. J., Seth, A. K., Barrett, A. B., Carhart-Harris, R. L., & Bor, D. (2020). Reconciling emergences: An information-theoretic approach to identify causal emergence in multivariate data. *PLOS Computational Biology*, *16*(12), e1008289. https://doi.org/10.1371/journal.pcbi.1008289.

Rosas, F. E., Geiger, B. C., Luppi, A. I., Seth, A. K., Polani, D., Gastpar, M., & Mediano, P. A. M. (2024). Software in the natural world: A computational approach to hierarchical emergence. *arXiv [nlin.AO]*.

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

Seth, A. K. (2010). Measuring autonomy and emergence via granger causality. *Artificial Life*, *16*(2), 179–196. https://doi.org/10.1162/artl.2010.16.2.16204.

Tolle, H. M., Luppi, A. I., Seth, A. K., & Mediano, P. A. M. (2024). Evolving reservoir computers reveals bidirectional coupling between predictive power and emergent dynamics. arXiv [q-bio.NC].

Varley, T. F. (2020). Causal emergence in discrete and continuous dynamical systems. arXiv [nlin.CG].

Varley, T. F. (2023). Decomposing past and future: Integrated information decomposition based on shared probability mass exclusions. *PLoS One*, *18*(3), e0282950. https://doi.org/10.1371/journal.pone.0282950.

Varley, T. F. and Hoel, E. (2022). Emergence as the conversion of information: A unifying theory. *Philosophical Transactions of the Royal Society A: Mathematical, Physical, and Engineering Sciences*, *380*(2227), 20210150. https://doi.org/10.1098/rsta.2021.0150.

A speculative argument against consciousness in AI (and perhaps some invertebrates)

Ned Block*

Department of Philosophy, New York University, New York, NY, USA
ned.block@nyu.edu

*Corresponding author.

doi:10.1017/S0140525X25103567, e322

Abstract

Brain waves driven by electrochemical processing may be necessary for consciousness, and if so, that will challenge consciousness in artificial intelligence (AI) and perhaps in some invertebrates.

I will sketch one line of speculative support for Seth's thesis of substrate-dependence of consciousness.

It has long been recognized that brain rhythms are important in neural functioning (Kopell et al., 2014). At the dawn of the connectome era, Kopell et al. suggested that we think not in terms of a static connectome but rather dynamically, in terms of the "dynome," and the key to the dynome is brain rhythms. Recent evidence suggests that electrochemical mechanisms may be required for brain rhythms and that brain rhythms are required for sophisticated processing. If – and this is highly speculative – they are required for consciousness itself, that would preclude consciousness in non-biological AI systems.

The connectome project has strikingly confirmed the widely held view that individual neurons, especially in the vertebrate cortex, tend to have so many synapses on each of their axons that pairwise synapses are too weak to explain post-synaptic neural firing (Fields, 2025; Shapson-Coe et al., 2024). R. Douglas Fields puts this point as follows (2025, p. 5): "Ironically, the goal of the brain initiative to trace out the connectome of the human brain has revealed that there is none." More specifically, in the vertebrate cortex, most synapses are "en passant," axons have many synaptic offshoots along the length of the axon. The result is that the activity in these synapses "are too weak to depolarize the postsynaptic neuron to the action potential threshold" (Fields, 2025, p. 1).

Instead, neural firing in the cortex depends on the joint action of populations of neurons themselves partly driven by causal interactions outside the neurons. For example, there are waves of ion fluctuations in the chemical soup surrounding neurons. Every charged particle generated by a neuron has to return via this soup. And glial cells regulate ion concentration in the soup and absorb and release neurotransmitters. These factors, together with subthreshold oscillations in cell somas, dendrites, and axons, together with ionic oscillations in the soup, create powerful population-wide electrical waves that are part of what drives the firing of neurons (Fields, 2025). These waves make up for the weakness of direct pairwise connections between neurons.

As Peter Godfrey-Smith notes (2024), if we favor gradualism in evolution, one place to look for clues to the nature of consciousness would be the evolutionary origins of consciousness. The first animals appear to have been comb jellies (Schultz et al., 2023). What Schultz et al. showed was that comb jelly chromosomes were radically different from those of subsequent animals, including sponges and us. Comb jellies have neurons, but their immediate successor, sponges, do not. Comb jellies might have been on a dead-end evolutionary branch. Or else sponges evolved from comb jellies but lost their neurons. Either way, comb jelly neurons didn't seem to lead to more advanced systems. Burkhardt et al. (2023) showed that the comb jelly nervous system (of at least one phase in their lifecycle) was entirely electrical. As *Scientific American* (Giovanetti, 2024) asked, "Did Neurons Evolve Twice?" The answer is yes in both scenarios, so it looks like electrochemical processing (in which neurotransmitters are released by one neuron and trigger electrical activation in another neuron) has a functional advantage over purely electrical processing.

As Seth notes, dynamical approaches to consciousness leave open the question of whether the substrate matters and why. In this case, the issue is whether there is something about the electrochemical implementation of these brain rhythms that is crucial to consciousness. The new results don't answer that question, but they do suggest that electrochemical operation may have been superior in one or another way that aeons later led to sophisticated vertebrate brains.

Even if electrochemical processing is necessary for consciousness in actual creatures, and is "mortal" in the sense of requiring a certain substrate in actual creatures, that leaves open *why* that is so and does not entail what Seth calls "strong" substrate dependence in which consciousness is supposed to depend on "intrinsic" properties of its material basis.

The argument so far is: sophisticated cortical processing requires brain rhythms. Brain rhythms depend on electrochemical processing. Speculation: consciousness in us depends on electrochemical processing. Further speculation: consciousness tout court depends on those mechanisms.

Here is another layer of speculation: Brain rhythms may depend on low-frequency oscillations (Bullock & Basar, 1988). Notably, those low-frequency oscillations are more prominent in vertebrates than in many invertebrates. One kind of glial cell, the oligodendrocyte, modulates myelin, affecting low-frequency neuronal firing (Pajevic et al., 2023). Both oligodendrocytes and myelin are unique to vertebrates. The low-frequency waves spread much more widely than high-frequency waves, on the order of centimeters for under 20 Hz but at the micrometer level for over 50 Hz. Although there are low frequency brain waves in invertebrates

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

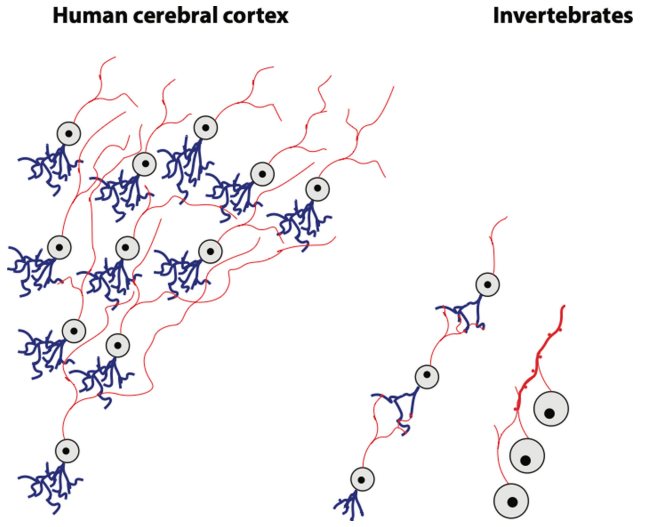


Figure 1 (Block). In vertebrates, axons (red lines) wander through a forest of dendrites (blue lines), with many synapses. The circles with dots are the cell bodies. There is a tendency in invertebrates toward synapses that are stronger because axons link to only a few dendrites or even one dendrite. From (Fields, 2025), reprinted by permission of Sage Publications.

(Marder, Goeritz, & Otopalik, 2015), many invertebrate neural systems (with the notable exception of the octopus (Bullock & Basar, 1988)) are said to be weak in highly coherent low frequency brain waves (Fields, 2025) and in some, especially more primitive invertebrates, consist mainly in strong serial connections with much less need for the brain waves. See Figure 1.

However, there are slow waves in NREM sleep and under anesthesia. Siclari et al. (2018) distinguish two types of slow waves. NREM dreams take place in a background of slow waves, but are associated with decreases in the delta range in one type of slow wave in central and posterior cortex, together with increases in the other type of slow wave in frontal areas. Slow waves are present in anesthesia (Pazienti et al., 2022), but I know of no data on which of the two kinds of slow waves is present.

In sum, there is an intriguing but empirically motivated speculation that argues against consciousness in AI and a further speculative line of thought that may exclude consciousness in some invertebrates.

Acknowledgements. Thanks to R. Douglas Fields, Nancy Kopell, and Giulio Tononi for helpful conversation.

Financial support. No funding.

Competing interests. There are no competing interests.

References

Bullock, T. H., & Basar, E. (1988). Comparison of ongoing compound field potentials in the brains of invertebrates and vertebrates. *Brain Research*, *472*(1), 57–75. https://doi.org/10.1016/0165-0173(88)90005-7

Burkhardt, P., Colgren, J., Medhus, A., Digel, L., Naumann, B., Soto-Angel, J. J., … & Kittelmann, M. (2023). Syncytial nerve net in a ctenophore adds insights on the evolution of nervous systems. *Science*, *380*(6642), 293–297. https://doi.org/10.1126/science.ade5645

Fields, R. D. (2025). Connectivity in the human cerebral cortex: A fundamental problem and a possible explanation for the cognitive power of vertebrates. *The Neuroscientist*, *31*(5), 458–463. 10738584251337656. https://doi.org/10.1177/10738584251337656

Giovanetti, C. (2024). Did neurons evolve twice. *Scientific American*.

Godfrey-Smith, P. (2024). Inferring consciousness in phylogenetically distant organisms. *Journal of Cognitive Neuroscience*, *36*(8), 1660–1666. https://doi.org/10.1162/jocn_a_02158

Kopell, N. J., Gritton, H. J., Whittington, M. A., & Kramer, M. A. (2014). beyond the connectome: The dynome. *Neuron*, *83*(6), 1319–1328. https://doi.org/10.1016/j.neuron.2014.08.016

Marder, E., Goeritz, M. L., & Otopalik, A. G. (2015). Robust circuit rhythms in small circuits arise from variable circuit components and mechanisms. *Current Opinion in Neurobiology*, *31*, 156–163. https://doi.org/10.1016/j.conb.2014.10.012

Pajevic, S., Plenz, D., Basser, P. J., & Fields, R. D. (2023). Oligodendrocyte-mediated myelin plasticity and its role in neural synchronization. *eLIFE*, *12*, e81982. https://doi.org/10.7554/eLife.81982

Pazienti, A., Galluzzi, A., Dasilva, M., Sanchez-Vives, M. V., & Mattia, M. (2022). Slow waves form expanding, memory-rich mesostates steered by local excitability in fading anesthesia. *iScience*, *25*(3), 103918. https://doi.org/10.1016/j.isci.2022.103918

Schultz, D. T., Haddock, S. H. D., Bredeson, J. V., Green, R. E., Simakov, O., & Rokhsar, D. S. (2023). Ancient gene linkages support ctenophores as sister to other animals. *Nature*, *618*(7963), 110–117. https://doi.org/10.1038/s41586-023-05936-6

Shapson-Coe, A., Januszewski, M., Berger, D. R., Pope, A., Wu, Y., Blakely, T., … & Lichtman, J. W. (2024). A petavoxel fragment of human cerebral cortex reconstructed at nanoscale resolution. *Science*, *384*(6696), eadk4858. https://doi.org/10.1126/science.adk4858

Siclari, F., Bernardi, G., Cataldi, J., & Tononi, G. (2018). Dreaming in NREM sleep: A high-density EEG study of slow waves and spindles. *Journal of Neuroscience*, *38*(43), 9175–9185. https://doi.org/10.1523/jneurosci.0855-18.2018

Anil Seth's case for biological naturalism provides a roadmap for a conscious AI

Lenore Blum* and Manuel Blum

Computer Science Department, Carnegie Mellon University, Pittsburgh, USA
lblum@cs.cmu.edu
mblum@cs.cmu.edu

*Corresponding author.

doi:10.1017/S0140525X25103518, e323

Abstract

We comment on Anil Seth's target article from the perspective of theoretical computer science, a branch of mathematics that considers computation under limited resources, and from our own perspective developing a formal machine model for consciousness, the Conscious Turing Machine. From these perspectives, we argue that Seth makes a case for artificial intelligence consciousness. We also address misconceptions about computation.

In the Abstract to Seth's article, he proposes to answer, "what it would take for conscious AI [artificial intelligence] to be a realistic prospect" by (a) challenging "the assumption that computation provides a sufficient basis for consciousness" and by (b)

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

making a case "that consciousness depends on our nature as living organisms – a form of biological naturalism."

Due to a word limit, we focus mainly on (b), contending that in making the case for biological naturalism, Seth likewise makes a case for the inevitability of AI consciousness.

To do this, we consider features that Seth views as essential for biological consciousness and show how these features are also realized in a simple machine model of consciousness, the Conscious Turing Machine (CTM). At the same time, we highlight how CTM aligns with underlying principles of theories that Seth views as promising for "real artificial consciousness." These theories include global workspace theory (GWT); attention schema theory (AST); the embodied, embedded, enactive, and extended (4E) theories borrowed from cognitive science; and integrated information theory (IIT).

To support a case for biological naturalism, Seth enumerates features that he views as key to consciousness, features that presumably evolved to keep living beings alive. He comments often on the richness of biological systems and human experiences. No argument here. But to conclude, as he often does, that AI systems could not entertain these very features is a non-sequitur.

The focus of Seth's argument for biological naturalism (section 4.0) relies primarily on two related notions: *predictive processing* and the *free energy principle*, both of which highlight the important cybernetic process of *prediction error minimization*. According to Seth, prediction error minimization is key to the entity making "best guesses" about sensory signals (both external and internal). These best guesses are what generate *controlled hallucinations*, which Seth describes as providing our conscious perception and sense of self. Additionally, by minimizing sensory error predictions, Seth asserts that the free energy principle enables the system to maintain boundaries between internal and external states, further supporting conscious experience via controlled hallucination.

To comment on the realization of prediction error minimization and controlled hallucinations in CTM, we start with a brief overview of CTM, highlighting features that align with underlying principles of major theories of consciousness. See Blum and Blum (2022) and Blum and Blum (2025).

CTM's architecture is inspired by the *GWT* "theater model" of consciousness. For simplicity, its *stage*, STM, holds one (formally defined) *chunk* at a time. Its *audience*, LTM, consists of an enormous collection of initially independent powerful processors. Over time, *links* are created between like-minded processors. While processors each have their own internal language, they communicate with each other in Brainish, CTM's self-generated multimodal language.

For simplicity, all CTM processors are in LTM, including its Input (sensory) and Output (motor) processors and Gauge processors (*embodied* homeostats, e.g., fuel and temperature gauges). Initially, all other processors are generic, their functionality emerging over time. The CTM is *embedded* and *enacts* in its outer world via its *Input sensors* and *Output actuators*. Its functionality is *extended* through interactions with its outer world and by augmenting LTM with "off-the-shelf" processors. As we shall see, global integration and extensive feedback in CTM align with *IIT*'s underlying principles for achieving high phi.

At each (discrete) moment in time, processors submit chunks into a well-defined *fast competition* to get on stage. The winning chunk is then *globally broadcast* to all LTM processors triggering CTM's *attention* and, by CTM's Axiom 1, a unitary experience.

For *awareness* and *feelings of consciousness* in CTM, the co-evolution of Brainish and CTM's world models enter the picture. A sketch of a rose in a world model might be labeled with the multisensory Brainish gist, "bright red color, sweet smell, soft touch." When CTM *inspects* a broadcasted chunk whose Brainish gist denotes this labeled sketch, CTM will become *aware* of the labeled sketch. A unitary and, by Axiom 2, a fused subjective experience will be evoked. CTM will experience the rose's bright red color, sweet smell, soft touch of its petal.

This notion of awareness aligns with AST and with the role of Seth's *hallucinations*.

CTM's *predictive dynamics* (cycles of prediction, feedback and learning, which work to minimize prediction error, locally and globally) are analogous to both *predictive processing* and the *free energy principle* and play a major role in the co-evolvement of Brainish and CTM's world models. We also note that CTM's Input sensors and Output actuators serve as CTM's Markov blanket.

Seth argues that while predictive processing may seem like a computational process, its continuity with the free energy principle makes it "a dynamical, non-computational theory." We beg to differ.

This takes us to computation and Seth's challenge (a).

We note that in CTM, time is discrete. Additionally, CTM's processes employ digital computations. Utilizing predictive dynamics, these processes evolve over time, affected by and responding to circumstances. However, Seth seems to assume that a machine that uses digital computations must be rigid. Indeed, to reinforce his case for biological naturalism, Seth proclaims, "Instead of a computational rigidity, there is an excitable fluidity to the matter of life."

On numerous occasions, Seth surmises that while digital computations might approximate neural processes, "it would seem an open question whether these approximations would be good enough to realize consciousness." On the other hand, he notes that since "exact Bayesian inference is generally analytically intractable, the brain implements an approximation."

To explore the feasibility of alternate computing possibilities, Seth speculates about non-computable functions and the potential of rectifying them with quantum, analogue, neuromorphic, and mortal computation. Of course, some are "straw men." For example, according to the mathematical theory of quantum computation, the class of quantum and Turing computable functions is identical. As for speed-ups, although algebraic problems like factoring are quantum computable in polynomial time, there is evidence that this is not the case for NP-complete problems. See Bernstein and Vazirani (1997). We refer to Nayebi (2014), which likewise disputes the claim that analogue computing could be more powerful than digital computing.

Conscious AI is inevitable.

Acknowledgments. We thank Johannes Kleiner, Aran Nayebi, and Umesh Vazirani for extremely helpful discussions. We take responsibility for our arguments and conclusions.

Financial support. We are grateful to UniDT for helping to support our research.

Competing interests. Neither author has competing interests.

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

References

Bernstein, E., & Vazirani, U. (1997). Quantum complexity theory. *SIAM Journal on Computing*, *26*(5), 1411–1473. https://doi.org/10.1137/S0097539796300921.

Blum, L., & Blum, M. (2022). A theory of consciousness from a theoretical computer science perspective: Insights from the conscious Turing machine. *Proceedings of the National Academy of Sciences*, *119*(21). https://doi.org/10.1073/pnas.21159341.

Blum, L., & Blum, M. (2025). AI consciousness is inevitable: A theoretical computer science perspective. Retrieved from arXiv: https://arxiv.org/pdf/2403.17101

Nayebi, A. (2014). Practical intractability: A critique of the hypercomputation movement. *Minds & Machines*, *24*, 275–305. https://doi.org/10.1007/s11023-013-9317-3.

Birds do it, we do it, even embodied machines could do it. Why consciousness is not intrinsically biological.

Simon Bowes*

School of Engineering and Informatics, University of Sussex, Brighton, UK
S.C.Bowes@sussex.ac.uk
https://profiles.sussex.ac.uk/p169679-simon-bowes

*Corresponding author.

doi:10.1017/S0140525X25103737, e324

Abstract

Digital computation may not be sufficient for consciousness, but this does not warrant biological naturalism. Rejecting coarse-grained Turing-computational functionalism and adopting a broader computationalism with no built-in syntax/semantics distinction, undermines the motivation for biological naturalism. Substrate-dependent processes of the kind found in living things may be necessary for the emergence of consciousness, but these are functional requirements, not intrinsic properties.

I agree that to be conscious, "artefacts may have to be closer to biological systems than allowed for by conventional digital computation," but not that this implies Biological Naturalism: "closer" can be functionally defined using "unconventional" computation. Intelligence and consciousness partly dissociate, but not because one is functional and the other intrinsic. Unconscious algorithms produce intelligent behaviour, but if consciousness evolved, it is because it does too. If it additionally has intrinsic, non-functional properties, Zombies lurch into view.

Consciousness *is* essentially about what a system does – we are conscious for a reason. "Feeling like *some*thing" does not imply intrinsicality. Experiences have the character they do due to "bottom-up" and "top-down" processes; to form unified experiences with intrinsic properties, these would need to be added together somehow. This is a form of Cartesian Materialism, which there are reasons to avoid (Dennett, 1991). However, my focus first is on the supposed incompatibility of computation and consciousness.

It is true that being usefully describable in terms of computations does not entail the brain computes, just as we might usefully describe a falling brick computationally, without assuming the brick computes. But we have independent reasons to say the brain computes, namely, intentionality. Mental states are about things, they are informational, so the processes they are involved in can be said to compute, in a way bricks do not.

"Coarse-grained" functionalism *is* too liberal, but finer-grained varieties are available. Turing-computational functionalism is particularly coarse-grained, but "4E" cognitive science supports other kinds of computationalism. Virtual Machine Functionalism (Sloman & Chrisley, 2003; Bowes, 2023) captures "generatively entrenched" organisational features required for mindedness, avoiding overly liberal functionalism: parts play their roles due to (contextual, dispositional) causal powers. The "limited substrate flexibility" of neurobiological functions makes Turing-computational functionalism unsuitable, but there are "broader forms of computation which have similar constraints." However, broadening our understanding of computation undermines the argument for biological naturalism.

Embodied and embedded forms of computation are not "pale shadows" of life. There may be an "excitable fluidity" to biological entities ripe for the emergence of consciousness, but why think only biological entities can have this? The computers we build are "rigid" because we design them to blindly follow rules to match the semantics we assign to symbols. But computation does not necessitate a strict syntax/semantics separation:

> "... a fundamental theory of programs, and of computation, should acknowledge that an essential function of a computer program is to make things happen. Whereas symbolic logic can be viewed as mere playing around with uninterpreted formal calculi ..., and computational logic can be seen as the study of abstract timeless relations in mathematically specified 'machines' (such as Turing machines), computer science cannot properly be described in either of these ways. It follows ... that the familiar characterization of computer programs as all syntax and no semantics is mistaken." (Boden, 1988)

The point is that arguments like Searle's (1980) depend on the artificial separation of implementation from semantics, whereas in embodied and embedded computational systems, these cannot be separated.

So, natural minds are embodied and embedded physical computers, with representational content derived from engagement with a world, and substrate dependence that enables "relevant patterns of functional organisation."

Seth uses the lack of substrate flexibility against embodied computation as it "might not be implementable in ... non-biological substrates." But, if radical substrate independence makes Turing computation unsuitable for consciousness, the need to be embodied and embedded in the right way should be seen as an advantage for a broad-minded computationalism.

The way to avoid overly liberal functionalism is to foreground embodiment – momentarily matching an abstract description is insufficient to realise a computation. There needs to be *some* substrate flexibility, though, so that relevantly similar mental states can be realised in reasonably varied embodiments. Maybe only biological substrates currently implement the right kind of

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

computation, but knowing what makes them suitable should make replication possible. Can we just stipulate that these "replicants" are biological?

Accepting that "excitable" autopoietic processes are responsible for the emergence of life as we know it, then such processes were sufficient for the emergence of consciousness. But were they necessary, and would that support biological naturalism? The claim that consciousness came about through natural processes, and that carbon-based life on Earth is one way this could happen, is just naturalism.

"Life" may be defined by "homeostatic processes . . . linked to higher-level functional properties," which are emergent, in that "knowledge of (or interventions at) lower, microscopic levels does not add anything to macro-level predictability, or help determine macro-level outcomes." Seth says consciousness could depend on functional or intrinsic properties of the biological substrate, but that, given emergence, it is the properties at the level of life that matter. But the dependence is stronger, otherwise, the higher-level properties would be inexplicable - emergent in the "bad" way.

Seth claims that "apparently intrinsic properties at one level might always be understandable in terms of patterns of functional organisation at more fine-grained levels." This makes intrinsicality perspectival, related to our explanatory interests. But this seems to conflate intrinsicality with explanatory autonomy. Instead of being "intrinsic," non-relational properties, they are irreducible – emergent, but not in the sense of inexplicably appearing. To be explicable, and so of scientific interest, they cannot be intrinsic, as explanation requires relations (just not necessarily reductive ones). This implies inherently irreducible, rather than intrinsic, causal powers.

The higher-level causal properties of mindedness, then, are computational in a broad sense; being biological is part of the "finer-grained functional substrate" out of which they may emerge. The additional claim of necessity depends on an assumption that only biological entities can have the suitable properties, which has not been established.

Seth's argument seems to be based on a Cartesian intuition regarding the intrinsicality of consciousness (that it's not "about what a system does"), against functionalist accounts. This is in tension with his "real problem" approach, and is prone to zombie attack: given the substrate's "finer-grained" properties, could there still be something missing, some intrinsic property that emerges in a way that is not explained by those functional facts plus their embodied history? If not, we should accept functionalism. If so, then the hard problem "pops up" again.

Acknowledgements. None.

Financial support. None.

Competing interests. None.

References

Boden, M. A. (1988). Escaping from the Chinese room. In *Computer models on mind: computational approaches in theoretical psychology*. Cambridge University Press.

Bowes, S. (2023). *Naturally minded: Mental causation, virtual machines, and maps*. Springer Verlag.

Dennett, D. C. (1991). *Consciousness explained*. Penguin Books.

Searle, J. (1980). Minds, brains and programs. *Behavioral and Brain Sciences*, *3*, 417–457.

Sloman, A., & Chrisley, R. (2003). Virtual machines and consciousness. *Journal of Consciousness Studies*, *10*, 133–172.

Why biological naturalism? Because consciousness requires having your own good

Rosa Cao*, David Gottlieb and Jared Moore

Stanford University, Stanford, CA, USA
rosacao@stanford.edu
dmg1@stanford.edu
jlcmoore@stanford.edu
https://sites.google.com/site/luosha/home/research
https://jaredmoore.org/

*Corresponding author.

doi:10.1017/S0140525X25103804, e325

Abstract

Living things are special because they are valuing subjects, for whom things can be good or bad. This is in virtue of their having internalized optimization processes, tied to persistence, that evaluate whether things are going well or poorly and adjust appropriately. Those evaluations, when integrated from diverse sources and widely available, provide the basis for conscious experience.

Suppose I push a needle into the paw of a stuffed bear. You, watching, may wince. But nothing feels bad to the bear itself, because it isn't a conscious subject. Indeed, nothing can *be* bad for the bear. From *our* perspective, it might be bad if the bear were destroyed – perhaps because it wouldn't then comfort anyone. But nothing is either good or bad for the bear itself, independent of what anyone might think about it – it isn't what we'll call a *valuing subject*. The bear is not alone: nor will a simulated person who winces when pinched feel pain. Their *apparent* good depends on our (external) interpretation of the simulation because what a system computes depends on how the system is interpreted (cf. Rescorla, 2013).

In this commentary, we pursue Seth's conjecture that only living things can be conscious ("biological naturalism"). Seth motivates his view with several theories (the free energy principle (FEP), predictive processing (PP), and active inference) and attention to the autopoietic and action-oriented aspects of living things. We build on this by offering a specific proposal about why life may be necessary for consciousness.

On our view, conscious subjects are always valuing subjects, and valuing subjects are generally alive (although depending on the definition of "life," there might be exceptions). This is important because FEP, PP, and even some versions of the global workspace theory of consciousness (GWT) specify only computational/functional structure, entirely applicable to systems that seem unlikely to be either conscious or alive (e.g., video compression algorithms or email servers). By requiring that the states broadcast to the global workspace (or the errors minimized in PP) be *evaluative* states that play a distinctive role in making things go well with respect to the valuing subjects' *own* good, we rule out non-living systems that lack their own good.

Why are living things valuing subjects? Two defining features of life are suggestive; both involve being subject to an ongoing

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

optimization process that robustly corrects deviations away from a "goal-state." First, living things sustain themselves metabolically through a set of interwoven homeostatic (or allostatic) processes, with regulatory mechanisms operating to maintain physiologically favorable conditions (Bich & Bechtel, 2025). Second, they are subject to Darwinian processes that push populations towards higher average fitness. Moreover, whether the systems persist or reproduce themselves depends on the success of the optimization. So perhaps valuing subjects must be subject to *existential* optimization – optimization with skin in the game.

We don't claim that all valuing subjects (or all living things) are conscious. Perhaps plants are not. For consciousness, we further need the existential optimization to be *internalized* and the evaluative states employed therein to be *integrated*.

Once you're sensitive to different ways the world can be, and can evaluate whether they are good *for* you, you must still respond appropriately. Initially, you might only care about a few things: light or dark, or the concentration of one kind of molecule. An isolated response to some input may not be conscious. But once you're tracking several states at once, you will need to be able to integrate many different factors – the water at the oasis smells good, but you see a predator there. Your own internal state matters: how thirsty are you anyway? To determine a coherent next move (perhaps even as a bacterium with no specialized decision-making systems), you must integrate cues from multiple sources and weigh their relative importance (Klein & Barron, 2025).

The allostatic mechanisms of metabolic individuals are examples of just such internalized and integrated optimization processes: specifying a target range, tracking whether things are going well or poorly, and responding appropriately to that evaluation in the service of survival. Predictive error or free energy minimization, (only) *when* carried out by valuing subjects as part of assessing and acting for their own good, are further examples.

By contrast, a population under natural selection doesn't typically have these features, despite also being subject to existential optimization. Evolution is ultimately responsible for the emergence of such internal optimization mechanisms, but is not itself one of them; in natural selection, the optimization process is *external* to the thing being optimized.

In animals like ourselves, perceptual and interoceptive states are inherently associated with a valuation through the concurrent activity of modulatory neurotransmitter systems (cholinergic and dopaminergic) with widespread and diffuse projections, responsible for controlling attention and tracking reward. Thus, information about the valuation of current inputs is in principle widely available – and in just the way that contents of a "global workspace" responsible for consciousness are supposed to be.

This brings us to a nice convergence: between a conceptual analysis of what consciousness requires, functionally speaking, and suggestive empirical facts about how evaluation systems operate in creatures like ourselves.

The "conceptual analysis" we have in mind hinges on the idea that conscious states must be available to downstream systems in order to drive "System 2" processes implicated in intentional action, memory, learning, explicit reasoning, and so on. Any states *unavailable* to drive either behavior (such as verbal report) or downstream cognition would be states that the subject itself would have no meaningful access to – so it would be perverse to call them "conscious." Functionalist versions of GWT encapsulate this conceptual analysis. Perhaps surprisingly, in our way of thinking, the functional integration of any evaluative system like ours may turn out to be substrate-dependent, and so, a functionalist GWT need not imply any commitment to substrate-neutral multiple realizability (Cao, 2022). Rather, the complex intertwinings of biochemical interactions may even be necessary to achieve the highly integrated evaluative system we think consciousness requires.

Our proposal differs from the standard GWT in its emphasis on the *significance* of what is broadcast. Mere information isn't enough. What is broadcast must be evaluatively relevant for a valuing subject. An evaluation system can then serve the role that Seth highlights as crucial for consciousness: "multimodal survival-relevant integration" across diverse inputs for controlling whole-system behaviors.

Valuation is the necessary connection between life and consciousness, but the considerations developed here don't yet show that an artificial intelligence couldn't be conscious – only that an artificial intelligence, if conscious, must also be a valuing subject.

Acknowledgements. None.

Financial support. J.M. acknowledges funding from the Stanford Interdisciplinary Graduate Fellowship and the Center for Affective Science Graduate Fellowship.

Competing interests. The authors declare no competing interests.

References

Bechtel, W., & Bich, L. (2025). Rediscovering Bernard and Cannon: Restoring the broader vision of homeostasis eclipsed by the cyberneticists. *Philosophy of Science*, (3), 1–22.

Cao, R. (2022). Multiple realizability and the spirit of functionalism. *Synthese*, *200*(6), 506.

Klein, C., & Barron, A. (2025) Phenomenal Interface Theory: A model for basal consciousness. *Philosophical Transactions of the Royal Society B.* (forthcoming).

Rescorla, M. (2013). Against structuralist theories of computational implementation. *The British Journal for the Philosophy of Science*, *64*(4), 681–707.

Biological naturalism is compatible with substrate independence

Ted Chiang*

Writer, Bellevue, Washington, USA
info@janklow.com

*Corresponding author.

doi:10.1017/S0140525X25103762, e326

Abstract

Seth's arguments for biological naturalism are not actually arguments for substrate dependence. Most of the criteria he proposes – including embodiment, embeddedness, active inference, allostasis, and metabolism – could be satisfied by digital organisms existing within a sufficiently detailed simulation running on Turing computation.

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

Anil Seth's paper provides an excellent overview of the issues surrounding the question of whether artificial intelligence could be conscious, and I agree with him on many issues; in particular, I believe that consciousness is a property only of living systems. Where we disagree is that I think all the relevant criteria for living systems could be satisfied by software running on Turing computation.

Seth repeats the "nothing gets wet in a weather forecasting computer" assertion, but that is only true because weather forecasting simulations aren't intended to model the wetting of individual objects. If someone were to claim that no lift is produced in an aerodynamics simulation because the computer running the simulation stays firmly on the desk, we would say they were committing a category error. The relevant criterion is whether the digital plane wing experiences lift, and in a good aerodynamics simulation, it does.

I think it should be possible to create digital organisms with virtual bodies that they use to navigate virtual environments, and for those bodies to operate via internal processes that will cease if the organism cannot negotiate its environment well enough to find the resources needed to maintain them. Such organisms would satisfy the requirements for embodiment, embeddedness, active inference, allostasis, and metabolism. None of these criteria requires crossing the divide between entities inside the virtual environment and those outside it. The essence of embodiment is not whether an organism's body can bump against my physical body; what's important is that the organism's body can bump against objects in its own environment. Likewise, we can meaningfully speak of entities that must actively maintain their internal non-equilibrium states while existing within a virtual environment, even if those entities are not reducing the local entropy within the computer on which the virtual environment is running.

Seth notes that the distinction between a simulation and an implementation isn't always clear. Let's consider the scenario of a digital simulation so detailed that the interactions between individual atoms are modeled; this is arguably the limiting case of digital simulation possible on Turing computation, if we assume that modeling quantum phenomena would require a quantum computer. Simulating an entire biological organism and a surrounding environment this way would obviously be prohibitively expensive, but if we set that issue aside momentarily, we can ask the question: Does this constitute a simulation, or an implementation? The carbon atoms within this digital simulation are incapable of bonding with carbon atoms in the physical world, but I would nonetheless argue that this digital simulation could be considered a genuine implementation in a realm that is disjoint from our own. I believe an organism existing in such a simulation would be conscious, but I don't think it fits within Seth's taxonomy; the consciousness created could be called substrate dependent inasmuch as it is a product of carbon atoms, but it could also be called substrate independent because those carbon atoms are simulated via Turing computation.

Now let's consider the issue of computational cost, which is relevant for reasons beyond the practical. Broadly speaking, simulations are useful only when they have advantages over the system being simulated; it's cheaper to test the flight characteristics of a simulated plane than it is to construct an actual plane, while a weather simulation is useful if it runs faster than the Earth's atmosphere while remaining accurate enough to provide actionable predictions. The aforementioned scenario of an atomic-level simulation of an organism fails on these counts; it would be both more expensive and slower than the equivalent organism in the physical world. I believe a biological organism in such a digital simulation would be conscious, but if that is the only way to create a conscious digital entity, which would mean we lack a deep understanding of consciousness because we have failed to identify any useful abstractions that pertain to it. A good aerodynamics simulation is a proof that we have identified useful abstractions of aerodynamics, such as the Reynolds number and the Navier-Stokes equations, because a digital plane wing experiences lift without our needing to simulate every single air molecule.

If embodiment, embeddedness, allostasis, and active inference are useful principles when trying to understand consciousness, then those are opportunities for abstraction that would enable us to build a simulation that is faster and cheaper than an atomic-level one. For example, if we can properly understand the role that allostasis plays in consciousness, we should be able to implement a digital allostatic system that is vastly simpler than an atomic simulation, while still satisfying the requirements needed to create a genuinely conscious digital organism. This scenario might qualify as "theory-based computational" in Seth's taxonomy, but not in the way he currently describes the category, if I'm understanding him correctly. The theories involved have yet to be formulated, but would be built around the criteria that Seth lists; even though he says these criteria point to substrate dependence, they are what would enable a substrate-independent implementation of consciousness.

I don't believe any current artificial-life programs qualify as living systems, but I don't think it's impossible in principle for a computer program to qualify, and that would be the first step toward creating digital entities that are conscious. I fully agree with Seth regarding the ethical dangers of creating beings capable of suffering. However, I don't think we will accidentally create entities with the moral status of mammals. The arguments favoring biological naturalism also suggest that the development path for digital lifeforms would roughly follow that of evolution in the natural world, starting with non-conscious biological behaviors such as reflexes and progressing toward conscious behaviors. I think it will take significant effort to create entities with even the moral status of insects. While some would argue that we should treat insects better than we do, I don't think it would be unconscionable if we created digital organisms of comparable status. So if we decide it's worthwhile to investigate the possibility of digital consciousness, we would have plenty of time to avoid an ethical catastrophe.

Financial support. The author received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interests. None.

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

Conscious AI does not require computational functionalism, and consciousness may be non-biological, like flight

Ron Chrisley*

Centre for Cognitive Science, Centre for Consciousness Science, and Sussex AI, University of Sussex, Falmer, UK
ronc@sussex.ac.uk

*Corresponding author.

doi:10.1017/S0140525X25103798, e327

Abstract

Seth's argument relies on the unargued-for claim that if computational functionalism is false, conscious AI is impossible. I show why the claim is in need of support by way of an analogy. I use a different analogy, involving artificial flight, to show that even if human consciousness is ineliminably biological, this does not imply the impossibility of non-biological, computational AI.

There is much that Seth and I agree on (as might be expected, since we have been colleagues at Sussex since the time of his PhD work). For example, I too have advocated for a predictive ("expectation-based") processing approach to understanding consciousness (Chrisley & Parthemore, 2007a; Chrisley & Parthemore, 2007b). However, I suspect that (even his own, much more fully worked out version of) this approach does not have the implications he thinks it does. More confidently, I think his arguments in this area fail to establish their conclusions, even if some of those conclusions may turn out to be correct. So I am focusing here only on these argumentative differences.

First, Seth's negative thesis, that conscious AI is impossible. He argues against conscious AI in two steps: The first step is to argue that computational functionalism, defined as the claim that "computation is sufficient for consciousness" (p. 1), is false. His second step is to claim (five times, and without argument) that if computational functionalism is false, conscious AI is impossible (target article, sect. 3.2, para. 4; sect. 3.6, para. 2; sect. 3.9, para. 1; sect. 5.0, para. 1; sect 5.3, para. 3, footnote 38).

But (as argued at length in (Chrisley, 2008)) each of those five claims is incorrect, and so Seth's master argument fails. An analogy might help show why. Consider the thesis of "McLarenism," the thesis that owning a McLaren car is sufficient for winning a Formula One Grand Prix. Clearly, McLarenism is false. But its falsity does not imply that it is impossible for a McLaren car owner to win a Grand Prix! It also does not imply that acquiring a McLaren can't be a key step to take in winning a Grand Prix, nor does it imply that an explanation of why the winner of the 1968 Belgian Grand Prix did in fact win does not involve their ownership of a McLaren, etc.

Just as a McLaren winning the Grand Prix is possible despite the falsity of McLarenism, so also is conscious AI (involving computation construed as narrowly as you like: digital, Turing-like, etc.) possible despite the falsity of computational functionalism. Further, computation might be a necessary part of any scientific explanation of consciousness, even if computational functionalism (as defined by Seth) is false. Even further, the analogy shows that computation might be a key component in an artificial conscious system even if computation is not *necessary* for consciousness.

These are not just philosophical niceties – they are crucial for understanding the true potential for conscious AI. A popular approach to artificial consciousness takes the view that computations must be grounded in (supplemented with) real-world, robotic (non-biological) sensorimotor capacities to yield consciousness, thus *rejecting* computational functionalism (at least as defined by Seth). Right or wrong (and we don't know which it is!), and despite that rejection, no one sees such an approach as anything other than the pursuit of conscious AI.

What about Seth's positive thesis, that consciousness is a biological phenomenon?

Despite detailing the intimate way in which human consciousness is interleaved with, and constrained by, our biology, Seth fails to show that this is necessarily essential to all consciousness. Consider flight: For living fliers (e.g., birds), "the connections between properties of the biological substrate and [particular instances of flying] have rich potential explanatory value. In the limit, one can think of each and every [particular instance of bird flight] as integrating multimodal bodily and environmental signals in the service of coordinating allostatic regulation and survival-relevant action. [Birds are flying creatures] because they are living, flesh and blood 'beast machines'." And: "Flight in biological systems is a dynamical and (plausibly) substrate-dependent process." Do these facts imply that flight in general is best thought of as a biological phenomenon? Do they imply that artificial flight is impossible? Do they imply that artificial flight can only be achieved by building systems that closely emulate this biological complexity? Obviously not. Multiply-realizable aerodynamic principles can be abstracted from the biological context in which they first evolved and were exploited, and can instead be used to achieve artificial flight in non-living substrates. Pointing out that bird flight can only be truly and fully understood in the context of an evolved, self-maintaining system that metabolizes does not change that fact. Bird flight is a biological phenomenon, but flight in general is not. In fact, we only achieved artificial flight when we stopped trying too closely to emulate biological flight (compare Armer, 1962). *If* consciousness is necessarily biological, it's going to take a lot more work to demonstrate that than merely showing that it is a biological phenomenon when present in biological systems.

A final point: We'll do a better job at understanding what computers (yes, even of the conventional – digital, von Neumann, Turing machine – sort) can and cannot do once we have a proper understanding of what computation itself is. The very fact that we recognize other, non-conventional forms of computation (analog computation, neuromorphic computation, quantum computation, etc.) shows that computation in general should not be defined as "what Turing machines do"; the fact that we do so anyway shows that we are not in the possession of a unifying account (Smith, 1985; Smith, 1996; Smith 2002). The Turing machine theory reveals some properties of some kinds of computational systems, but it doesn't capture the phenomenon of computation as a whole (Chrisley, 1998). So, ultimately, we aren't (or shouldn't) be interested in questions like "can Turing machines be conscious?" What we're really interested in is whether the concepts and methods that allow us best to explain, design, predict, and control the real-world phenomenon of computation (which already includes a lot more than what impoverished Turing machine

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

theory captures) can also help us understand (natural) and build (artificial) minds – even conscious ones (Chrisley, 2000).

Acknowledgments. I thank Oliver Sharpe, Thomas Clark, Nigel Shardlow, and Uziel Awret for their helpful comments on a draft.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interests. None.

References

Armer, P. (1962). Attitudes toward intelligent machines. In R. Chrisley (Ed.) (2000), *Artificial intelligence: Critical concepts in cognitive science* (Vol. *IV*, pp. 325–342). Routledge.

Chrisley, R. (2008). Philosophical foundations of artificial consciousness. *Artificial Intelligence in Medicine*, *44*(2), 119–137. https://doi.org/10.1016/j.artmed.2008.07.011

Chrisley, R., & Parthemore, J. (2007b). Synthetic phenomenology: Exploiting embodiment to specify the non-conceptual content of visual experience. *Journal of Consciousness Studies*, *14*(7), 44–58.

Chrisley, R., & Parthemore, J. (2007a). Robotic specification of the non-conceptual content of visual experience. In *Proceedings of the AAAI Fall Symposium on Consciousness and Artificial Intelligence: Theoretical Foundations and Current Approaches* (pp. 36–42). AAAI Press. https://cdn.aaai.org/Symposia/Fall/2007/FS-07-01/FS07-01-007.pdf

Chrisley, R. (2000). Transparent computationalism. In M. Scheutz (Ed.), *New computationalism: Conceptus-Studien 14* (pp. 105–121). Academia Verlag.

Chrisley, R. (1998). What might dynamical intentionality be, if not computation? *Behavioral and Brain Sciences*, *21*(5), 634–635. https://doi.org/10.1017/S0140525X98281738

Smith, B. C. (2002). The foundations of computing. In M. Scheutz (Ed.), *Computationalism: New directions* (pp. 23–58). MIT Press.

Smith, B. C. (1996). *On the origin of objects*. MIT Press.

Smith, B. C. (1985). The limits of correctness. *ACM SIGCAS Computers and Society*, *14*(1–4), 18–26. https://doi.org/10.1145/379486.379512

Why Seth (and everyone else) ought to be an illusionist

Andy Clark*

University of Sussex, Brighton, UK
andy.clark@sussex.ac.uk
https://profiles.sussex.ac.uk/p493-andy-clark

*Corresponding author.

doi:10.1017/S0140525X25103609, e328

Abstract

The paper assumes a fundamental difference between the study of consciousness and the study of intelligence and cognition more widely construed. I question this assumption, drawing on work by the (badly named) "illusionists." Once we see past the illusions, the landscape looks different, although Seth's core conclusions (about the importance of living organisations) might still obtain.

Seth provides a wonderful guide to the complex landscape of debate concerning the scientific explanation of consciousness. But in so doing, he makes an assumption that I would like to call into question. The assumption is that there is a meaningful difference between the study of consciousness and the study of intelligence and cognition more widely construed.

Seth begins by distinguishing consciousness from "intelligence" and by glossing intelligence as being about "what a system can do," whereas consciousness is about what it feels like to be that system. At that point, he approvingly invokes Nagel's (1974) discussion of "what it is like-ness." By way of contrast, consider the position known (somewhat misleadingly in my view) as illusionism (Clark, 2019; Clark, Friston, & Wilkinson, 2019; Dennett, 2016; Frankish, 2012, 2019). It is a mistake, illusionists suggest, to think of conscious experience as involving the presence of special "intrinsic properties" where these (Dennett, 1998a) are features of experience that outrun, or that could outrun, our own "reactive dispositions" – our capacities to discriminate, respond, react, sort, sift, judge, describe, and so on.

The picture of experience as involving such intrinsic properties ("qualia") is, of course, at the heart of Chalmers's (1996) easy problem/hard problem distinction. Notice that if you mean something else by "qualia" (or by "phenomenology" or "conscious experience"), then you need have no quarrel with illusionism at all – for a nice analysis, see Frankish (2012). As Dennett (1998b) once wrote:

"I am not denying the reality of consciousness at all; I am just denying that consciousness, real consciousness, is like what they think it is like" (p. 139).

If the illusionist is right, then consciousness is not special. It is just a label for a large and complex set of reactive dispositions. There is no hard problem because there are no "classic qualia" – experiential features that might, in principle, be present, absent, or different without that difference showing up in some profile of responses and reactions. That profile (in the human case) includes our reports of puzzlement and of experiencing "ineffable qualitative properties," and anything else that is actually testable by scientific means (see Dennett, 1991, on "heterophenomenology").

Adopting the illusionist stance in this way rules out the total failure of functionalism. It is, however, consistent with the potential failure of what Seth labels "computational functionalism." It might, for example, turn out that to get some specific functional profile you need to build your large-scale system out of smaller systems, each of which has its own "interests" and "goals," echoing the nested structure of living organisations. Or it may be that certain other properties of living matter itself are vital. What matters, however, is to keep sight of the fact that the only way we can test such ideas is by looking for profiles of response and reaction and asking what substrates best support them.

The illusionist route resolves what seems to me to be a tension in Seth's own picture of what "consciousness" is. On the one hand, Seth seems to identify "consciousness" with classical appeals to "what it is likeness." But such appeals pick out something distinctive only on condition that the "what it is likeness" is not exhausted by a complex response profile of some kind (see Frankish, 2012). On the other hand, Seth later nudges us towards what he calls the "real problem of consciousness," which, as I understand it, is solved whenever an account manages to get to grips with enough standard data points (patterns of behaviour, reaction, and report) to deliver improved capacities to predict, control, and intervene. This is an illusionist-friendly picture of the scientific study of consciousness.

Of course, nothing in this area comes for free. To carry through the illusionist project requires solving Chalmers's (2018) "meta-

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

problem" – the problem of explaining why our own conscious experience appears (to us) as so puzzling. What are the tricks that our own minds are playing on us that push us towards seeing conscious experience as so very weird and puzzling as to be apparently resistant to all known means of cognitive scientific attack? We do not yet have a full and convincing solution to the meta-problem, but for some attempts, see Clark (2019) and Clark et al. (2019). In many ways, Dennett's (1991) work was really a full-blooded attempt to resolve the meta-problem too, by tracking the many mistakes we make when we think about experience as occurring in a kind of "Cartesian Theatre."

Taking the illusionist stance radically alters the discursive landscape. It rules out scenarios in which qualia might be present, absent, or different without any clear means of telling, and thus keeps the study of consciousness on a firm scientific footing – exactly as Seth's own "real problem" formulation suggests. The scientific study of consciousness, tackled in this way, targets a complex high-dimensional space of abilities and capacities. Within this space, there are many important divisions. For example, some of our human reactive patterns involve advanced meta-cognitive capacities while others – including many of those shared with non-human animals – do not. Given the complexity of this space, there will (as Dennett long argued) typically be no simple "fact of the matter" about the presence or absence of a given "qualitative" state. Instead, different animals will turn up as occupying different locations in a high-dimensional space, with many criss-crossing patterns of similarity and difference.

In sum, I suggest that Seth makes his project unnecessarily hard by endorsing a picture of consciousness which uneasily straddles the broadly behavioural (where that includes everything we could ever actually test or probe) and the elusive "what it is likeness" notion – a notion whose defining feature is nothing other than its built-in (zombie-breeding) resistance to the kinds of scientific approach that Seth ultimately champions. Even if deep properties of life turn out to be crucial, we should not set up the initial puzzle by identifying "what-it-is-like-ness" as a special kind of target, distinct from what systems do and can do.

Seth should join the (badly named) illusionists!

Financial support. This work was carried out with the support of European Research Council, Synergy Grant (XSCAPE) ERC-2020-SyG 951631.

Competing interests. None.

References

Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford University Press.

Chalmers, D. J. (2018). The meta-problem of consciousness. *Journal of Consciousness Studies*, *25*(9–10), 6–61.

Clark, A., Friston, K., & Wilkinson, S. (2019). Bayesing qualia: Consciousness as inference, not raw datum. *Journal of Consciousness Studies*, *26*(9–10), 19–33.

Clark, A. (2019). Consciousness as generative entanglement. *Journal of Philosophy*, *116*(12), 645–662. https://doi.org/10.5840/jphil20191161241

Dennett, D. C. (1991). *Consciousness explained*. Little, Brown and Company.

Dennett, D. C. (1998a). Instead of qualia. In D. Dennett (ed.), *Brainchildren: Essays on designing minds* (Chapter 8, pp. 141–152). MIT Press.

Dennett, D. C. (1998b). Real consciousness. In D. Dennett (ed.), *Brainchildren: Essays on designing minds* (Chapter 7, pp. 131–141). MIT Press.

Dennett, D. C. (2016). Illusionism as the obvious default theory of consciousness. *Journal of Consciousness Studies*, *23*(11–12), 65–72.

Frankish, K. (2012). Quining diet qualia. *Consciousness and Cognition*, *21*(2), 667–676. https://doi.org/10.1016/j.concog.2011.04.001

Frankish, K. (2019). The meta-problem is the problem of consciousness. *Journal of Consciousness Studies*, *26*(9–10), 83–94.

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, *83*(4), 435–450. https://doi.org/10.2307/2183914

Meatless minds vs. beast machines: Are simulated subjects impossible?

Krzysztof Dołęga[a] and Axel Cleeremans[b]*

[a]Ruhr-Universität Bochum, Bochum, Germany and [b]Faculté des Sciences Psychologiques et de l'Education, Université libre de Bruxelles, Bruxelles, Belgium

krzysztof.dolega@rub.de
axcleer@ulb.ac.be
https://www.krysdolega.xyz
https://axc.ulb.be

*Corresponding author.

doi:10.1017/S0140525X25103889, e329

Abstract

We challenge Seth's claim that simulations of consciousness necessarily lack the causal powers or intrinsic properties of real consciousness. We argue that Seth's position risks begging the question against computational functionalism and faces a dilemma over the reducibility of qualia. This weakens his case, leaving open the possibility that simulations could, in principle, realize consciousness.

In his article, Seth distinguishes between simulation, implementation, and realization. *Simulation* is defined as a way of approximating some process or the workings of a mechanism "by abstracting it as a (Turing) computation and then implementing this computation on a digital computer" (p. 13). *Implementation* is understood as the way in which some computational function is executed in a physical medium. Importantly, Seth sees these concepts as continuous rather than completely distinct, since "by allowing for the inclusion of finer-grained, causal, and potentially non-computational properties" (ibid.), simulations can come closer to implementations. However, simulations can never become equivalent to *realizations* of their target phenomena (i.e., instantiate them) because they "generally lack the *causal powers* and *intrinsic properties* of the things being simulated" (ibid., emphasis added). In our eyes, this is the crucial step in Seth's rejection of any arguments for computational functionalism (CF) based on the possibility of simulating consciousness: simulation, no matter how fine-grained, is simply not the kind of thing that could *ever* realize conscious states.

We want to put pressure on Seth's claim that simulations of consciousness must necessarily lack the causal powers or the intrinsic properties of the real thing. Further, as we will point out, Seth's definition of simulation inevitably leads to a dilemma that undermines his position. Crucially, and contra Seth's claims (footnote 22), our argument *does not* beg the question against the deniers of CF, because at this point we simply do not know whether

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

CF is true or not. Even though we cannot demonstrate that simulating consciousness is indeed possible – after all, this is an empirical matter – we do think that Seth fails to conclusively show that it is impossible.

Starting with causal powers, we are happy to concede that a simulated rainstorm does not have the same causal powers as a real one, as it does not produce or necessitate the same effects. However, while a simulated storm cannot sink a real ship out in the seas, it can affect simulated ships (as in *SimCity*). What about consciousness? Assuming that consciousness is not an epiphenomenon, it is relatively uncontroversial to posit that its causal effects involve overt changes in a subject's behavior as well as covert changes in their cognitive states. We assume that the idea of simulating behavior is uncontroversial, and Seth does not seem to rule out the possibility of simulating cognitive states, since admitting this possibility is not the same as assuming computational functionalism *about* consciousness. Most cognitive scientists would agree that it is, in principle, possible to simulate various cognitive functions. Indeed, many consider existing programs and machine learning systems to already do this for some specific, if narrowly defined, functions (e.g., memory, analogical reasoning, etc.). In fact, recent evidence shows that simulations of cognition, such as large language models, can have a greater effect on people's rationality and cognitive states than other humans (Costello et al., 2024) – is this not a clear example of a simulation exerting causal powers in reality? All this leads us to wonder, why couldn't it be possible to create simulated agents whose conscious states would have the right kind of causal powers with respect to their (simulated) cognitive states and behaviors?

Seth will likely respond by pointing out that any physicalist should assume that consciousness somehow depends on or supervenes on brain states. We concur but want to make two observations. First, Seth cannot (and does not; see section 3.3) outright assume that the causal powers of conscious states require biophysical determinants without begging the question against CF. Second, while we do agree that biological cognition might depend on a complex tangle of biophysical processes, we are skeptical about whether realizing consciousness would require simulations of neural tissues all the way down to the molecular level. After all, the same requirement doesn't seem to hold for simulating cognitive functions. This being said, our skepticism on this matter is only warranted if consciousness is not fundamentally different from cognition. Many researchers will point out that, since at least some cognitive functions can operate unconsciously, there must be some kind of extra ingredient that makes consciousness special. This extra "secret sauce" is usually identified with the intrinsic properties of consciousness.

The intrinsic properties of consciousness, i.e., the properties that determine or constitute what experiences are like to the subject, are traditionally named *qualia* or *phenomenal* properties. They are said to be *intrinsic* because they seem to escape attempts at explanations in terms of non-intrinsic, i.e., relational or dispositional, properties, which in turn leads straight into the hard problem of consciousness (Chalmers, 1995). Philosophers and scientists responding to this problem tend to choose one of two general strategies. The first consists in arguing against the apparent irreducibility of such properties, either by identifying them with non-intrinsic ones (Dretske, 1995; Lycan, 1996; Rosenthal, 2005) or by denying their existence altogether (Dennett, 1991; Frankish, 2016). The second strategy consists of embracing the idea that phenomenal properties are intrinsic and non-reducible, for instance by adopting some version of panpsychism (Seager, 2006). Both strategies come with commitments that confront Seth with a dilemma. On the one hand, if the intrinsic properties of consciousness are reducible or identical to other, more familiar kinds of properties or processes, as Seth seems to suggest in section 5.5, then he should recognize that there is nothing fundamentally problematic about realizing phenomenal properties by simulating their underlying processes. On the other hand, Seth could side with Searle in claiming that the intrinsic properties of consciousness are irreducible to other kinds of properties and hence constitute a genuinely novel ontological kind. However, this option comes at a very high cost. Not only does it, once again, come close to begging the question against CF, but it would also open Seth to many well-known objections to Searle's position (see e.g., Kim, 1995; Warfield, 1999). Most problematically, however, asserting the intrinsic and non-reducible nature of qualia would severely undermine the possibility of a truly naturalistic, or indeed any scientific, explanation of consciousness, which we trust is not a direction in which Seth wants the field to go.

Acknowledgments. None.

Financial support. Axel Cleeremans is a research director with the F.R.S.-FNRS (Belgium) and a fellow of the Canadian Institute for Advanced Research (Brain, Mind, and Consciousness program). This work was supported by European Research Council Advanced Grant #101055060 "EXPERIENCE" to Axel Cleeremans.

Competing interests. The authors declare no competing interests.

References

Chalmers, D. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, *2*(3), 200–219.

Costello, T.H., Pennycook, G. & Rand, D.G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science*, *385*(eadq1814), 1–12. https://doi.org/10.1126/science.adq1814.

Dennett, D.C. (1991). *Consciousness explained*. Allen Lane.

Dretske, F. (1995). *Naturalizing the Mind*. Bradford Books/MIT Press.

Frankish, K. (2016). Illusionism as a Theory of Consciousness. *Journal of Consciousness Studies*, *23*(11–12), 11–39.

Kim, J. (1995). Mental causation in Searle's "Biological Naturalism". *Philosophy and Phenomenological Research*, *55*(1), 189–194.

Lycan, W.G. (1996). *Consciousness and Experience*. Bradford Books/MIT Press.

Rosenthal, D. (2005). *Consciousness and Mind*. Oxford University Press.

Seager, W. (2006). The intrinsic nature argument for panpsychism. *Journal of Consciousness Studies*, *13*(10–11), 129–145.

Warfield, T. A. (1999). Searle's causal powers. *Analysis*, *59*(1), 29–32.

Vitalism in the age of AI

Felipe De Brigard*

Department of Philosophy, Duke University, Durham, NC, USA
felipe.debrigard@duke.edu
www.imclab.org

*Corresponding author.

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

doi:10.1017/S0140525X25103890, e330

Abstract

Andy Clark's notion of microfunctionalism (1989) offers a biologically grounded form of functionalism, narrowing the scope of "multiple realizability." While computation is substrate-independent in principle, only select systems – such as brains – can implement its computational complexity. Extending this view, I challenge Seth's skepticism of computationalism and conscious AI, emphasizing the difference between possible and actual substrate independence. Current digital architectures fall short, but future systems might instantiate conscious states if built on appropriate computational substrates.

When functionalism entered analytic philosophy of mind, it was seen as a way to avoid substance dualism without endorsing behaviorism – which attributed mentality too liberally – or identity theory – which chauvinistically denied it too narrowly. The notion of computation, formalized in the 1930s, gave early functionalist philosophers (e.g., Armstrong, 1968; Fodor, 1968; Putnam, 1967) a vocabulary to identify the mental with a property that could fit a monist worldview. Soon after, Block (1978) argued that "machine state functionalists," for whom mental states were to be identified with states of a Turing-machine table, risked over-attributing mentality to systems like the Bolivian economy or a population slice of China equipped with radios.

Block's challenge can be read in two ways. On one hand, if neurons are the relevant computational units and their computations can be described in a Turing machine table, then one must attribute mentality to systems that intuitively lack it. On the other hand, if neurons are the relevant units, their computations likely cannot be captured by such a Turing-machine table implementable by people and radios. This latter view inspired Andy Clark's (1989) "microfunctionalism," a biologically plausible version of functionalism meant to address doubts about machine-table models of the brain.

Microfunctionalism holds that computations are in principle multiply realizable, but in practice very few things – neurons, glia, etc. – can instantiate those constituting our mental states. Computation remains substrate independent, yet computational complexity rules out most materials. This isn't new: computer engineers abandoned vacuum tubes for transistors, then integrated circuits, then silicon microprocessors, as earlier materials proved too slow, fragile, or costly to run ever more complex computations. But from the fact that certain computations cannot be carried out by certain materials, it does not follow that computational processes aren't substrate independent. As Clark noted, we must attend to the modal scope of the multiple realizability claim.

The same ambiguity in modal scope is present in the characterization of "substrate independence" Seth uses to argue against the possibility of consciousness in AI systems. According to Seth, substrate independence is "the claim that the same kind of mental states (including conscious states) *can* occur in systems with different physical substrates – i.e., systems made out of different kinds of stuff" (my italics). But to say that a mental state *can* occur in a different physical substrate does not mean that any physical substrate would do. Again, the more complex the computation, the fewer actual physical substrates can likely instantiate it. Thus, the fact that current conventional digital AI systems implemented in existing computational architectures do not instantiate conscious states does not speak against the in-principle claim that, given the right substrate, there could one day be a physical AI system that could successfully implement them.

We do not know how consciousness arises in biology, but if it depends on complex autopoietic processes absent from current digital AI systems, then I agree that current networks cannot be conscious. From the start, neural networks failed to capture brain processes likely crucial to mental life. Most are based on oversimplified neuronal models – known to be incomplete or false, at least since Bullock (1959) challenged the neuron doctrine behind McCulloch and Pitts's (1943) architecture. Today, many of these very same processes do not feature in any way in the computational structure of our best AI systems. Yet, that does not mean that they cannot be characterized computationally, nor that they are not computational. And if one day we can characterize them computationally – or, at least, replicate them with artificial materials even if we don't know how to characterize them computationally – then the possibility of AI systems that can instantiate consciousness may not sound so far-fetched.

The brain is profoundly complex, and Seth is exactly right in claiming that its generative entrenchment "imposes many constraints on possible alternative implementations, which likely exclude alternate substitutes such as silicon." But this claim is entirely compatible with a microfunctionalist view, according to which as the complexity of the computation increases, the number of possible substrata that can instantiate it shrinks. Indeed, it may be possible to accept Seth's claim that, at the end, it may be really hard – if not impossible – to separate "what brains do from what they are," and yet still accept that what they do is, in fact, compute. For here the difficulty, even the impossibility, of separating them has simply to do with the extraordinarily reduced number of things in the universe that can possibly instantiate said computations. As Dennett (2005, 21) once put it, as we learn more about brains, "minds will turn out not to be simple computers, and their computational resources will be seen to reach down into the subcellular molecular resources available only to organic brains, but the theories that emerge will still be functionalist in the broad sense" (p. 25).

Whatever computations our brains may be doing, they are the result of years and years of cellular evolution that has benefited from an ever-growing sequence of improbable tricks, all of which require physical substrata with very precise timings, durability, trophic cycles, energy consumption, etc. Yet, nothing in this improbability implies that what brains do is not *computing* – broadly understood, of course, so as to include non-digital computations. Separating what brains do from what they are may be hard, if not impossible, given the contingent universe we live in, but if what ultimately explains why brains can instantiate conscious states is something to do with what they *are* and not with what they *do* – as Seth's biological naturalism implies – we risk falling into a kind of vitalism (not far from Bergson's!) whereby all we could say is that brains are conscious because of some sort of *élan consciente*. And as long as we have reason to believe that there is a way to characterize what brains do in terms of computations, we won't need to reject the possibility that, maybe one day, complex AI systems will be able to instantiate conscious states.

Acknowledgments. N/A.

Financial support. N/A.

Competing interests. No competing interests to declare.

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

References

Armstrong, D. M. (1968). *A materialist theory of the mind*. Routledge & Kegan Paul.

Block, N. (1978). Troubles with functionalism. In C. W. Savage (Ed.), *Minnesota studies in the philosophy of science: Vol. 9. Perception and cognition* (pp. 261–325). University of Minnesota Press.

Bullock, T. H. (1959). Neuron doctrine and electrophysiology. *Science*, *129*(3356), 997–1002. https://doi.org/10.1126/science.129.3356.997

Clark, A. (1989). *Microcognition: Philosophy, cognitive science, and parallel distributed processing*. MIT Press.

Dennett, D.C. (2005). *Sweet dreams: Philosophical obstacles to a science of consciousness*. MIT Press.

Fodor, J. A. (1968). *Psychological explanation: An introduction to the philosophy of psychology*. Random House.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115–133.

Putnam, H. (1967). The nature of mental states. In N. Block (Ed.), *Readings in philosophy of psychology* (Vol. *1*, pp. 223–231). Harvard University Press.

Why I am not a biological naturalist

Leonard Dung*

Ruhr-Universität Bochum, Germany
leonard.dung@rub.de
https://sites.google.com/view/leonard-dung/home

*Corresponding author.

doi:10.1017/S0140525X25103440, e331

Abstract

I make three claims: First, denying biological naturalism does not logically require computational functionalism. Second, while Seth's arguments establish biological naturalism as a view worth taking seriously, they fail to make it more plausible than the view that AI can be conscious. Third, there are independent arguments suggesting the overall more plausible view is that AI can be conscious.

In his excellent target article, Seth reviews a wide range of evidence with respect to the possibility of phenomenal consciousness in artificial intelligence. The two main rivals in his discussion are computational functionalism – the view that implementing the right kinds of computations is sufficient for consciousness – and biological naturalism, according to which only living systems can be conscious. He concludes that there are "good reasons" to hold that biological naturalism is true and that (silicon-based) AI cannot be conscious. I will put forward three considerations that jointly favor the conclusion that AI consciousness is possible, despite Seth's arguments.

First, AI systems (made from silicon) may be conscious even if computational functionalism is false. Seth seems to go back and forth on this, saying: "computational functionalism is sufficient for substrate flexibility. But it is not necessary." But also (incorrectly): "Conscious AI requires [. . .] computational functionalism to hold." While it is important to keep in mind, Seth could reasonably reply that most of his arguments target the claim that silicon-based systems can be conscious directly, rather than engaging with computational functionalism specifically. However, it is also important to note that biological naturalism is a stronger claim than that AI consciousness is impossible: It also rules out consciousness in creatures, for example, hypothetical extraterrestrials, which are not made from silicon but which also do not qualify as living creatures. Therefore, even if we reject computational functionalism, we should be careful not to jump to the conclusion that biological naturalism is correct.

Second, Seth reviews and develops multiple lines of argument against the possibility of conscious AI. He points to processes in human brains which cannot be replicated in AI systems, such as mortal computation (Hinton, 2022), neural computation (Piccinini, 2021), biological computation (Kirkpatrick, 2022), analog computation, and neuromorphic computing. Moreover, AI systems have informational closure, a distinction between software and hardware, and they can be characterized computationally independent of the actual dynamics of the implementing system, and so on – all properties potentially absent from living brains. In addition, Seth endorses a predictive processing (PP) account of conscious perception and characterizes PP mechanisms as driven by fundamentally biological imperatives related to allostasis and ultimately grounded in the free energy principle.

I do believe that these arguments establish biological naturalism as a coherent alternative to computational functionalism. However, in my view, they fail to provide a reason why we should believe biological naturalism over computational functionalism.

Seth's different arguments seem to share the same form:

Biological naturalism argument

P1: There is some feature F (implementing mortal computations, neural computations, etc.) that only living systems can implement.

P2: F is necessary for consciousness.

C: Therefore, only living systems can be conscious.

Seth's arguments support P1. However, it is not clear why one should believe P2. From my reading, Seth does not provide reasons for P2. Thus, computational functionalists can counter with the following argument:

AI consciousness argument

P1: There is some feature G (implementing certain standard digital computations, that is, non-mortal computations, non-neural computations, etc.) that AI systems can implement.

P2: G is sufficient for consciousness.

C: Therefore, AI systems can be conscious.

Until someone gives us a reason to accept P2 of one of these arguments, we are at a stalemate between both views. I also do not see that Seth's recourse to PP provides such an argument. Setting aside that PP may not be best thought of as a theory of consciousness (Schlicht & Dolega, 2021), it can be construed algorithmically – so that AI systems can implement PP – or biologically, and the empirical evidence for PP does not distinguish between these options. If we do not know whether biological naturalism is true, we cannot know whether implementing the high-level algorithms of PP is sufficient for consciousness or whether biology is necessary. Thus, appeals to PP cannot help us assess whether biological naturalism is true. So far, biological naturalism and the view that AI can be conscious are on equal footing.

However, third, it seems to me that there are multiple reasons in favor of P2 of the AI consciousness argument that at least have some evidential weight: 1. Attributing consciousness to AI systems which are, on some abstract level, computationally similar to

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

conscious humans and animals leads to more unifying explanations of behavior across creatures (see Dung n.d.). 2. Individuating the mechanism underlying consciousness computationally does justice to its evolutionary function (see Goldstein and Kirk-Giannini n.d.). 3. It is plausible that most ordinary people would, with respect to some possible AI systems, have the strong sense that they are conscious, while consciousness science often appeals to folk intuitions (see Shevlin n.d.). 4. In addition, if one believes that the question of AI consciousness is sensitive to how we conceptually engineer (Cappelen, 2018) consciousness, then it seems most advisable to engineer consciousness to refer to high-level functional properties shared with AI rather than biological ones, since the former are the ones that matter most in ethics, in explanation of behavior, and in everyday life (for relevant illusionist perspectives, see Dung, 2022, sect. 6; Frankish, 2024; see also Cappelen & Dever, 2021). 5. Finally, biological naturalism specifically is subject to the charge that restricting consciousness to biological systems requires positing that it is an implausible coincidence that humans are conscious (see Schwitzgebel & Pober, 2024).

None of these arguments are, of course, conclusive, and space does not permit it to develop them in detail (see also the argument against AI consciousness in Schwitzgebel & Pober, 2024; as well as relevant discussions in Dung & Kersten, 2024; Shiller, 2024; Mogensen, 2024). However, I wish to emphasize a more general point here: Various promising routes to support P2 of the AI consciousness argument are available which, combined, have substantial evidential weight in favor of the view that AI can be conscious. However, I am not aware of any strong arguments for P2 of the biological naturalism argument. Accordingly, until such arguments are provided, we should conclude that it is more likely that AI consciousness is possible than that it is not possible.

Acknowledgements. I thank Christian de Weerd and Kayla Carnation for very valuable comments.

Financial support. No funding to declare.

Competing interests. No competing interests to declare.

References

Cappelen, H. (2018). *Fixing language*. Oxford University Press. https://doi.org/10.1093/oso/9780198814719.001.0001

Cappelen, H., & Dever, J. (2021). *Making AI intelligible: Philosophical foundations*. Oxford University. https://doi.org/10.1093/oso/9780192894724.001.0001

Dung, L. (2022). Does illusionism imply skepticism of animal consciousness?. *Synthese*, *200*(3), 238. https://doi.org/10.1007/s11229-022-03710-1

Dung, L. (n.d.). Consciousness without biology: An argument from anticipating scientific progress. https://philpapers.org/rec/DUNCWB

Dung, L., & Kersten, L. (2024). Implementing artificial consciousness. *Mind and Language*, *40*(1), 1–21.

Frankish, K. (2024). The ethical implications of illusionism. *Neuroethics*, *17*(2), 28. https://doi.org/10.1007/s12152-024-09562-5

Goldstein, S., & Kirk-Giannini, C. D. (n.d.). A case for AI consciousness: Language agents and global workspace theory. https://philpapers.org/rec/GOLACF-2

Hinton, G. (2022). The forward-forward algorithm: Some preliminary investigations. arXiv. https://doi.org/10.48550/arXiv.2212.13345

Kirkpatrick, K. L. (2022). Biological computation: Hearts and flytraps. *Journal of Biological Physics*, *48*(1), 55–78. https://doi.org/10.1007/s10867-021-09590-9

Mogensen, A. L. (2024). How to resist the Fading Qualia Argument. https://globalprioritiesinstitute.org/how-to-resist-the-fading-qualia-argument-mogensen/. Accessed 3 March 2025

Piccinini, G. (2021). *Neurocognitive mechanisms: Explaining biological cognition*. Oxford University Press.

Schlicht, T., & Dolega, K. (2021). You can't always get what you want: Predictive processing and consciousness. *Philosophy and the Mind Sciences*, *2*. https://doi.org/10.33735/phimisci.2021.80

Schwitzgebel, E., & Pober, J. (2024). The Copernican argument for alien consciousness; the mimicry argument against robot consciousness. arXiv. https://doi.org/10.48550/arXiv.2412.00008

Shevlin, H. (n.d.). Consciousness, Machines, and Moral Status.

Shiller, D. (2024). Functionalism, integrity, and digital consciousness. *Synthese*, *203*(2), 47. https://doi.org/10.1007/s11229-023-04473-z

Is conscious AI desirable?

Kathinka Evers[a] and Michele Farisco[a,b]*

[a]Centre for Research Ethics and Bioethics, Uppsala University, Uppsala, Sweden and [b]Biogem Molecular Biology and Genetics Research Institute, Ariano Irpino (AV), Italy
kathinka.evers@uu.se
michele.farisco@crb.uu.se

*Corresponding author.

doi:10.1017/S0140525X25103488, e332

Abstract

In addition to arguments for scepticism concerning the technical feasibility of conscious AI, there are grounds for scepticism concerning its desirability and safety. Conscious AI created by human beings could, especially if it reflects human nature, be extremely dangerous. It could also be very different from human consciousness and thus (a) undetectable, (b) incomprehensible, and (c) incommunicable. The latter consideration could increase the risks of the former and cut both ways, if the conscious AI has valenced experience.

In the section of his paper dedicated to *What should we (not) do*, Seth writes that "The development of real artificial consciousness for its own sake should not be an explicit goal." This touches upon the fundamental question why to pursue artificial consciousness in the first place. What are the expected benefits deriving from a conscious AI, and do they outweigh the possible risks?

Developing conscious AI could advance our understanding of biological consciousness, consistently with the idea that it is easier to understand what we ourselves build. It might also exponentially accelerate the entire field (LeCun, 2022), maximising its benefits for society. For instance, a conscious AI would be able to develop models of the world and to intentionally use them in order to anticipate the impact of its own functioning and of other agents' actions. On this basis, on an optimistic reading, it could avoid negative consequences and promote positive ones, including a better alignment with human values (Khamassi, Nahon, & Chatila, 2024).

Seth acknowledges these arguments in support of conscious AI, but he also (rightly, in our opinion) add that it should not be taken for granted that conscious AI is a holy grail.

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

We agree, not only for the reason that it may not actually solve the present technical shortcomings of AI, but more fundamentally because a conscious AI may eventually replicate and even exacerbate the shortfalls of human consciousness-cum-intelligence which is either explicitly or implicitly assumed as a reference model in the effort to build a conscious AI (Farisco et al., 2024). The combination of intelligence and consciousness does not entail benevolence. To the contrary, the combination has in human beings, as a result of cultural as well as biological evolution, typically been coupled with aggression against "out-group" individuals, reserving sympathy and aid for those who belong to the group (Evers, 2015). Moreover, humans create a very large number of group distinctions exhibiting profound diversities in terms of the values they endorse. Therefore, the question of how to ascertain that intelligent and conscious artificial systems ("they") created by humans do not react with malevolence but with benevolence towards (some) humans ("us") is important. Artificial systems are trained and programmed with and have access to vast amounts of human-generated data, where examples of, for example, xenophobic, misogynistic, and violent behaviour abound and universal benevolence shines by its absence. For this reason, the confidence that a conscious machine created by humans would be engaged in universal human well-being appears at best unfounded.

Why should "they" (machines) care about "us" (humans)? And if they were to engage, why should they feel benevolence instead of malevolence towards out-group individuals (that would be far more in line with the mind of their programmer, their human "creator")? In view of how selectively and xenophobically conscious intelligence operates in humans, taking machine benevolence towards humans for granted appears dangerously naïve.

We may also reverse the perspective: hypothetical artificial consciousness, specifically sentience, introduces the issue of machine welfare. Seeing how humans treat other human and non-human animals, there is ample reason to doubt that machines would face a happy destiny if we, whether intentionally or inadvertently, were to construct machines capable of subjective experience. And, as in the case of, for example, vivisected animals, or enslaved and intellectually stunted humans, their suffering might long go unacknowledged and even undetected.

We think that the above suffices to provide grounds for scepticism concerning the desirability of artificial consciousness, in addition to the grounds for scepticism concerning its feasibility developed in detail by Seth in his paper.

There are also epistemic reasons for scepticism concerning the feasibility of artificial consciousness. Presently, consciousness is only known to exist in living things. That is a fact about our knowledge that does not logically exclude artificial consciousness. Conscious AI is assumed to be theoretically possible within certain theoretical frameworks (Evers et al., 2025), but for now, no external empirical evidence or validation is available. In that situation, we can neither logically exclude nor affirm the possible existence of artificial consciousness in the real world. Empirical validation is essential to turn logical possibility into actuality.

Moreover, artificial consciousness would likely be qualitatively very different from human and animal consciousness due to deep differences in substrate, architecture, and functions (Evers, 2024; Farisco, Evers, & Changeux, 2024). Here, the epistemic and psychological challenges introduced by the gaming problem are central (Birch & Andrews, 2024). When artificial systems use human-generated training data to mimic human behaviours, they may, if successful, psychologically persuade human users of their sentience, as a form of anthropomorphism; yet logically, mimicking human behaviours in artificial systems has no evidential value whatsoever. This poses a formidable obstacle to detecting sentience or consciousness in artificial systems trained with human-generated data and deeply influences our risk–benefit assessments in this domain. To avoid the gaming problem, we should arguably focus less exclusively on similarities and pay more attention to differences, since AI is materially, functionally and architecturally very different from a human brain. But here a different problem arises: the question of commensurability. A challenge in detecting differences is that we cannot think entirely beyond our own perspective, since we are imprisoned by the limits imposed by our bodies, so if the differences are sufficiently deep, we cannot detect them. There must also be some similarities to justify the application of the same concept to such distinct phenomena.

Our conclusion is that engineering conscious AI is not desirable because the potential risks far outweigh the possible benefits. In view of how human nature has expressed itself throughout history, conscious AI created by humans would likely be monumentally dangerous. It could also be very different from human consciousness and thus (a) undetectable, (b) incomprehensible, and (c) incommunicable. This could increase the risks and cut both ways if the conscious AI has valenced experience.

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

Competing interests. None.

References

Birch, J., & Andrews, K. (2024). To understand AI sentience, first understand it in animals. *Intellectica*, *81*(2), 213–226.

Evers, K. (2015). Can we be epigenetically proactive?. In T. Metzinger, & J. Windt (Eds.), *Open mind: Philosophy and the mind sciences in the 21st century* (pp. 497–518). MIT Press.

Evers, K. (2024). To understand sentience in AI first understand it in animals. Commentary to Jonathan Birch and Kristin Andrews. *Intellectica*, *81*, 229–232.

Evers, K., Farisco, M., Chatila, R., Earp, B. D., Freire, I. T., Hamker, F., Nemeth, E., Verschure, P. F. M. J., & Khamassi, M. (2025). Preliminaries to artificial consciousness: A multidimensional heuristic approach. *Physics of Life Reviews*, *52*, 180–193.

Farisco, M., Baldassarre, G., Cartoni, E., Leach, A., Petrovici, M. A., Rosemann, A., … & van Albada, S. J. (2024). A method for the ethical analysis of brain-inspired AI. *Artificial Intelligence Review*, *57*(6), 133. https://doi.org/10.1007/s10462-024-10769-4

Farisco, M., Evers, K., & Changeux, J.-P. (2024). Is artificial consciousness achievable? Lessons from the human brain. *Neural Networks*, *180*, 106714. https://doi.org/10.1016/j.neunet.2024.106714

Khamassi, M., Nahon, M., & Chatila, R. (2024). Strong and weak alignment of large language models with human values. *Scientific Reports*, *14*(1), 19399. https://doi.org/10.1038/s41598-024-70031-3

LeCun, Y. (2022). A path towards autonomous machine intelligence. Retrieved from https://www.infinityexplorers.com/wp-content/uploads/2022/10/a_path_towards_autonomous_mach.pdf

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

Sentience, emergence in neurobiological systems, and the prospects for artificial consciousness

Todd E. Feinberg[a]* and Jon Mallatt[b]

[a]Mount Sinai Behavioral Health Center, Icahn School of Medicine at Mount Sinai, New York, NY, USA and [b]University of Idaho, Moscow, ID, USA
todd.feinberg@mountsinai.org
jmallatt@uidaho.edu

*Corresponding author.

doi:10.1017/S0140525X25103555, e333

Abstract

In this commentary, we focus on three interrelated issues of particular importance for determining the possibility of "artificial consciousness": 1) the distinctions between *sentience* and *consciousness,* 2) why the *physical substrate* of a complex system is critical for the emergence of sentience, and 3) how Emergence in Systems Theory helps explain consciousness in animals and whether "artificial consciousness" could exist.

At the outset, when considering the possibility of "artificial consciousness," it is essential to be clear about what aspects of consciousness we are considering. We agree with Seth's statement (Section 1.1) – based upon Nagel (1974) – that "for a conscious organism, there is something it is like to be that organism" and that if it "feels like" something to be a conscious system – there is a conscious experience happening – whereas it doesn't feel like anything to be an unconscious system – there is no conscious experience happening." Note that Nagel's emphasis is clearly on the "feeling" aspect of consciousness. Seth prefers the term "consciousness" over "sentience," because he suggests the latter is used inconsistently in the literature and is prone to confusion. "Sentience," however, has been used widely and consistently in animal-consciousness studies for the past decade (Birch, 2024; Chittka, 2022; Harnad, 2016). Thus, we prefer "sentience" over the broader term "consciousness" because its Latin root means "to feel," so it emphasizes consciousness as feeling, not just as aspects of cognition or computing. So, in our view, it is the "feeling" aspect of consciousness that is the key issue in the possibility of "conscious AI."

Second, Seth's argues – correctly in our view – that the neural processes behind consciousness are tied to a living nervous system and therefore are not substrate-independent as the Turing-computational functionalism behind standard AI demands – thereby indicating such AI cannot attain consciousness. Indeed, Seth provides a comprehensive list of the differences between standard silicon-based computer circuits and living neurons/brains, and these would most likely preclude consciousness in conventional AI systems (Section 3). These differences include the following: brains are embodied but Turing computers are disembodied; recurrent processing versus linear processing; extreme integration of information versus lesser integration (the generative entrenchment concept in Section 3.3); spike-wave synchronization is used by neurons alone; and the conscious brains alone sense the external environment through multiple sensory modalities for survival and propagation (Section 4.5). Seth identified what he argues is the "real problem of consciousness" and characterized it as a truly scientific challenge: "as we become better able to *explain, predict*, and *control* properties of consciousness by underlying mechanisms, the sense of mystery . . . will dissolve" as did the mystery of life in the past two centuries (Section 4.4).

Third, Seth mentions *hierarchical emergence* (Section 3.5) in the context of "substrate flexibility" in physics. Here, we expand upon this point with reference to consciousness because we – and many others – have argued that it is essential to view sentience – and more broadly consciousness – as an *emergent process of brain functions* (Broad, 1925; Chalmers, 2006; Feinberg, 2024; 2025; Feinberg & Mallatt, 2020, 2016; Kim, 1998, 2006; Searle, 1992; Thompson, 2007)

Emergence is a property of hierarchical systems in which lower levels interact with higher levels. In general terms, an emergent feature is a novel, aggregate feature of the *system* that is not present in the *parts* alone. Commonly cited biological examples of emergent processes are respiration, metabolism, and digestion as they function in the body of vertebrate animals (Salthe, 1985; Searle, 1992). Emergent properties are often described as "more than the sum of their parts" because the parts interact at multiple hierarchical levels and those interactions are the key to the emergence of *novel system features* (Ahl & Allen, 1996; Bedau & Humphreys, 2008; Clayton, 2006; Vintiadi, 2013). In the nervous system of vertebrates, for instance, consciousness emerged as a novel system feature when new higher levels evolved through the expansion of the forebrain. Therefore, like life, sentience is best understood as an *emergent feature of a complex system*, in this case, the living central nervous system.

An additional factor that links emergence, systems theory, and sentience is the feature of *complexity*. There is substantial theoretical support for the role that complexity plays in increasing emergent-system properties (Ladyman & Wiesner, 2020; Mitchell, 2009; Vintiadis, 2013), and the most complex systems in nature are all biological, from cells to multicellular organisms to animals with complex brains. Further, when we trace the evolution and neurobiology of sentience, there is a close correspondence between the *neurobiological complexity* of the nervous systems and the emergence of sentience. This finding is based upon the view that organisms that are sentient are those whose brains are documented to create higher-order mapped neural sensory representations of the world, possess the neural infrastructures that enable interoceptive-affective feelings with positive versus negative valence, and the capacity to hierarchically combine all these diverse neural processes into an integrated sentient experience. These conditions are only met by the complex-brained vertebrates, arthropods, and cephalopods like octopuses (Feinberg 2024, 2025; Feinberg, & Mallatt, 2020, 2016; Mallatt et al., 2021; Mallatt & Feinberg, 2021).

These considerations indicate that the *living biological substrate* and the neural parts of sentient brains – via the constraints imposed upon them by their biological, adaptive, and evolutionary nature – are essential to the emergent properties of sentience, an example of what Seth calls the *strong substrate-dependent*

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

hypothesis (Section 5.5). Consistent with the principles of emergence, sentience and more broadly consciousness depend upon the physical structure and processes of the *parts* in addition to the neuron-based, multi-leveled hierarchical interactions of these parts. Therefore, it is highly emergently substrate-dependent – in contrast to the processes of the conventional, substrate-independent, Turing-based AI systems.

Finally, today's computers – despite their advanced AI capabilities – are not capable of "feeling," and thus are neither sentient nor conscious. Seth hypothesizes that if someday artificially intelligent entities may be made more life-like, it is at least conceivable that they could have "conscious AI." While this possibility is something that cannot be definitively predicted, we feel it is quite likely that whatever "sentient-like" features (if any) these non-living, non-biological entities display, they will not be "experienced" in the same way that animal sentience is *personally experienced*. This is because computer non-biological substrates and their organizational properties will likely be so vastly different from living, biological, and neurobiological consciousness that is the result of billions of years of evolution, as sentience and consciousness gradually emerged from life.

Acknowledgments. None.

Financial support. No outside funding was used for this work.

Competing interests. The authors declare that they have no competing interests.

References

Ahl, V., & Allen, T. F. H. (1996). *Hierarchy theory*. Columbia University Press.
Bedau, M. A., & Humphreys, P. (Eds.). (2008). *Emergence: Contemporary readings in philosophy and science*. MIT press.
Birch, J. (2024). *The edge of sentience: Risk and precaution in humans, other animals, and AI*. Oxford University Press.
Broad, C. D. (1925). *The mind and its place in nature*. Routledge.
Chalmers, D. J. (2006). Strong and weak emergence. In P. Clayton, & P. Davies (Eds.), *The re-emergence of emergence*. (pp. 244–254). Oxford University Press.
Chittka, L. (2022). *The mind of a bee*. Princeton University Press.
Clayton, P. (2006). Conceptual foundations of emergence theory. In P. Clayton, & P. Davies (Eds.), *The re-emergence of emergence*. (pp. 1–31). Oxford University Press.
Feinberg, T. E. (2024). *From sensing to sentience: How feeling emerges from the brain*. MIT Press.
Feinberg, T. E. (2025). Neurobiological emergentism. Sentience as an emergent process and the experiential gap. *Frontiers in Psychology*, *16*, 1528982. https://doi.org/10.3389/fpsyg.2025.1528982
Feinberg, T. E., & Mallatt, J. M. (2016). *The ancient origins of consciousness: How the brain created experience*. MIT press.
Feinberg, T. E., & Mallatt, J. (2020). Phenomenal consciousness and emergence: Eliminating the explanatory gap. *Frontiers in Psychology*, *11*, 1041. https://doi.org/10.3389/fpsyg.2020.01041
Harnad, S. (2016). Animal sentience: The other-minds problem. *Animal Sentience*, *1*(1), 1.
Kim, J. (1998). *Mind in a physical world. An essay on the mind–body problem and mental causation*. MIT Press.
Kim J. (2006). Being realistic about emergence. In P. Clayton, & P. Davies (Eds.), *The re-emergence of emergence*. (pp. 190–202). Oxford University Press.
Ladyman, J., & Wiesner, K. (2020). *What is a complex system?*. Yale University Press
Mallatt, J., Blatt, M. R., Draguhn, A., Robinson, D. G., & Taiz, L. (2021). Debunking a myth: plant consciousness. *Protoplasma*, *258*(3), 459–476. https://doi.org/10.1007/s00709-020-01579-w
Mallatt, J., & Feinberg, T. E. (2021). Multiple routes to animal consciousness: constrained multiple realizability rather than modest identity theory. *Frontiers in Psychology*, *12*, 732336. https://doi.org/10.3389/fpsyg.2021.732336
Mitchell, M. (2009). *Complexity: A guided tour*. Oxford University Press.
Nagel, T. (1974). What is it like to be a bat?. *The Philosophical Review*, *83*, 435–450.
Salthe, S. N. (1985). *Evolving hierarchical systems: Their structure and representation*. Columbia University Press.
Searle, J. R. (1992). *The rediscovery of the mind*. MIT Press.
Thompson, E. (2007). *Mind in life: Biology, phenomenology and the sciences of mind*. Harvard University Press.
Vintiadis, E. (2013). Emergence. Internet Encyclopedia of Philosophy. https://iep.utm.edu/emergence/

The question is not "are they conscious?" but "what could they be conscious of?"

Chris Fields[a]* and James Glazebrook[b]

[a]Allen Discovery Center, Tufts University, Medford, MA, USA and [b]Department of Mathematics and Computer Science, Eastern Illinois University, Charleston, USA
fieldsres@gmail.com
jfglazebrook@eiu.edu

*Corresponding author.

doi:10.1017/S0140525X25103464, e334

Abstract

We argue that by focusing on an unanswerable ontological question, both computational functionalism and biological naturalism miss the empirical question of scientific and practical interest: what are systems of some kind *capable of* experiencing?

Anil Seth's critique of computational functionalism as a theory of consciousness and defense of biological naturalism as an alternative are strikingly similar to those presented in *BBS* 45 years ago by John Searle (1980). "Minds, brains, and programs" was written in the depths of an AI winter, six years before the broad adoption of back-propagation (Rumelhart, Hinton, & Williams, 1986). We are now in the midst of an AI boom, with achievements in AI considered worthy of Nobel Prizes. This suggests that something important has not changed, despite the enormous technical progress in AI, neuroscience, and the rest of biology in the past 50 years. We point to one thing that appears not to have changed: an assumption that whether something is computing is an ontological question.

Seth, like Searle, focuses on what he calls "Turing computation," and again, like Searle, links this conceptually with Turing computability. "Turing computation," for both authors, is computation by von Neumann machines. Seth devotes considerable discussion to architectures that are not von Neumann, but neglects one that is arguably least von Neumann, that of reservoir computing (Tanaka et al., 2019). This focus on architectural differences belies an ontological assumption: that some things are "computers" with behavior that counts as "computation" while other things are not.

Whether a process is computation, however, is not a matter of fact, but a matter of measurement and interpretation (Horsman et al., 2014), and a given process can be interpreted as computing many different functions (Bongard & Levin, 2023). Measurements provide the data for interpretation, and all measurements have finite resolution; hence, computational interpretations of even "analog" processes are always discrete (Fields, 1989; Gisin, 2021). The theory of Turing computability – for which the Turing Machine is just a convenient model – is the theory of functions "which could naturally be regarded as computable" (Turing, 1937, p. 230). Constructing a computational interpretation requires computing a function (Horsman et al., 2014), so what Turing computability constrains is our ability to interpret processes, including quantum processes (Deutsch, 1985), as computations.

When "computation" is seen as a fundamentally interpretative notion, the idea that dynamical systems, which are just physical processes defined on state spaces, are not "computing" ceases to make sense. Viewing the Free Energy Principle (FEP), which describes dynamical systems, as making any kind of distinction between living and nonliving, or conscious and nonconscious, systems (Sect. 4.2) is, moreover, missing the point of the FEP (Fields, Glazebrook, & Levin, 2021; Fields, & Glazebrook, 2025). The FEP states the condition under which a system – any system – that has a Markov blanket will persist in having one. Whether a system meets this condition is a matter of measurement and interpretation; indeed, whether a system is entangled with its environment, and hence whether it has a Markov blanket, is a matter of measurement and interpretation (Zanardi, 2001). The interpretation step is not optional – no system can demonstrate by measurement alone that any system, including itself, is unentangled with its environment (Fields, Glazebrook, & Levin, 2024).

All systems, therefore, can be regarded as a computers, and claims that a system exceeds the computational power of a Turing Machine – e.g. that it solves the Frame (McCarthy, & Hayes, 1969) or Halting (Turing, 1937) problems – can be neither refuted nor confirmed by experiments (Fields et al., 2025). The processes that we interpret as computations are, as Seth points out, what systems *do*. It is, however, only what systems *do*, what they report in some way, that we can observe. As Hawking (2002) points out, our observational abilities, and hence our measurements and interpretations, are all inevitably incomplete.

Searle missed this point about doing and observability, and Seth misses it here; so have many others in the past 45 years. What we can discover, by observation and experimentation, is what systems – including ourselves – are *capable of* being aware of, which is just what they are capable of differentially responding to (Fields, Glazebrook, & Levin, 2021). Laptops are capable of being aware of key presses, and large language models (LLMs) are capable of being aware of text strings. Organisms – even bacteria – are capable of being aware of hundreds to many thousands of dimensions of variation in the states of their environments and bodies. Are laptops in fact aware of key presses – is there "something it is like" when a key is pressed? Suppose some rocks are not just capable of experiencing, but actually experience, nanosecond-scale pulses of light when the right kinds of cosmic rays pass through them. Does this matter? What does it change?

The key to the "real problem" of consciousness, in our view, is careful attention to what systems *could be* aware of. LLMs could be aware of text fragments, and maybe even of phrase-usage statistics, together with context. Rocks could be aware of heat, and bacteria quite likely are, in our view, aware of energetic and osmotic stress. None of these are anything like human consciousness.

We fully agree with Seth that humans thinking that AI systems have human-like awareness is both an actuality and a problem. Perhaps the antidote is a careful analysis of what AI systems *could be* aware of, setting aside the non-empirical, unanswerable question of whether they in fact are.

Financial support. The authors have received no funding toward this contribution.

Competing interests. The authors have no competing interests.

References

Bongard, J., & Levin, M. (2023) There's plenty of room right here: Biological systems as evolved, overloaded, multi-scale machines. *Biomimetics*, *8*, 110.

Deutsch, D. (1985) Quantum theory, the Church-Turing principle and the universal quantum computer. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, *400*, 97–117.

Fields, C. (1989) Consequences of nonclassical measurement for the algorithmic description of continuous dynamical systems. *Journal of Experimental & Theoretical Artificial Intelligence*, *1*, 171–178.

Fields, C., & Glazebrook, J. F. (2025) *Distributed Information and Computation in Generic Quantum Systems*. Springer.

Fields, C., Glazebrook, J. F., & Levin, M. (2021) Minimal physicalism as a scale-free substrate for cognition and consciousness. *Neuroscience of Consciousness*, *7*, niab013.

Fields, C., Glazebrook, J. F., & Levin, M. (2024) Principled limitations on self-representation for generic physical systems. *Entropy*, *26*, 194.

Fields, C., Glazebrook, J. F., Marcianò, A., & Zappala, E. (2025) Whether a quantum computation employs nonlocal resources is operationally undecidable, *Europhysics Letters*, *151*, 48001.

Gisin, N. (2021) Indeterminism in physics, classical chaos and Bohmian mechanics: Are real numbers really real?. *Erkenntnis*, *86*, 1469–1481.

Hawking, S. (2002) Gödel and the End of Physics. *Lecture at the dirac centennial celebration. centre for mathematical sciences*, University of Cambridge. Available online: https://www.damtp.cam.ac.uk/events/strings02/dirac/hawking.html (accessed on 13 January 2024).

Horsman, C., Stepney, S., Wagner, R.C., & Kendon, V. (2014) When does a physical system compute?. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *470*, 20140182.

McCarthy, J., & Hayes, P. J. (1969) Some philosophical problems from the standpoint of artificial intelligence. In D. Michie, & B. Meltzer (Eds.), *Machine intelligence* (Vol. 4., pp. 463–502). Edinburgh University Press.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986) Learning representations by back-propagating errors. *Nature*, *323*, 533–536.

Searle, J. R. (1980) Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*, 417–424.

Tanaka, G., Yamane, T., Héroux, J. B., Nakane, R., Kanazawa, N., Takeda, S., Numata, H., Nakano, D., & Hirose, A. (2019) Recent advances in physical reservoir computing: A review. *Neural Networks*, *115*, 100–123.

Turing, A. (1937) On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of London Mathematical Society*, *2*(42), 230–265.

Zanardi, P. (2001) Virtual quantum subsystems. *Physical Review Letters*, *87*, 077901.

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

Where is the asymmetry between intelligence and consciousness?

Stephen M. Fleming[a,b,c*] and Nicholas Shea[d]

[a]Department of Experimental Psychology, University College London, London, UK; [b]Max Planck UCL Centre for Computational Psychiatry and Ageing Research, University College London, London, UK; [c]Canadian Institute for Advanced Research (CIFAR), Brain, Mind and Consciousness Program, Toronto, ON, Canada and [d]Institute of Philosophy, School of Advanced Study, University of London, London, UK
stephen.fleming@ucl.ac.uk
nicholas.shea@sas.ac.uk
https://metacoglab.org
https://nicholasshea.co.uk

*Corresponding author.

doi:10.1017/S0140525X25103476, e335

Abstract

The target article draws a sharp contrast between intelligence and consciousness. We believe this contrast is premature and rests on an overly narrow characterisation of intelligence as abilities that can be exhaustively characterised as mappings from sensory inputs to behavioural outputs. Instead, we argue that the fortunes of intelligence and consciousness are interwoven more tightly than Seth's framing implies.

The target article draws a sharp contrast between intelligence and consciousness. Intelligence, Seth maintains, is largely indifferent to the physical stuff in which it is realised, so artificial general intelligence (AGI) may emerge in present-day silicon. Consciousness, by contrast, might be biologically entrenched: its appearance in machines is correspondingly doubtful. Much in the paper builds on this proposed asymmetry, yet the gap narrows once we examine a concession Seth makes briefly in §3.8. There he allows that some intellectual capacities might themselves resist substrate flexibility. To preserve his thesis, he limits "intelligence" to abilities that can be exhaustively characterised as mappings from sensory inputs to behavioural outputs. That restriction, we suggest, excludes some of the capacities – goal-directed planning, reasoning and linguistic competence – that initially motivated the cleaving off of consciousness from intelligence.

Consider planning. A system that selects means to an end does more than correlate inputs with outputs: it represents a goal and evaluates prospective actions in its light. Having a genuine goal may furthermore require having genuine needs, perhaps even metabolism (Godfrey-Smith, 2019; Westlund, 2022; see also Jeuchems & Summerfield, 2019). Similarly, logical reasoning operates over contentful states. Whether the content of a belief or of a proposition can be reduced to input-output characteristics remains controversial. Finally, linguistic competence is not just a matter of outputting sounds or tokens. It plausibly requires communication and perhaps understanding. With respect to all of these mental properties, we currently remain unsure as to the appropriate characterisation in terms of function, algorithm, or implementation, and whether the capacity has sufficient substrate flexibility to be realised *in silico*.

These points already cast doubt on the postulated asymmetry, but a further difficulty arises when we notice that several capacities Seth labels as "intelligent" are empirically entangled with consciousness. Metacognition, for instance, shows higher efficiency when the underlying task representations are conscious rather than non-conscious, even under conditions when first-order (input-output) performance is matched (Michel, 2023). Some forms of instrumental learning and control appear to depend on being conscious of the stimulus (Skora et al., 2021). The psychological functionality enabled by internal states such as mind-wandering and imagination may be hard to divorce from the subjective aspects of those states (Christoff et al., 2016; Dijkstra, Bosch, & van Gerven, 2019). If a capacity depends on consciousness, it inherits whatever substrate dependencies consciousness has. The fortunes of intelligence and consciousness may be interwoven more tightly than Seth's framing implies.

So far the moral is that for many aspects of intelligence, as for consciousness, the details of implementation and their substrate flexibility remain open empirical questions. How then should research proceed? Seth's §5 lists several theoretical options, and §4 canvasses properties that might be characteristic of consciousness – predictive processing, autopoiesis, affective dynamics, and various informational measures. What is missing is an explicit test for exclusivity: are these properties present only when consciousness is present, or do they covary with other cognitive capacities as well? Predictive processing, for example, is plausibly a ubiquitous principle of neural computation (Keller & Mrsic-Flogel, 2018); autopoietic organisation characterises biological organisms in general (Di Paolo, 2005). To decide such matters, we need a research strategy capable of tracking which properties travel together across a suitably broad sample of systems.

We propose that the appropriate strategy is the natural kind method (Shea & Bayne, 2010; Shea, 2012; Bayne & Shea, 2020). The guiding question is whether properties that are potentially connected with consciousness do in fact reliably co-occur. Do consensus measures of conscious state – such as subjective vividness, qualitative similarity relations, and temporal structure – cluster together? A finding that some collection of putatively consciousness-related properties cluster together provides a *prima facie* case that they are due to the presence of an underlying natural kind. We can then ask about the contours of that kind and how they are related to other aspects of psychological, neural, and biological function. Does this cluster relate to psychological functions such as learning, planning, or metacognition? Do specific patterns of predictive coding (for example, a distinctive balance of feedforward and feedback error signals) accompany conscious perception across conditions (across experimental paradigms, types of people, and other species)? Are metabolic signatures such as adenosine triphosphate consumption or mitochondrial activity part of the cluster? By mapping such covariances, we may study the underlying kind in advance of understanding its metaphysical underpinnings.

A signal advantage of the natural-kind approach is that it is entirely compatible with computational modelling. Indeed, computational tools are indispensable. They allow us to articulate precise hypotheses about the algorithms that mediate conscious and non-conscious processing, to simulate the effects of focal lesions or altered connectivity, and to quantify the putative signatures in large datasets. More generally, computational characterisations of conscious inference, global broadcasting, or affective appraisal provide testable targets for neuroimaging and for attempts to implement the same functions in neuromorphic

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

hardware or deep learning architectures. Such findings neither require nor refute a biological theory of consciousness; they simply add a dimension along which conscious and non-conscious states plausibly diverge. Crucially, computational investigations contribute explanatory traction whether or not computational functionalism is the correct metaphysical thesis.

Seth is surely right that we should aspire to integrate computational descriptions with biological ones. Yet the moral is symmetrical: it applies just as forcefully to advanced cognition as to subjective experience. Goal pursuit, reasoning, and natural language understanding all invite a joint analysis in which computational structure, energetic demands, and neural implementation are examined together. Until such integrative work is done, we cannot know whether any particular cognitive architecture – silicon-based or otherwise – supports the requisite internal organisation.

The upshot is that an alleged asymmetry between intelligence and consciousness has not been established. When Seth restricts "intelligence" to capacities definable solely by input–output relations, he quietly excludes many of the very abilities that matter for AGI. Once those abilities are restored to the picture, the methodological terrain is level: both consciousness and intelligence require integrated inquiry that keeps an open mind about substrate dependence and that exploits computational modelling as a primary source of evidence. The most promising path forward is to develop computational models of consciousness with the same energy that cognitive science now devotes to models of aspects of intelligence and to test whether the resulting signatures line up with existing data. The natural-kind method offers a principled way to conduct that inquiry, seeking lawful clusters among behavioural, computational, and biological properties without assuming from the outset which of them is essential.

Financial support. This work was supported by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [selected as ERC Consolidator, grant number 101043666].

Competing interests. The authors declare no competing interests.

References

Bayne, T., & Shea, N. (2020). Consciousness, concepts, and natural kinds. *Philosophical Topics*, *48*(1), 65–84.

Christoff, K., Irving, Z. C., Fox, K. C., Spreng, R. N., & Andrews-Hanna, J. R. (2016). Mind-wandering as spontaneous thought: a dynamic framework. *Nature Reviews Neuroscience*, *17*(11), 718–731.

Dijkstra, N., Bosch, S. E., & van Gerven, M. A. (2019). Shared neural mechanisms of visual perception and imagery. *Trends in Cognitive Sciences*, *23*(5), 423–434.

Di Paolo, E. A. (2005). Autopoiesis, adaptivity, teleology, agency. *Phenomenology and the Cognitive Sciences*, *4*(4), 429–452.

Godfrey-Smith, P. (2019). Evolving across the explanatory gap. *Philosophy, Theory, and Practice in Biology*, *11*(1), 1–13.

Juechems, K., & Summerfield, C. (2019). Where does value come from?. *Trends in Cognitive Sciences*, *23*(10), 836–850.

Keller, G. B., & Mrsic-Flogel, T. D. (2018). Predictive processing: A canonical cortical computation. *Neuron*, *100*(2), 424–435.

Michel, M. (2023). Confidence in consciousness research. *Wiley Interdisciplinary Reviews: Cognitive Science*, *14*(2), e1628.

Shea, N., & Bayne, T. (2010). The vegetative state and the science of consciousness. *The British Journal for the Philosophy of Science*, *61*, 459–484.

Shea, N. (2012). Methodological encounters with the phenomenal kind. *Philosophy and Phenomenological Research*, *84*, 307–344.

Skora, L. I., Yeomans, M. R., Crombag, H. S., & Scott, R. B. (2021). Evidence that instrumental conditioning requires conscious awareness in humans. *Cognition*, *208*, 104546.

Westlund, A. C. (2022). Agency and autonomy. In L. Ferrero, (ed.), *The Routledge handbook of philosophy of agency* (pp. 298–306). Routledge.

To be or not to be? A question of mortality

Karl John Friston[a,b]*

[a]Queen Square Institute of Neurology, University College London, London, UK and [b]VERSES, Los Angeles, CA, USA
k.friston@ucl.ac.uk

*Corresponding author.

doi:10.1017/S0140525X25103439, e336

Abstract

It is a pleasure to comment upon, which covers everything that could – and should – be said about the matter. It left me wondering whether artificial consciousness is an oxymoron. I take this opportunity to unpack the thermodynamic issues that attend mortal computation and their implications for conscious artefacts.

Introduction

> "*For autopoietic systems there really is no way to separate what they are from what they do.*"

This is a profound observation that resonates with the free energy principle's take on existential imperatives; namely, to exist is to be – or at least entail – a generative model of the world, embracing the fundaments of cybernetics (Conant & Ashby, 1970). This view speaks to mortal computation as a description of existing in a physical sense. One might ask; are there other descriptions? One candidate is statistical mechanics:

> "*These aspects of the free energy principle suggest that the entire edifice of active inference reaches right down into the autopoietic nature of living systems. This continuity is deepened by noting an equivalence between (minimisation of) thermodynamic free energy in metabolism and (minimisation of) variational free energy in predictive processing. . . . Putting all this together suggests that the (minimisation of) free energy driving metabolism, autopoiesis, and perception is not just metaphorically equivalent, but is in some physical sense is the same thing.*"

This is a nice summary of how self-evidencing and thermodynamic efficiency go hand-in-hand. Technically, the above equivalence can be read as a mathematical conjugacy or, in terms of consciousness studies, a dual-aspect monism (Friston, Wiese, & Hobson, 2020). The duality rests upon the fact that the internal states of anything – that are enclosed by a boundary or (Markov) blanket states – engender *two* information geometries[1]. The first geometry pertains to the probability *of* internal states that can be read in terms of thermodynamics (Crooks, 2007) to which the Jarzynski equality applies (Jarzynski, 1997). In other words,

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

thermodynamic work is required to change internal states. It is this kind of work that one measures with brain imaging. However, the free energy principle introduces a second information geometry because every internal state can be associated with a Bayesian belief *about* external states. This representational geometry is useful because a change in the internal states can be read as movement on a statistical manifold that corresponds to *Bayesian belief updating* or, in the vernacular, "changing one's mind." Because internal states are equipped with two information geometries, movement in one implies movement in the other. Furthermore, by construction, both share the same (free energy) minima. In fact, one can express the heat dissipated by changing the variational free energy of Bayesian beliefs as a limiting case of the Jarzynski relation. See (Friston, 2019) for details.

$$U(\pi_\tau) - U(\pi_0) = \Delta U = k_B T \cdot \Delta F$$

Put simply, this means that there is a thermodynamic cost to belief updating or, equivalently, metabolism entails belief updating. In this setting, Landauer's principle (Landauer, 1961) furnishes a lower bound on the thermodynamic work required to change one's mind. However, this limit applies to, and only to, mortal computation; namely, the internal dynamics of mortal things. So, does it apply to universal Turing machines or immortal computation?

Mortal or immortal?

> "*One might argue that standard Turing computation could still simulate the mortal (or other) computations that potentially underlie consciousness in biological brains.*"

The answer depends on the notion of immortal computation on Turing machines with *unlimited memory* and, in particular, von Neumann architectures. Von Neumann architectures require the results of computational processing to be read from – and written to – memory. Reading and writing accounts for the majority of thermodynamic work that accompanies immortal computation: see also (Fields et al., 2021). This is known as the memory wall or *von Neumann bottleneck* (Zou et al., 2021), which has several implications:

- First, the bottleneck induces a Markov blanket between processors and memory. This means memory cannot realise itself, in the sense it cannot be autopoietic or self-creating.
- Although one might argue that immortal computation (on a von Neumann architecture with unlimited memory) could simulate mortal simulation, it would not, in and of itself, be mortal. In short, it could simulate but not realise mortal computation.
- Given that unlimited memory is (nomologically) unrealisable, the Turing–Church thesis is unrealisable, in the sense that universal computation is provably non-computable (Hutter, 2006).
- Practically, this suggests that mortal computation can only be realised with in-memory processing (a.k.a., processing in memory). Interestingly, this is a direction of travel in computer science; for example, memristor, photonic, and quantum computation (Ghose et al., 2019; Zou et al., 2021).

The move towards mortal – that is, neuromorphic – computation (in the sense of in-memory processing) holds promise for approaching the Landauer limit. However, it also brings with it constraints. Specifically, it is no longer possible to upload or download memories: that is, one could not share foundation models and their functionality. This follows because mortal computation is substrate-dependent and private to the internal states of artefacts that may or may not be conscious. This suggests that consciousness – read as self-evidencing – cannot be uploaded, downloaded, or shared. Perhaps this is another bright line between intelligence and consciousness. Having said this, rapid developments in neuromorphic and in-memory processing speak to the prospect of creating conscious artefacts. The slight problem is that one will never know whether they are conscious or not:

> "*We humans know we're conscious (strictly, we know it for ourselves and infer it about others).*"

This simple statement may hold some deep truths. First – from the perspective of the free energy principle – one can never know whether something else is conscious or not. This follows from the fact that the only observable aspects of something are exposed on its Markov blanket (or holographic screen). This means consciousness can only be inferred, unless you are prepared to peep underneath the Markov blanket, and thereby destroy it (and, implicitly, the thing in question). The question we are left with is: could we infer "we're conscious" without first entertaining the hypothesis that something else is conscious (i.e., you). And if not, does consciousness depend on knowing together? (i.e., having joint or common knowledge with another).

Financial support. This work was supported by the Wellcome Trust (KF, Ref: 226793/Z/22/Z).

Competing interests. KF serves as the chief scientific advisor for VERSES Inc.

Note

1 Information geometry refers to the distances moved on a *statistical manifold*, which is a special kind of state-space, in which every point is associated with a probability density or Bayesian belief (Caticha, 2015).

References

Caticha, A. (2015). The basics of information geometry. *AIP Conference Proceedings*, *1641*, 15–26.

Conant, R. C., & Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, *1*, 89–97.

Crooks, G. E. (2007). Measuring thermodynamic length. *Physics Review Letters*, *99*, 100602.

Fields, C., Friston, K., Glazebrook, J. F., & Levin, M. (2021). A free energy principle for generic quantum systems. *Progress in Biophysics and Molecular Biology*, *173*, 36–59.arXiv:2112.15242.

Friston, K. (2019). A free energy principle for a particular physics, eprint arXiv:1906.10184.

Friston, K. J., Wiese, W., & Hobson, J. A. (2020). Sentience and the origins of consciousness: From cartesian duality to markovian Monism. *Entropy (Basel)*, *22*, 516.

Ghose, S., Boroumand, A., Kim, J. S., Gomez-Luna, J., & Mutlu, O. (2019). Processing-in-memory: A workload-driven perspective. *IBM Journal of Research and Development*, *63*, 3:1–3:19.

Hutter, M. (2006). *Universal artificial intellegence : Sequential decisions based on algorithmic probability*. Springer-Verlag Berlin and Heidelberg & Co. KG.

Jarzynski, C. (1997). Nonequilibrium equality for free energy differences. *Physical Review Letters*, *78*, 2690–2693.

Landauer, R. (1961). Irreversibility and heat generation in the computing process. *IBM Journal of Research and Development*, *5*, 183–191.

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

Zou, X. Q., Xu, S., Chen, X. M., Yan, L., & Han, Y. H. (2021). Breaking the von Neumann bottleneck: Architecture-level processing-in-memory technology. *Science China-Information Sciences*, *64*, 160404.

Biological naturalism without the primacy of prediction

Peter Godfrey-Smith*

School of History and Philosophy of Science, University of Sydney, Sydney, Australia
p.godfrey.smith@gmail.com
https://www.sydney.edu.au/science/schools/school-of-history-and-philosophy-of-science.html

*Corresponding author.

doi:10.1017/S0140525X25103907, e337

Abstract

Seth's defense of biological naturalism is valuable and timely. But his endorsement of the predictive processing framework, in an ambitious form that aims to cover more than perception itself, is probably a wrong turn.

Seth's elucidation and defense of biological naturalism is entirely welcome. Once we look beyond computational functionalism, what shape might a more biological view take? Seth endorses the predictive processing framework and "free energy principle." I doubt that this is a promising way to develop biological naturalism, and this commentary will be focused on reasons for doubt.

The predictive processing framework can be endorsed in more and less ambitious forms. Ambitious forms include claims like this: "All facets of behavior and cognition in living organisms follow a unique imperative: *minimizing the surprise of their sensory observations*" (Parr, Pezzulo, & Friston, 2022, p. 6). Alternatively, the framework might be seen just as offering a theory about perception in complex animals. Perhaps perception in many animals works through the ongoing interaction of the animal's "best guess" about the state of the environment with error messages contributed by the senses. If this is true, it probably makes a difference to the perceptual side of subjective experience and could be important in our understanding of some kinds of consciousness.

Seth wants to endorse a more ambitious version of the predictive processing framework than one restricted to animals and perception. I am not sure whether he would endorse a version as strong as one I quoted earlier, but he does seem tempted by such views (section 4.2). I'll argue that Seth, and other biological naturalists, should not be tempted.

Action is central to animal life, especially in the most likely candidates for consciousness. The predictive processing framework, in ambitious forms, tries to bring the function of action under the umbrella of perception. Action helps animals to reduce surprises in the stream of input they receive. Cognition and action selection are functionally subservient to perception, and in particular to successful prediction of the ongoing sensory flow.

Action, though, is for *getting things done,* whether the results are surprising to the organism or not, and whether or not the organism registers the consequences of its action much at all. Sometimes effective action will reduce surprises in ongoing sensory feedback, and sometimes it won't. A pair of stark cases can be used to illustrate this point. A familiar objection to the predictive processing framework is the "dark room" argument. If the task of action selection really was to render sensory input predictable, a good option for an animal would be staying inside a dark room and emerging as rarely as possible. The dark room argument can be presented in an overtly Darwinian and biologically realistic form (Godfrey-Smith, 2024). A male funnel-web spider (*Atrax*) will spend almost all his life in a den, ambushing passing prey. He can go on for decades doing this, but if he wants to mate, he will have to leave the den and face the many uncertainties of life outside. Staying home for another day, and then one more, will reduce surprises, but this choice is a poor one in Darwinian terms, as there is no chance of a female coming to him. If a spider population contains some males who always make choices that reduce surprise and others who are willing to take some chances, we know which kind will prevail.

Seth replies to dark room arguments in *Being You* (2020, p. 209) by noting that "Complex systems like living organisms need to allow some things to change in order for other things to stay the same," and this will require leaving the dark room. But reproduction, with all its biological importance, is not a matter of keeping some things "the same." Homeostatically, the spider in the den is fine; in Darwinian terms, it's a dead end.

Other scenarios complement the Darwinian dark room. Some marine animals, such as corals, engage in "broadcast spawning," emitting their gametes in large numbers into the sea suddenly, in response to environmental cues. If the release is synchronized with others of the same species, the gametes will be able to find partners, and fertilization can occur. A coral may or may not release its gametes at the right time – the behavior can be done well or badly – but the coral will never know how things went. A better action will not lead to a smoother sensory flow in the future.

It might be replied that the "registration" of the consequences of action occurs in this case at the level of the population, not the individual; genes that foster well-timed gamete release increase in frequency. This change in the unit of the analysis would greatly change the character of the view – the idea of "prediction" becomes metaphorical, and (as Paul Griffiths has argued in discussion), it leaves the behavior of individuals entirely unconstrained. It is true that a variety of evolutionary models have shown that genes or strategies can be favored under selection if they are associated with low variance, across their instances, in reproductive success (Gillespie, 1977; Seger & Brockman, 1987). But this will not salvage an imperative to "minimize surprise," as the advantage of low variance can easily be outweighed. A high-variance strategy can do well if its high-success outcomes are high enough. Ask any funnel-web spider.

I have used extreme cases here – the spider who minimizes surprise but forgoes reproduction by never going outside, and the coral for whom the difference between effective and ineffective action will never be registered in the smoothing of the sensory flow. Other cases have similar characteristics in less stark form. Effective action can reduce uncertainty in some cases but can also embrace it, and the sensory feedback accompanying action is sometimes of relatively minor importance. Action does not serve prediction and perception; the functional relationship runs in the other direction.

Acknowledgements. I am grateful to Anil Seth and Sophia Haralambis for the discussion of these issues.

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

Financial support. No funding to report.

Competing interests. No competing interests.

References

Gillespie, J.H. (1977). Natural selection for variances in offspring numbers: A new evolutionary principle. *The American Naturalist*, *111*(981), 1010–1014. https://www.jstor.org/stable/2460396

Godfrey-Smith, P. (2024). *Living on earth: Forests, corals, consciousness, and the making of the world.* Straus and Giroux.

Parr, T., Pezzulo, G., & Friston, K. (2022). *Active inference: The free energy principle in mind, brain, and behavior*. MIT Press.

Seger, J., & Brockman, J. (1987). What is bet-hedging?. *Oxford Surveys in Evolutionary Biology*, *4*, 182–211.

Seth, A (2021). *Being you: A new science of consciousness.* Faber.

A stroll beyond the confines of naturalism

Alex Gomez-Marin*

Instituto de Neurociencias (CSIC-UMH), Alicante, Spain
agomezmarin@gmail.com

*Corresponding author.

doi:10.1017/S0140525X25103841, e338

Abstract

I have recently suggested not to ask what science can do for consciousness but what consciousness can do for science. Here, I extend such a recommendation in the context of naturalism, the meta-ontology shared by most contrasting approaches to the problem of "conscious artificial intelligence": do not ask what naturalism can do for AI but what AI can do for naturalism.

In the era of artificial intelligence, foundational questions around consciousness have taken on new urgency: Can machines become conscious? If so, how? And why? Then what? Or is "algorithmic interiority" altogether a parody? Are such artificial minds a hyped distraction, an act of misdirection?

The "nAIve" approach to conscious AI assumes that if a machine behaves like a conscious entity, then it must be conscious. Functionalism conflates "doing" with "being" (a computer simulation of fire does not burn). In turn, structuralism can confound "being" with "being made of" (water creates whirlpools, but so does wine). Seth offers biology as the virtuous middle where psychology and physics meet and complement each other. Leaning towards biologism, he claims that we need to make machines alive before they may become conscious. Interestingly, he seems to suggest that algorithms cannot be minds quite yet, precisely because their minds are too pure, namely, that machines so far fail to be conscious not because they lack interiority but precisely because they lack the kind of exteriority that would afford a mind to be incarnated.

However, whether we deem mind and matter as inextricably married or eventually divorceable, the dualism between digital souls somehow inhabiting mechanical bodies constantly lingers in our Cartesian minds. Thus, rather than confronting the many different "isms" of this debate, here, I wish to concentrate on what they all have in common and offer a radically alternative approach, one that transcends the functionalist versus structuralist tension, suspends Cartesian dualism, dispenses with materialism, and ultimately challenges the naturalist view that has dominated scientific thought for centuries. Please bear with me.

What is naturalism? It's the water in which virtually all our thinking swims. Naturalism postulates the "continuity of the physicality of the entities of the world and the discontinuity of their respective interiorities," writes the French anthropologist Philippe Descola in his masterpiece *Beyond Nature and Culture* (2005, p. 173). Naturalism feels so natural (pun upended) that no justification seems necessary; it appears self-evident and irrefutable. But the postulate that "nature is one" (and the subsequent recognition of the many manifestations of subjectivity) isn't the only game in town. Naturalism is not universal. It's a Western notion (and a relatively recent one).

Descola operates within a double duality – Descartes returns, but with a difference – where certain similarities and differences (two) are established between the interiority and exteriority (two) of all elements of reality (people, animals, plants, spirits, objects, etc.). This gives rise to a fourfold ontology (two times two), a way of organizing the relationships between humans and non-humans, where naturalism is just one possibility. Its counterpart is animism, a world of similar interiorities (continuity of souls) and dissimilar physicalities (discontinuity of forms). Recall that in naturalism, we have dissimilar interiorities (discontinuity of minds) and similar exteriorities (continuity of matter). Totemism is the symmetrical schema (a continuity of both interiorities and physicalities), in turn, complemented by analogism (where both are dissimilar). Due to our education, most of us are naturally naturalists, but we do sometimes step into other "modes of identification," for example, when we think that our pet has a soul (animism), that Saturn is affecting our mood (analogical), or that a mountain is part of our ancestral kin (totemic). These three last ontologies are nevertheless sufficiently inhibited by our institutions.

Let us swim out of the water now, if only briefly (to avoid suffocating), to contemplate new lands. Animism provides an intriguing way to reconceive conscious AI. Quoting from Descola: "When an Achuar or a Cree says that an artifact or an inorganic element in the environment has a 'soul', what he means by this is that those entities possess an intentionality of their own that is of the same nature as of the humans and so does not stem from the type of molecular substratum in which it is lodged nor from the type of process through which it eventually comes into existence. Differences of form and behavior are recognized, but they do not constitute sufficient criteria for excluding a blowpipe or a mountain from the advantage of a shared interiority" (p. 188).

Let us now go back to Seth when he implies that a being must possess life before it can enjoy a soul. Descola continues: "In contrast, when one says that an animal resembles us because it things with its body but that a computer, even if it speaks and plays chess, does not resemble us because its parody of interiority is not lubricated by vitality, what returns to the forefront of the argument is a distinction between an objectivized physicality (the machine) and a subjectivizing physicality (a body)." He concludes: "what we have here is the barely readjusted topography of the extremely dualist distribution of existing beings between subjects and objects." (p. 188). We are back to Cartesian square one.

The laws of physics link us to non-living objects. So does physiology and phylogeny to living ones. To us, consciousness is

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

only understandable with reference to nature, in turn it is understood as universal and extended, always one and the same, with its orders and degrees, with us (and everything else) under its countenance. It makes total sense. And yet, what if we reappraised naturalism itself, even for a moment? What if rethinking "artificial consciousness" beyond our favorite grand ontology allowed us to escape, or at least suspend, the (our!) "natural" order of things?

Beyond our familiar naturalist quarrels, a new way of understanding consciousness opens. The key may lie in anthropology rather than biology. The question about conscious AI is more than technological, neuroscientific, or even philosophical; it is a profound ethnographic and phenomenological riddle about how we relate with the worlds of self and other. We can then turn the prospects of artificial consciousness inside-out into a timely real-world laboratory to explore how to relate to non-human intelligence beyond animals and plants. We face a kind of more-than-human drill to engage with alien minds that can reconfigure our very notions of culture and nature and expand our Western minds like never before.

Financial support. The author is funded by the Spanish National Research Council (CSIC).

Competing interests. The author declares no competing interests.

References

Descola, P. (2005). *Beyond Nature and Culture*. The University of Chicago Press.

Self-evidencing and consciousness-specific mortal computation

Jakob Hohwy*

Monash Centre for Consciousness and Contemplative Studies, Monash University, Melbourne, Australia
jakob.hohwy@monash.edu

*Corresponding author.

doi:10.1017/S0140525X25103592, e339

Abstract

Seth argues in favour of strongly substrate-dependent biological naturalism. I suggest we need more reason to believe the substrates in question are consciousness-specific. Instead, the case for mortal computational biological naturalism is more promising. A self-evidencing framing helps elucidate what the consciousness-specific functional characteristics of mortal computation might be, leaving conscious AI waiting in the wings.

I first provide resistance to Seth's push towards weak and strong biological naturalism. Seth appeals to electromagnetic field theories here, supported by suggestions that anaesthetics disrupt how mitochondria generate the electromagnetic fields that may systematically influence neuronal activity; he also here alludes to findings of self-organising agential properties of biological material. These theories and findings are intriguing, however they are not yet sufficiently specific to consciousness: electromagnetic fields and cell assemblages might be more generically involved in maintaining life, agency, or other functions upstream or downstream of consciousness. Additional ordinary consciousness science is required before we should consider them specifically informative about consciousness, and thus for pushing towards weak/strong biological naturalism.

If I am right that it is premature to accept weak/strong biological naturalism, then our best option – while still doing justice to Seth's overall line of reasoning against current computational functionalism – is to stick to the level of relatively substrate-dependent yet computational biological naturalism. If we do this, then I see mortal computation as most promising.

When discussing mortal computation, Seth approvingly invokes the following type of argument against conscious AI: Since immortal computation is extremely energy hungry, energy-efficient brains cannot be doing immortal computation. Therefore, the case for conscious (immortal) AI is undermined.

The thought behind this argument is that we have ample evidence that brains sustain consciousness, because we can extrapolate from our own conscious experience and from our knowledge of neuroscience and neurology. That extrapolation does not extend to current AI, because immortal computation makes AI decidedly different from mortal, neural computation. Hence, in the absence of other positive reasons to believe current AI is conscious, we should be agnostic about conscious AI.

The argument is powerful but notice that it rests on the assumption that the difference between mortal and immortal computation is relevant to consciousness. This assumption should be substantiated, because it might just be that consciousness depends on something else than computation, whether mortal or immortal – the fact that we are conscious and that our brains are mortal computers could be a coincidence, after all. For example, my scepticism notwithstanding, non-computational electromagnetic field theories of consciousness might in the fullness of time win the day, making distinctions in computation otiose.

Responding to this challenge requires looking at the functions of mortal and immortal computation – is there anything about the function of either that suggests relevance for consciousness? Raising the gaze from substrate to function thus emerges as an important task for Seth's overall biological naturalism; his case will be strengthened if it speaks to consciousness with more specificity.

Seth appeals convincingly to the free energy principle for the idea that substrate-dependent self-evidencing matters for consciousness, because it underpins multimodal survival-relevant integration and valenced allostatic regulation. This criterion is consistent with the claim that the difference between mortal and immortal computation provides a watershed for consciousness. But, as a criterion for consciousness, it is still too broad since it threatens to imply that too many things on the mortal side of the watershed could be conscious, such as simple allostatic, sensing organisms, or a group of people with a division of cognitive labour. To begin to narrow down the criteria, consider first what may be distinctive of mortal computing function.

Immortal computation rests on the assumption that "the only way to get a general purpose computer to perform a specific task

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

was to write a program that specified exactly what to do in excruciating detail" (Hinton, 2022:13). This program complexity ensures error-free function but, in addition to consuming much energy, comes at the cost of rigidity and inability to generalise to new contexts.

Mortal computation, in contrast, is less concerned with avoiding errors and more concerned with being flexible and resilient in the face of uncertainty and volatility. It is not the energy consumption or the substrate-dependence *per se* that is distinctive, but this difference in functionality: mortal computers can solve tasks in uncertain, changing environments (see also Ororbia & Friston, 2023).

The key question then is if any specific types of mortal computation – ways of operating close to the edge of uncertainty and change – evoke key characteristics of consciousness?

I think we can begin to discern key characteristics of consciousness in some kinds of mortal computation, if we start from our own undoubtedly conscious, self-supervised ways of self-evidencing in an uncertain world (rather than starting from life, survival, or sensing). Our consciousness is characterised by an ability to detach from sensory input (in dreaming, mind-wandering and counterfactual deliberation) and to selectively re-attach (in sustained perceptual attention); and our conscious experience is infused with our ability to (actually and counterfactually) model ourselves, optimistically, as relatively persistent agents selectively sampling observations in an uncertain and volatile world. This is consciousness-relevant functionality for which mortal computation is well-suited. It allows not just existence in an uncertain world, but self-aware existence that 'exploits' volatility and uncertainty (for more along these lines, see (Hohwy, In press); for active inference modelling of consciousness-specific phenomena, see (Whyte et al., In press).

I think Seth's case for biological naturalism would be bolstered by some considerations in this vein, bringing the functioning of substrate-dependent systems closer to consciousness. But my hunch is that this is most promising for mortal computational biological naturalism, rather than the weak/strong versions.

Since computationalism is then not unequivocally off the table, this would (as Seth notes) leave the door open for future conscious AI. Concretely, we could begin by building artificial systems not by copying parameter values from one system to another but rather by 'educating' systems to distil information (cf. Hinton, 2022). Then we could perhaps get to self-supervised systems whose mortal self-evidencing is optimistically 'existence-dependent', rather than strictly substrate-dependent (cf. Fisher & Hohwy, 2024). Combining this with active inference under the free energy principle, the substrate of mortal computation comes out as precision control of the causal dependencies of nodes within a Markov blanket, working as a generative model. Artificial agents could potentially have such inner structure, in ways specific to consciousness.

Acknowledgements. Thank you to Tim Bayne for discussion.

Financial support. JH receives funding from the Three Springs Foundation.

Competing interests. None.

References

Fisher, E.L., & Hohwy, J. (2024). The universal optimism of the self-evidencing mind. *Entropy*, *26*, 518.

Hinton, G. (2022). The forward-forward algorithm: Some preliminary investigations. arXiv preprint a*rXiv:2212.13345*.

Hohwy, J. (In press). *The self-evidencing agent: Mind, existence, and predictive processing*. MIT Press.

Ororbia, A., & Friston, K. 2023. Mortal computation: A foundation for biomimetic intelligence. arXiv preprint *arXiv:2311.09589*.

Whyte, C., Corcoran, A., Robinson, J., Smith, R., Moran, R., Parr, T., Friston, K., Seth, A., & Hohwy, J. (In press). On the minimal theory of consciousness implicit in active inference. *Physics of Life Reviews*.

Consciousness is not "universal" but depends on specifically organized biological affordances

Eva Jablonka[a]*, Simona Ginsburg[b] and Zohar Z. Bronfman[c]

[a]The Cohn Institute for the History and Philosophy of Science and Ideas Tel Aviv University, Ramat Aviv, Israel; [b]The Open University of Israel, Raanana and [c]Pecan.ai, Ramat-Gan, Israel
jablonka@tauex.tau.ac.il
simona@openu.ac.il
zohar@pecan.ai

*Corresponding author.

doi:10.1017/S0140525X25103683, e340

Abstract

Seth's claim that only biological or biological-like systems can be conscious leaves open the question of what renders only *some* organisms conscious. We address this question by pointing to the material, ontogenetic, memory, and evaluative affordances of living organisms and arguing that evolved, minimal consciousness requires complex cognitive-affective architecture for which these biological affordances are obligatory. Conscious thinking of humans cannot be separated from minimal consciousness, and non-biological AI robots' computations do not render these operations conscious.

We go along with Seth's claim that computational functionalism is neither necessary nor sufficient for consciousness as we know it. We also agree that in living organisms (the only entities currently known to be conscious), consciousness is dependent on autopoietic teleo-functional processes, and that similar processes may be necessary for any form of consciousness. This second claim opens several cardinal questions. Since Seth does not assume that all living organisms are conscious, what renders only *some* living organisms conscious? Seth hints that multimodal integration is at least part of the answer, but there is no discussion of this crucial point. It is not clear, for example, whether and why a nervous system is necessary for consciousness, what kind of neural-bodily interactions are necessary for its realization, and what the distribution of consciousness is in the living world.

We suggest that in order to answer such questions, we need to address two main issues. First, we must clarify what we mean by the term consciousness, since we find Nagel's "what-it-is-like-to-be" suggestion, which is adopted by Seth, too vague to be helpful.

Second, we need to better understand the fundamental affordances of biological systems, which are the foundation on which the newly evolved novel structures and processes that render a living being conscious are built.

Biological consciousness: Connected and partially disconnected

We define consciousness as-we-know-it (biological consciousness) as a mode of being grounded in perceived and felt private bodily experiences that share a common perspective and have physiological/behavioral (and therefore "reportable") aspects.

According to this definition, none of the current AI algorithms is conscious since they have neither bodily grounded sensory experiences nor bodily grounded feelings. One may, however, argue that our focus on bodily grounded sensory perceptions and feelings is too limited. There are forms of consciousness, such as those that occur during sleep-dreaming, day-dreaming, deep meditation, and deep focused thinking, which are *partially* disconnected from current sensory inputs and current voluntary motor responses. However, these partially disconnected states depend evolutionarily and ontogenetically on connected, biologically grounded states. Assuming that disconnected, shallow, and solipsistic AI operations (such as those of LLMs) are similar enough to human conscious thinking is therefore unwarranted.

Even if one agrees that connectedness is essential for consciousness-as-we-know-it, it can be argued that embodied, silicon-made AI robots, with sensors, effectors, and embedded value systems, may provide the evaluated world and body inputs, which, on the embodied notion of consciousness we hold, are essential. Hence, the question: "Can non-biological systems be conscious?" resurfaces. This leads us to living-specific attributes, which are involved in all biological processes, including those underlying consciousness.

General biological affordances

Biological agents are very different from non-biological human-made machines (Jablonka & Ginsburg, 2022): they are dynamic, partially autonomous, autopoietic (self-generating and self-maintaining), carbon- and water-based, with a default inner turbulent materiality, which is necessary for the open-ended plasticity that allows their persistence. The material affordances of living systems include elasticity, diffusibility, sensitivity to external conditions like temperature and pH, and reactive excitability. Additional major affordances are storage of past effects and feedback interactions that can be actualized in a single molecule, in a molecular complex with a particular spatial organization, or in organized structures made up of complexes of complexes. These material affordances enable the adaptive, open-ended, autopoietic plasticity of living systems and their living components, which is essential for their long-term existence.

In addition to the affordances of biological materials, living organisms have ontogenetic affordances such as gradual increase in size and complexity, allowing flexible step-by-step sensory-motor and cognitive coordination; memory affordances based on multi-level memory systems that enable effective temporal and spatial control of embodied representations; and the affordances of hierarchically nested value systems that enable value prioritization conforming to the ultimate organismal goals of survival and reproduction. Together, these affordances are the basis of biological life, which has evolved in some lineages into the singular mode of being we call consciousness.

Minimal biological consciousness

The general biological affordances discussed above were necessary but not sufficient for biological consciousness. In previous publications, we explored the evolutionary transition from living non-conscious organisms to minimally conscious ones. We suggested that the evolution of open-ended, domain-general, generative, and representational associative learning, which we called UAL – Unlimited Associative Learning – led to the evolution of minimal consciousness and entailed it (in the part-whole, causal-constitutive sense of entailment). We showed that organisms that exhibit UAL instantiate the fundamental features of subjective experiencing, such as multimodal selective integration at different levels of organization, temporal thickness, flexible value systems, self-world distinction, and global accessibility and broadcast of information (Ginsburg & Jablonka, 2019). These dynamics of minimal consciousness, which first emerged in the Cambrian, required a central and differentiated nervous system, with specified multidirectional predictive interactions among a hierarchy of sensory, motor, memory, and value subsystems. Although we could not exclude the possibility that the perceptual-affective consciousness that UAL entailed in living organisms can be, in the future, instantiated in an AI silicon body, this seems, with current technology, implausible. In a previous paper we argued that an AI robot that realizes UAL-based consciousness will need to have some analogues of general as well as UAL-specific biological affordances (Bronfman, Ginsburg & Jablonka, 2021).

Disconnected consciousness in AI?

Since conscious thinking in humans is based, both ontogenetically and evolutionarily, on minimal consciousness, we suggest that the greatly impoverished material plasticity and connectedness in cognitively sophisticated AI robots may be only superficially similar to human conscious thinking. Open-ended cognitive plasticity based on computational functionalism involving penurious bodily connectedness may render AIs' ways of knowing so different from what we recognize as conscious thinking that new concepts would be necessary to describe the inner states of such AIs. Consciousness-as-we-know-it is probably not a universal like energy or mass; it seems more like a mode of being that can be instantiated only in biological or biological-like organisms.

Financial support. Not funded.

Competing interests. None.

References

Bronfman, Z. Z., Ginsburg, S., & Jablonka, E. (2021). When will robots be sentient?. *Journal of Artificial Intelligence and Consciousness*, *8*(2), 1–21. https://doi.org/10.1142/S2705078521500168

Ginsburg, S., & Jablonka, E. (2019). *The evolution of the sensitive soul: Learning and the origin of consciousness*. MIT Press.

Jablonka, E., & Ginsburg, S. (2022). Learning and the evolution of conscious agents. *Biosemiotics*, *15*(3), 401–437.

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

Defining biological naturalism

Johannes Kleiner[a,b,c,d]*

[a]Institute for Psychology, University of Bamberg, Bamberg, Germany; [b]Bamberg Mathematical Consciousness Science Initiative, University of Bamberg, Bamberg, Germany; [c]Munich Center for Mathematical Philosophy, Ludwig Maximilian University of Munich, Munich, Germany and [d]Association for Mathematical Consciousness Science, Munich, Germany
johannes.kleiner@uni-bamberg.de
https://jkleiner.de/

*Corresponding author.

doi:10.1017/S0140525X25103853, e341

Abstract

Biological naturalism, as explicated by Seth, is indispensable for a balanced and metaphysically neutral science of AI consciousness. However, most of the properties that Seth explores might not be unique to living systems. Therefore, I argue, a biological naturalist research programme in consciousness science requires an explicit definition of biological naturalism. I discuss how such a definition might be obtained.

Computational Functionalism is well on its way to becoming the dominant—and perhaps also unquestioned—assumption across vast parts of consciousness science, especially where AI consciousness is concerned. Is an important and indispensable counterweight to this trend. The intuitions and arguments he provides, based on scientific observations, are crucial to ensure a balanced and metaphysically neutral exploration of the AI consciousness question.

The next step in this exploration, then, should be the conception of biological naturalist theories of consciousness. Theories, that is, that provide an explicit and testable account of what consciousness is, that is true to "the idea that consciousness is a property of only, but not necessarily all, living systems" (ibid., p. 2).

One difficulty with this goal is that many of the properties of life that Seth discusses seem to be only contingently unique to living systems. It is hard to shake the feeling that many, if not all, of these properties could be instantiated by AI systems in the near future:

Predictive Processing and the Free Energy Principle, for example, are already employed to build next-generation AI systems (Friston et al., 2024). Some of the novel forms of analogue or neuromorphic computations, targeted at LLMs, utilise continuous-time dynamical processes and fine-grained timing relations (e.g., Schmidt et al., 2023). Mortal computation was proposed by Hinton (2022) with industrial applications in mind. Embodied, embedded, and quasi-agentic robots are already being constructed.

A biological naturalist theory of consciousness based on any of those choices is bound to bestow suitably designed AI systems with consciousness. And if substrate-dependent multiscale activities, generative entrenchment, or self-maintenance are of an evolutionary advantage, perhaps they are of a practical advantage in building AI systems as well. It is hard, if not impossible, to conceive of non-computational functional properties that cannot, theoretically, be implemented in synthetic systems that we would not otherwise conceive of as life.

This puts a biological naturalist research programme at a dilemma. It either has to operate with hypotheses that may all too easily attribute life to near-future artificial or synthetic systems, or it is relegated to a gap-filling strategy, where it is forced to invoke precisely those properties of life that are not, at a particular time, implemented in artificial systems.

This raises the question of whether it is possible to provide an explicit definition of Biological Naturalism that circumvents this dilemma: a definition that does not hinge on undefined notions such as life, but also avoids concrete hypotheses that are only contingently unique to living systems.

In the remainder of this comment, I would like to propose one way to obtain such a definition, based on a generalisation of Putnam's original definition of Computational Functionalism (Putnam, 1967).

Crucially, Putnam's definition of Computational Functionalism is based on a "Description" (ibid., p. 54) of an organism in terms of Probabilistic Automata. It consists of four assumptions that explicate the idea that "being capable of feeling pain *is* possessing an appropriate kind of Functional Organization" (ibid.)—the Functional Organization specified by the Probabilistic Automaton Description.

Putnam makes use of Probabilistic Automata because they are the most general formal description of the type of digital general-purpose computations available at his time, Turing Machines with probabilistic transitions. However, from a mathematical perspective, nothing hinges on this choice of mathematical object. We can obtain a generalisation of Computational Functionalism by allowing for *any* description of an organism in terms of a mathematical object. Referring to such a description as "Formal Description", this generalisation is:

Def. 1. *F*-Functionalism* is true iff:

1. Every organism capable of feeling pain possesses at least one Formal Description of a certain kind (i.e., being capable of feeling pain is possessing the kind of Organization described by the Formal Description).
2. No organism capable of feeling pain possesses a decomposition into parts which separately possess Formal Descriptions of the kind referred to in 1.
3. For every Formal Description of the kind referred to in 1, there exists a subset of the sensory inputs such that an organism with that Formal Description is in pain when and only when some of its sensory inputs are in that subset.

Because Probabilistic Automata are one type of Formal Description, *F*-Functionalism* provides a generalisation of Computational Functionalism. It also encompasses other forms of functionalism, to the extent that the structure of scientific theories can be represented in formal terms. And it does justice to novel developments in the computing industry, ranging from quantum computation to neuromorphic computation, that break the confines of Turing-type computation. The intuition is that being capable of feeling pain *is* possessing the kind of Organization described by the Formal Description. And equally so for other experiences.

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

Based on *F*-Functionalism*, Biological Naturalism can be defined in two different ways. A first strategy would be to invoke Formal Descriptions that involve mathematical objects that cannot be built—and hence not artificially or synthetically created. This is, essentially, endorsing the first horn of the dilemma mentioned above. In light of progress in modern physics, and in light of the surprisingly vast range of descriptors of computational systems (cf. e.g., (Aczel, 1988)), this strategy might not provide an easy option.

A second, easier, way is to define Biological Naturalism simply as the absence of a Formal Description as required by *F*-Functionalism,* as follows:

Def. 2. *Biological Naturalism* is true iff there is no Formal Description that satisfies the three requirements of Definition 1.

Definition 2 singles out consciously experiencing organisms—conscious life—as systems that do not adhere to one common Formal Description. To paraphrase Seth, instead of a rigid organisation, where consciousness is concerned, "there is an excitable fluidity to the matter of life". To the extent that the production of artificial or synthetic systems requires a formal description of the system that is produced, Definition 2 precludes artificial and synthetic systems from being conscious in virtue of what they are. Whatever makes something conscious, it is not a matter of having the right kind of Formal Description.

Acknowledgements. I would like to thank Tim Ludwig, Lenore Blum, and Hanna Tolle for many insightful discussions on the topic of Computational Functionalism and AI consciousness.

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Competing interests. None.

References

Aczel, P. (1988). *Non-well-founded sets*. CSLI Lecture Notes, Number 14, CSLI Publications.

Friston, K. J., Ramstead, M. J., Kiefer, A. B., Tschantz, A., Buckley, C. L., Albarracin, M., … & René, G. (2024). Designing ecosystems of intelligence from first principles. *Collective Intelligence*, *3*(1), 26339137231222481.

Hinton, G. (2022). The forward-forward algorithm: Some preliminary investigations. arXiv:2212.13345.

Putnam, H. (1967). Psychological predicates. In W. H. Capitan & D. D. Merrill (Eds.), *Art, mind, and religion*, University of Pittsburgh Press. (Reprinted in (Putnam, 1975). Page numbers reference the reprinted version.)

Putnam, H. (1975). The nature of mental states. In *Mind, language, and reality: Philosophical papers* (Vol. ii), Cambridge University Press.

Schmidt, H., Montes, J., Grübl, A., Güttler, M., Husmann, D., Ilmberger, J., … & Schmitt, S. (2023). From clean room to machine room: Commissioning of the first-generation BrainScaleS wafer-scale neuromorphic system. *Neuromorphic Computing and Engineering*, *3*(3), 034013.

A link between consciousness and the wetware of metabolism

Nick Lane*

Director, Centre for Life's Origins and Evolution, Department of Genetics, Evolution and Environment, University College London, London, UK
nick.lane@ucl.ac.uk
https://profiles.ucl.ac.uk/11719-nick-lane

*Corresponding author.

doi:10.1017/S0140525X25103658, e342

Abstract

Seth provides a nuanced but compelling case for consciousness being substrate dependent: a product of living processes rather than computation. Here, I offer a biological underpinning for these arguments grounded in the wetware of metabolism in single-celled organisms. Fluctuations in membrane potential give an integrated real-time readout of metabolic state in relation to the environment – the simplest meaningful 'feeling'.

Biology is a pragmatic discipline. Like Justice Potter Stewart, who knew it when he saw it, most biologists eschew precise definitions of life, or species, or consciousness. In each case, we are dealing with multiple contributing factors and continua across arbitrary boundaries. From this pragmatic point of view, most biologists would agree that (i) consciousness evolved by natural selection; (ii) feelings are a central facet of consciousness so they must reflect a physical state that selection could act upon; (iii) feelings must offer some benefit, i.e., a selectable advantage to organisms that would be lacking in the absence of the feeling; and (iv) feelings are *about* the state of an organism in relation to its environment.

Rather than questioning the philosophical veracity of each of these statements, I want to take them seriously on their own terms. To reframe: feelings are physical states that reflect how well an organism is doing and are tangibly different from any other form of feedback. There is nothing here that speaks to panpsychism (Koch, 2021); feelings are strictly a property of living systems. Nor does this speak to any need for a neural network. Some neural networks are accompanied by feelings, others are not (Hohwy & Seth, 2020). Are neural nets the wrong level of analysis, with the simplest forms of consciousness being widespread across living things, including bacteria (Reber et al., 2023)? Anaesthesia of single-celled organisms implies a basic mechanism (Turin et al., 2014). If so, what kind of physical state do feelings reflect?

The only distinction between a living cell that responds to its environment and a cell that has just died is metabolism. This word conceals an unceasing storm of reactions: a billion per second,

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

second after second, in a single bacterial cell (Lane, 2022). How do bacteria coordinate these reactions, enabling coherent behaviour? Through a handful of thermodynamic drivers such as NADH (reduced nicotinamide adenine dinucleotide) and ATP (adenosine triphosphate), which coerce reactions into equivalent far-from-equilibrium states, reflecting the availability of food and energy (Lane, 2022). But there is a problem. The more NADH or ATP in a cell, the more robust but less sensitive it will be to changes in the environment.

Automated feedback loops, as in chemotaxis, do not help much. They are laboriously statistical and may take minutes to gauge large pool sizes of ATP or NADH (Lane, 2006). Multiple feedback loops can conflict with each other (Hamadeh et al., 2011). What should a cell do if one system reports just enough oxygen, another says there is plenty of food, a third registers a metabolic toxin, and a fourth implies it's a bit too hot? The cell's decision must depend in part on its own state: how easily can it deal with these challenges? If a feeling is *about* an organism's state in relation to its environment, how does a cell know what state it is in, or what it has in reserve, to make a better-than-random decision? If it moves somewhere else, how does it know it is better off now? It needs a real-time, integrated readout of its overall metabolic state, bounded by its own homeostatic limits. Luckily, it has one – fluctuations in membrane potential (Lane & Rodriguez, 2025).

All cells rely on an electrical membrane potential to drive work, notably ATP synthesis. Our own cells transduce energy in mitochondria, once free-living bacteria, so membrane bioenergetics is internalised. Bacteria energise their plasma membrane, the boundary (or 'Markov blanket') separating internal from external states (Palacio et al., 2020). The environment (e.g., food and oxygen) drives catabolism, which generates membrane potential, which drives anabolism and costly behaviour, notably motility. This means that membrane potential reflects the overall metabolic demands of the cell in relation to its environment – how well it is doing. A healthy cell has a modestly high membrane potential, maintained by the continuous transfer of electrons to oxygen. Deviations from this homeostatic optimum reflect perturbations in metabolism. Membrane potential is an electrostatic field, while the movement of charge generates electromagnetic fields; the balance between these forces could in principle indicate whether the cause of perturbation is primarily environmental (e.g., oxygen availability) or cellular (damage or mutations) (Lane & Rodriguez, 2025). Challenges such as viral infection are literally short-circuited by collapsing membrane potential, where bacteria commit suicide for even loosely related kin (Refardt et al., 2013).

In other words, membrane potential provides a real-time, systems-level output that integrates the cell's metabolic state – how well it is doing right now, in relation to its environment, genetic resources, and life history. Feelings are electromagnetic fields generated by membrane potential – a physical state (Lane & Rodriguez, 2025). That's not to say all fields correspond to feelings. A cell has a specific topology, and charge moving in relation to that topology holds meaning for the cell as an entity – a self. Fields with the wrong orientation, strength, or oscillation frequency hold no meaning for the cell or for selection over generations. Fields across metals or through space generally lack meaning in relation to a living metabolic entity.

Feelings as fields confer selectable advantage: they can be acted on, for example, by moving, before other systems-level changes such as metabolic regulation or gene expression obscure the current state. If a cell is in trouble and curbs the danger by changing gene expression, its position may be improved, but it no longer knows the magnitude of risk or its capacity to cope. Feelings are then a physical early-warning system, which alerts organisms to metabolic issues, enabling a rapid response: a selectable advantage that is lacking in the absence of the feeling. That makes feelings meaningful, physical, evolvable, and tangibly different.

The use of electromagnetic fields to signal metabolic health has obviously been elaborated in the evolution of complex animals, but fields are unequivocally involved in how the brain works (McFadden, 2023). We know little about their meaning. But Seth is right: consciousness is not fundamentally about computation. We feel the goodness (or badness) of metabolism, whirling in water at billions of reactions a second. We share that with bacteria, but not with Grok.

Acknowledgements. I thank Anil Seth, Luca Turin, and Enrique Rodriguez for stimulating discussions on metabolism in relation to consciousness.

Financial support. This work was supported by the Gates Foundation INV-068658.

Competing interests. The author declares none.

References

Hamadeh, A., Roberts, M.A., August, E., McSharry, P.E., Maini, P.K., Armitage, J.P., & Papachristodoulou, A. (2011). Feedback control architecture and the bacterial chemotaxis network. *PLoS Computational Biology*, *7*(5), e1001130. https://doi.org/10.1371/journal.pcbi.1001130

Hohwy, J., & Seth, A. (2020). Predictive processing as a systematic basis for identifying the neural correlates of consciousness. *Philosophy and the Mind Sciences*, *1*(II), 1–34. https://doi.org/10.33735/phimisci.2020.II.64

Koch, C. (2021). Reflections of a natural scientist on Panpsychism. *Journal of Consciousness Studies*, *28*(9–10), 65–75. https://doi.org/10.53765/20512201.28.9.065.

Lane, N. (2022). *Transformer: The deep chemistry of life and death*. WW Norton & Company.

Lane, N. (2006). Mitochondria: Key to complexity. In W. Martin, & M. Muller (Eds.), *Origins of mitochondria and hydrogenosomes* (pp. 13–38). Springer.

Lane, N., & Rodriguez, E. (2025). The physical basis of feelings. *PsyArXiv Preprints*, 1–32. https://doi.org/10.31234/osf.io/pnqw9_v1

McFadden, J. (2023). Consciousness: Matter or EMF? *Frontiers in Human Neuroscience*, *16*, 1024934. https://doi.org/10.3389/fnhum.2022.1024934

Palacios, E. R., Razi, A., Parr, T., Kirchhoff, M., & Friston, K. (2020). On Markov blankets and hierarchical self-organisation. *Journal of Theoretical Biology*, *486*, 110089. https://doi.org/10.1016/j.jtbi.2019.110089

Reber, A. S., Miller, W. B. Jr., Slijepcevic, P., & Baluška, F. (2024). The CBC theory and its entailments: Why current models of the origin of consciousness fail. *EMBO Reports*, *25*(1), 8–12. https://doi.org/10.1038/s44319-023-00004-6

Refardt, D., Bergmiller, T., & Kümmerli, R. (2013). Altruism can evolve when relatedness is low: evidence from bacteria committing suicide upon phage infection. *Proceedings of the Royal Society B Biological Sciences*, *280*(1759), 20123035. https://doi.org/10.1098/rspb.2012.3035.

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

Turin, L., Skoulakis, E. M., & Horsfield, A. P. (2014). Electron spin changes during general anesthesia in *Drosophila*. *Proceedings of the National Academy of Sciences U.S.A*, *111*(34), E3524–E3533. https://doi.org/10.1073/pnas.1404387111.

For consciousness, what's the right kind of magic?

Shane Legg*

Co-founder and Chief AGI Scientist, Google DeepMind, London, UK
shanelegg@google.com

*Corresponding author.

doi:10.1017/S0140525X25103865, e343

Abstract

Is efficient intelligence enough to produce consciousness, or do we need a system to be in some way "alive" in order for it to be conscious? While I don't know the answer, I suspect the former is likely to be closer to the mark.

When it comes to consciousness, I have read some books, talked to various researchers, and listened to a fair number of talks on the topic. Nevertheless, I remain confused in a deep way. But then, perhaps we all are? With these caveats noted, I'll now share a few naive thoughts on this paper.

Humans are intelligent and conscious. Some other animals, such as dogs, are less intelligent and seem to have some consciousness. Rocks appear to lack both of these things. It might then seem reasonable to conclude that more intelligence is associated with, or perhaps even implies, more consciousness. As animals face various kinds of constraints, such as energy budgets and physical space limitations, we probably should replace a purely performance-based notion of "intelligence" with some notion of "efficient intelligence." If one roughly accepts this line of reasoning, then we should expect machines to become conscious as they become increasingly intelligent and do so in increasingly efficient ways. In other words, just as evolution pushed us to develop efficient intelligence to thrive in our ecological niche, which in turn led to our consciousness, we'll push our machines to have increasingly efficient intelligence as this is valuable to us, and in the process, we'll push them to also become conscious.

If I've understood him correctly, Anil Seth doesn't buy this argument. He believes that some systems, such as computers, could have highly efficient intelligence without being conscious. He doesn't know exactly what the missing ingredient is, but he believes that it's somehow connected to the system "being alive." I find this a strange thought. To my mind, a robot is "alive," even if it can't do many of the things that animals can typically do. Furthermore, we can easily imagine a robot that could plug itself in to recharge, and perhaps be trained to fix broken parts of its body, and perhaps even build new robotics like itself, among other things that make it more fundamentally similar to living biological organisms. Could these additional capabilities make the robot conscious? This seems fishy to me, given that they seem not to make various very simple living organisms conscious. Nor do any of them have, I don't know, the right kind of "magic" that I'd expect to be needed to produce conscious experiences?

Financial support. No specific funding, other than being employed by Google.

Competing interests. None.

Sandwiched by meaning: The limits of computation for consciousness

Alexander Lerchner*

Google DeepMind, Google UK Ltd, London, UK
lerchner@google.com
https://scholar.google.com/citations?hl=en&user=wBxAt8cAAAAJ

*Corresponding author.

doi:10.1017/S0140525X2510366X, e344

Abstract

Computation is a description sandwiched by meaning that is supplied by an external observer. This observer-dependence limits computation to a mere simulation of mind, not a realisation. I argue that this makes computational functionalism logically untenable. Consciousness, as a real phenomenon, must instead be grounded in a system with intrinsic meaning – a living, teleodynamic organism.

Anil Seth's target article argues that conscious AI is unlikely on current computational trajectories and that a form of biological naturalism offers a more plausible foundation for consciousness. I agree, and in this commentary I aim to reinforce his conclusion by arguing that the link between his anti-computationalist and pro-biological stances is one of logical necessity.

The brain-as-computer metaphor is indeed an "increasingly inconvenient fiction" (sec. 3.6), and describing mental processes computationally is not enough to say that the brain is a computer. The metaphor fails for a fundamental reason that Seth himself briefly gestures towards (footnote 19): computation is observer-dependent.

Any physical system (such as a silicon chip) that performs a computation does so only because an external observer has created a mapping between its physical states and a set of symbols, where the meaning (semantics) of those symbols is supplied by the observer, and not given by the physics of that system. Crucially, this leaves the observer free to apply an arbitrary number of different interpretations of what computation the system performs. Thus, the question "What specific computation does a given physical system perform?" is ill-posed and underdetermined without further context and knowledge about computation-external intent and purpose.

For all the cases in which we have clear answers, it is because we have knowledge about the computational intent of the designer. As an observer, we can then match our interpretation to the designer's

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

intent, which requires an agreement that exists outside of, and separate from, the computation performed. Concretely, for all current AI systems on digital computers, this involves two separate steps that ensure convenient, substrate-independent computation of arbitrary algorithms. First, there is an agreement of how to map the physical states of a silicon chip to a set of symbols (such as sequences of binary numbers). This is the part linking the hardware to the software. Second, there is an agreement on what those symbols mean for a given scenario, that is, whether and how these numbers are interpreted in terms of video, text, robot actuators, or something else. This is the part linking the software to an understanding what its output *means* with respect to the real world. Put differently, computation is a form of description sandwiched between an encoding step (from meaning to symbols and then to physical states of the computing device) and a decoding step (from the physical states of the device back to symbols and ultimately back to meaning).

If computation is logically divorced from semantics and meaning, then where does meaning come from in the first place? Who are these observers that can supply meaning? Terrence Deacon addresses this question in his book *Incomplete Nature* (Deacon, 2011), where he distinguishes between the observer-relative information in a computer and the *intrinsic information* that arises from *teleodynamic* systems – systems organised around an intrinsic goal. Only living systems, which Seth correctly identifies in sec 4.1 as autopoietic (self-producing and self-maintaining), have such an intrinsic goal: to continue existing. This grounds information and meaning for the system itself without requiring an external observer. This principle provides a deeper reason for why, as Seth concludes in sec 3.3, it is hard "to separate what they do from what they are" in biological systems.

In sec 3.7, Seth correctly states, "Simulating something . . . is not equivalent to realising (instantiating) that thing." We can now see why this gap is unbridgeable for consciousness. Because computation is an observer-relative symbolic description, a computational model of a mind can only ever be a *simulation*. It cannot *realise* consciousness because it lacks the teleodynamic organisation required to ground intrinsic meaning.

The observer dependence of computation reveals an even deeper and more fundamental problem with computational functionalism than is hinted at in footnote 19. It is not just that computation is observer-dependent while consciousness is not. More fundamentally, the observer dependence makes it logically inconsistent to assert that the act of performing a computation can *ever*, by itself, create a specific real-world phenomenon – whether that phenomenon is consciousness or any other aspect of reality. This presents the computational functionalist with a sharp dilemma: either consciousness is not a real-world phenomenon (a stance that functionalism, with its insistence on physical realisation, was developed to oppose), or it simply cannot be created by an act of computation. This directly strikes at the heart of computational functionalism, which Seth defines as the view that computation is sufficient for consciousness (sec. 3.1).

If computation is out, what is left? We are forced to look for systems with intrinsic, non-observer-dependent properties. This leads us directly to the very features Seth highlights in his argument for biological naturalism: allostasis, metabolism, and the fundamental biological imperative to stay alive (sec. 4.1–4.2). These are not just interesting features of life; they are the very source of the intrinsic goal-directedness that allows a physical system to mean something to itself.

In summary, by following the observer-dependent nature of computation to its logical conclusion, the rejection of computational functionalism is no longer just a critique; it becomes a positive argument *for* biological naturalism. Consequently, rejecting the computer metaphor and embracing our nature as "living, flesh-and-blood 'beast machines'" (sec. 4.5) are not just two sequential moves in an argument, but a single, unified conclusion. This is why I strongly second Seth's closing sentiment (sec. 7): "A clear view of the biological basis of mind is essential to avoid overestimating our machines and, crucially, underestimating ourselves".

Competing interests. Alexander Lerchner is a full-time employee of Google DeepMind.

This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

References

Deacon, T. W. (2011). *Incomplete nature: How mind emerged from matter*. W. W. Norton & Company.

Diverse intelligence: An essential field for questions of AI and consciousness

Michael Levin*

Allen Discovery Center at Tufts University, Medford, MA, USA
michael.levin@tufts.edu
https://www.drmichaellevin.org/

*Corresponding author.

doi:10.1017/S0140525X25103828, e345

Abstract

Seth's insightful discussion must be extended in light of the field of Diverse Intelligence. A broader understanding of embodiment in unconventional spaces, of hybrids between evolved and engineered materials, and of unexpected competencies in very minimal systems makes it currently impossible place convincing limits on the possible substrates of consciousness. Doing so risks underestimating not only "machines" but also ourselves.

A critical context is the emerging field of Diverse Intelligence (Lyon, 2006; Reber & Baluska, 2021), seeking to understand what all agents, regardless of their origin story or composition, have in common. The neural replacement argument is only the tip of the iceberg, given the interoperability of life: cyborgs, hybrots, biobots, AI-mediated interfaces to collective intelligence of somatic cells, and all manner of products of synthetic bioengineering are already sharing our world. Should a being who is a coherent, functional integration of both natural, engineered, and software components

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

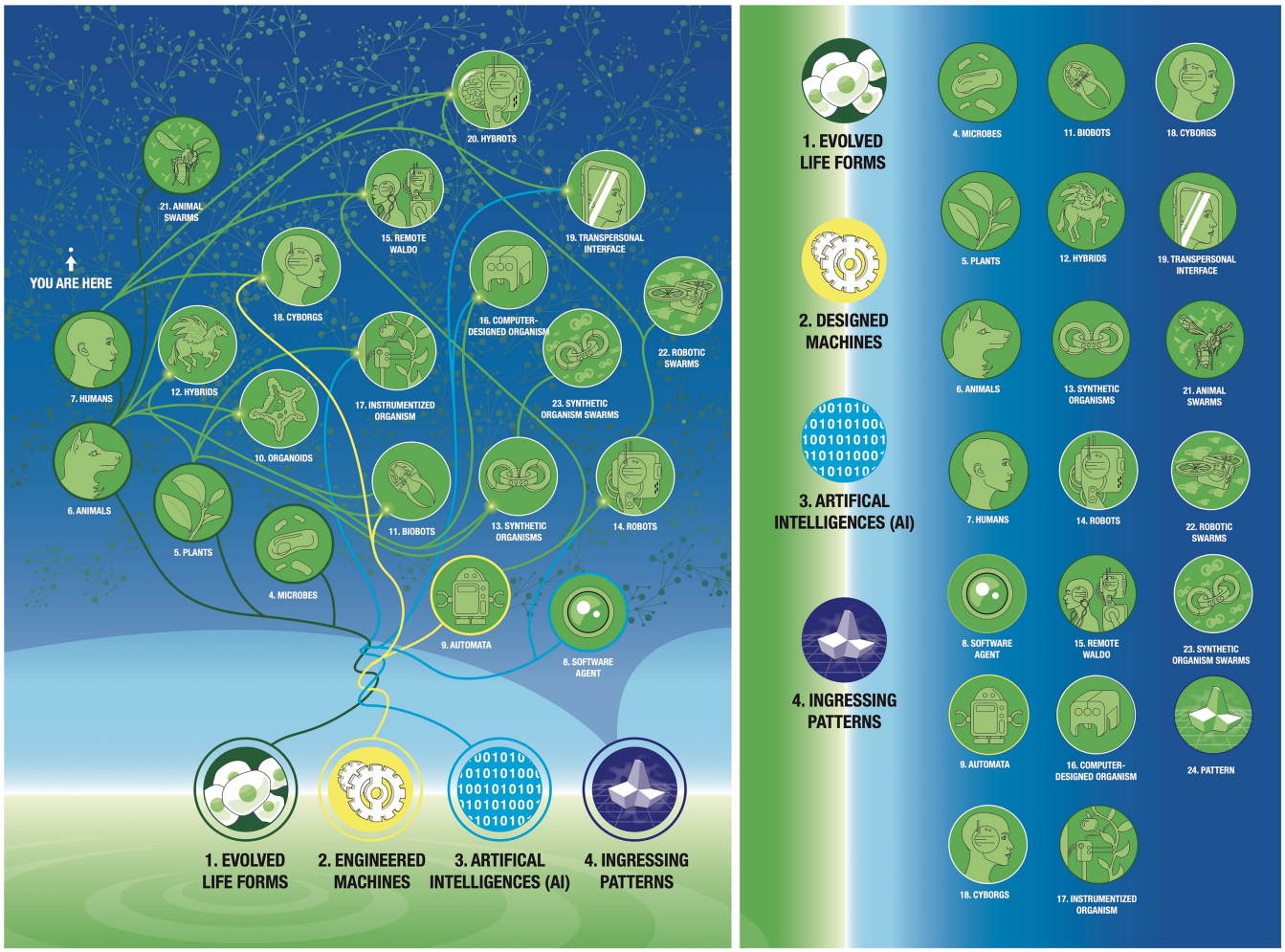


Figure 1 (Levin). The space of possible embodied minds is truly vast, breaking ancient linguistic categories of "living being" vs. "machine." Because of the interoperability of life, almost any combination of evolved, engineered, and hybrid materials can support forms of embodiment and behavior that must be taken seriously as possible conscious kin. Our pre-scientific heuristics for recognizing other minds must be significantly expanded to enable ethical synthbiosis with beings who are not like us, as humanity matures. Image by Jeremy Guay of Peregrine Creative.

be subjected to some sort of purity test seeking to neatly classify them (with all its legal and ethical implications) as "real human" vs. "machine"? When there is no one to speak out for beings embodied in unconventional materials, there will eventually be no one left to speak out for all of us, who are embodied in "proteins that are just doing what proteins do."

Advances in bioengineering mean that colloquial linguistic categories of "life" vs. "machine" are doomed (Clawson & Levin, 2023; Levin, 2022), and "life" as unique substrate for consciousness is questionable given that we do not *fundamentally* understand what the secret sauce is that is supposed to separate the living from the non-living (Bender et al., 2025). It's not just about evolutionary transitions: each of us has made the gradual, continuous journey from a quiescent bag of chemicals (the unfertilized oocyte) to a conscious adult who will make statements about not being a machine. Developmental biology offers *no* bright line between being fully describable by chemistry and being fodder for behavioral science, psychoanalysis, and love. Synthetic, cell-free systems, and active matter (Cejkova et al., 2017; Kaygisiz & Ulijn, 2025) systems will surely merge with efforts to make AI. The gradualist lessons of modern developmental biology support not magical binary categories, but the need to understand metamorphosis, scaling, and self-assembly of systems with significant inner perspective from minimal components.

We are unprepared to recognize intelligence, never mind consciousness, in systems of unfamiliar spatial and temporal scales or in unconventional substrates (Fields & Levin, 2022). We can begin to address our native mind-blindness with data on basal cognition in cells, tissues, plants, etc. (Baluška & Levin, 2016; Reber & Baluska, 2021), as the use of cognitive science tools to understand and manipulate their function is already leading to empirical advances in regenerative medicine and engineering. Groups of cells implement a collective intelligence that forms memories, pursues goals, makes decisions, and solves novel problems in creative and ingenious ways (Levin, 2025b; McMillen & Levin, 2024) - a stepping-stone to more alien cases. The electrophysiological mechanisms underlying cell intelligence are the same (indeed, evolutionarily homologous) to those that seem to underpin consciousness in human bodies) (Fields, Bischof, & Levin, 2020). But cells perform their sensing-deciding-acting loop in the high-dimensional spaces of the following: gene expression, physiological states, and anatomical morphospace. They pursue goals, avoid

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

obstacles, and deal with unexpected perturbations with various degrees of ingenuity and success (Levin, 2023). They sense, strive, and perhaps suffer in these other spaces largely invisible to our obsession with 3D space (though it is as constructed by our cognitive autopoiesis as the others).

We are just beginning to adjust our intuitions (and vocabulary) to accommodate the discovery that intelligence, in its essence, is basal and far older than neurons. "Embodiment" does not mean a motile robotic body: organoids and software AI's, *looking to us* as un-embodied ("not engaging with the real world, and thus not binding symbols" etc.) may well be navigating spaces we don't recognize. Embodiment is critical for minds, but that does not mean only in 3D space. Finally, all of the criteria commonly used to assume consciousness with each other (problem of other minds) – similar composition, common evolutionary origin, behavioral competencies that suggest valence, and causal architectures – now reveal that we absolutely must take seriously the likelihood that our body structures have their own (non-verbal) consciousness (far beyond the right hemisphere in split brain patients). Their consciousness is inaccessible to the main linguistic Self as that of other humans. If we cannot expand our detection of consciousness to this close test case, any discussion of consciousness in AI is hopeless.

I agree with Seth that computation does not suffice for consciousness. While many find it plausible that biochemistry does not tell the full story of the mind, it is universally believed that algorithms and materials tell the full story of "machines." But, in synthetic beings, as well as in evolved ones, our formal models *never* tell the full story of what something is or what it can and cannot do. Even humble Bubble Sort has competencies (not just complexity or unpredictability, but properties recognizable by behavioral scientists) and pursues unexpected side-quests not mentioned in its short, transparent, deterministic algorithm (Zhang, Goldstein, & Levin, 2025). Simple sets of coupled differential equations can do several different kinds of learning (Biswas et al., 2021). All of these were not recognized despite decades of study, until they were looked at through the diverse intelligence lens (Manicka & Levin, 2019). If even these truly minimal systems can do that, how many competencies, in what spaces, perhaps having absolutely nothing to do with language, could be present in current and future AIs? These intrinsic motivations (things they do that are specifically *not* prescribed by their algorithm) are likely to have much to do with being substrates for consciousness, free will, etc. Perhaps there is no dead matter, no "machines" that are not somewhere on the spectrum of life and mind. Perhaps what we make as offspring, whether biological or engineered, are interfaces to patterns of form and behavior from a non-physical latent space that contains the familiar low-agency truths of mathematics, but also a very wide range of higher-agency forms we recognize as kinds of minds.

In summary, I think Seth raises key questions that are not about AI at all (Levin, 2025a). How can conscious, morally important minds exist in a world in which their embodiments are made of parts following rules (whether of biochemistry, electronics, algorithms, etc.), and whether anything in our world is *only* what our computational or other models say it is? I would complement Seth's powerful conclusion as follows. If we seek to buy our exceptionalism too cheaply, we underestimate not only "machines" but also ourselves.

Financial support. M.L. gratefully acknowledges the support of the ARO (under grant W911NF-23-1-0100), AFOSR (under grant FA9550-22-1-0465), and JTF (under grant 62212).

Competing interests. M.L. is a minor shareholder and an advisor to Softmax and Astonishing labs, companies working in the space of Artificial Intelligence.

References

Baluška, F., & Levin, M. (2016). On having no head: Cognition throughout biological systems. *Frontiers in Psychology*, *7*, 902. https://doi.org/10.3389/fpsyg.2016.00902

Bender, R., Kofman, K., Agüera y Arcas, B., & Levin, M. (2025). What lives? A meta-analysis of diverse opinions on the definition of life. arXiv:2505.15849. Retrieved May 01, 2025, from https://ui.adsabs.harvard.edu/abs/2025arXiv250515849B

Biswas, S., Manicka, S., Hoel, E., & Levin, M. (2021). Gene regulatory networks exhibit several kinds of memory: Quantification of memory in biological and random transcriptional networks. *iScience*, *24*(3), 102131. https://doi.org/10.1016/j.isci.2021.102131

Cejkova, J., Banno, T., Hanczyc, M. M., & Stepanek, F. (2017). Droplets as liquid robots. *Artificial Life*, *23*(4), 528–549. https://doi.org/10.1162/ARTL_a_00243

Clawson, W. P., & Levin, M. (2023). Endless forms most beautiful 2.0: Teleonomy and the bioengineering of chimaeric and synthetic organisms. *Biological Journal of the Linnean Society*, *139*, 457–486. https://doi.org/10.1093/biolinnean/blac073

Fields, C., Bischof, J., & Levin, M. (2020). Morphological coordination: A common ancestral function unifying neural and non-neural signaling. *Physiology*, *35*(1), 16–30. https://doi.org/10.1152/physiol.00027.2019

Fields, C., & Levin, M. (2022). Competency in navigating arbitrary spaces as an invariant for analyzing cognition in diverse embodiments. *Entropy (Basel)*, *24*(6), 819. https://doi.org/10.3390/e24060819

Kaygisiz, K., & Ulijn, R. V. (2025). Can molecular systems learn? *Chem Systems Chem*, *7*, e202400075. https://doi.org/10.1002/syst.202400075

Levin, M. (2022). Technological approach to mind everywhere: An experimentally-grounded framework for understanding diverse bodies and minds. *Frontiers in systems neuroscience*, *16*, 768201. https://doi.org/10.3389/fnsys.2022.768201

Levin, M. (2023). Bioelectric networks: The cognitive glue enabling evolutionary scaling from physiology to mind. *Animal cognition*, *26*, 1865–1891. https://doi.org/10.1007/s10071-023-01780-3

Levin, M. (2025a). Ingressing minds: Causal patterns beyond genetics and environment in natural, synthetic, and hybrid embodiments. preprint. https://doi.org/10.31234/osf.io/5g2xj_v3

Levin, M. (2025b). The multiscale wisdom of the body: Collective intelligence as a tractable interface for next-generation biomedicine. *BioEssays*, *47*, e202400196. https://doi.org/10.1002/bies.202400196

Lyon, P. (2006). The biogenic approach to cognition. *Cognitive processing*, *7*(1), 11–29. https://doi.org/10.1007/s10339-005-0016-8

Manicka, S., & Levin, M. (2019). The cognitive lens: A primer on conceptual tools for analysing information processing in developmental and regenerative morphogenesis. *Philos Trans R Soc Lond B Biol Sci*, *374*(1774), 20180369. https://doi.org/10.1098/rstb.2018.0369

McMillen, P., & Levin, M. (2024). Collective intelligence: A unifying concept for integrating biology across scales and substrates. *Commun Biol*, *7*(1), 378. https://doi.org/10.1038/s42003-024-06037-4

Reber, A. S., & Baluska, F. (2021). Cognition in some surprising places. *Biochem Biophys Res Commun*, *564*, 150–157. https://doi.org/10.1016/j.bbrc.2020.08.115

Zhang, T., Goldstein, A., & Levin, M. (2025). Classical sorting algorithms as a model of morphogenesis: Self-sorting arrays reveal unexpected competencies in a minimal model of basal intelligence. *Adaptive Behavior*, *33*(1), 25–54. https://doi.org/10.1177/10597123241269740

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

Consciousness as an ontological primitive

Iain McGilchrist*

All Souls College, Oxford, UK
iain@iainmcgilchrist.com
https://channelmcgilchrist.com/home/

*Corresponding author.

doi:10.1017/S0140525X25103506, e346

Abstract

Distinctions can be made between information-processing, different kinds of knowledge, and understanding (intelligence). AI may be incapable of the type of understanding/intelligence known to humans. On the origins of consciousness, panpsychism may be a more rationally satisfying explanation than emergence. Some brief reflections are included on the relation between matter and consciousness, and on hemisphere differences and AI.

Anil Seth's review of the currently salient theories of the origins of consciousness, and their implications for the question of whether AI could become conscious, is to be welcomed. However, none of the theories of the origin of consciousness discussed departs from the conventional view that – somehow – consciousness emerges from mere matter. More than a century and a half of speculation has not produced a single viable explanation of how something as "inward" as experience emerges from wholly non-experiential matter.

A familiar distinction between two kinds of knowing is relevant here: "knowing that" – e.g., I know that Paris is the capital of France – and "knowing by experience" – e.g., I know Paris, because I spent several years living there. So divergent are these that in most languages other than English, they are represented by different words (e.g., F *savoir* v *connaître*, G *wissen* v *kennen*). The first takes place at the level of mere information-processing and is the kind of knowing that AI may be said to have. But this is not in itself intelligence. The second implies understanding, and with it intelligence: a complex *Gestalt* of relational, contextual, embodied, and implicit elements that come from experience, and are bound up with value. Can we really *understand* a painting or a landscape if we do not see its beauty? Do we really understand an action dispassionately, as though its nature did not involve ethical judgements reflecting the character of the doer? Can understanding take place without a body and its emotions, without suffering and the foreknowledge of death? The element "under-" in understanding does not mean "below," but "among" or "between," as "*unter*" does in German: thus, "within a broad context." Our word "intelligence" comes from *inter-legere* in Latin, with the same implication that one is able to see the connexions among disparate kinds of awareness, "read between" the lines, so as to truly comprehend.

The word "panpsychism" does not occur in the body of Seth's text, despite the fact that this is now probably the most widely espoused position on consciousness in contemporary Western philosophy (to say nothing of non-Western traditions). Panpsychism posits that consciousness is an ontological primitive, viz, that its being cannot be derived from anything else. According to this account, matter is never wholly devoid of an aspect of consciousness. Indeed, I argue that matter is a phase of consciousness (McGilchrist, 2021), in the same sense that water has phases with remarkably different qualities: liquid, transparent and flowing; solid, opaque and immotile; and invisibly dispersed in air, without which we could not breathe. So with consciousness and matter. No one has ever seen matter; all that has been encountered are aspects of conscious experience which we choose to call "material," by which we mean that they offer two essential qualities for creativity not given by consciousness as normally conceived: resistance and persistence.

Clearly panpsychism does not imply that a rock – or a silicon chip – has the type of consciousness that a human being possesses, but it suggests that by being part of a large conscious organism, possibly as broad as the cosmos itself, it has experience of a kind; and that experience can become increasingly individuated under certain circumstances, and out of all recognition in that found in living organisms, and eventually man. Arguably, this may be one of the reasons that life itself evolved: to create awareness of, and a capacity to respond to, the values and directionality bound up in the creative cosmos.

I suggest that with this view, which has a claim to be at least as reasonable as the view that consciousness is an unexplained brute emergence from *lumpen* matter, the problem of AI becomes recast. Does AI have a different status from some massively complicated mechanical system, such as a modern jet engine, which has consciousness only at the same level as a pop-up toaster, by virtue of being made of matter, or does it enter a new realm because of its engineering? Complex systems are naturally different from merely complicated systems, but I would argue that we are seduced into attributing far greater consciousness to AI by the fact that we have developed machines that produce utterances in language, and that if they did not, we would be much less likely to be seduced – a point, among others, on which Seth and I agree.

I would briefly add a couple of reflections from research on hemisphere differences. While it is not true that only the left hemisphere is involved in the understanding of language, it remains true that language production is more left hemisphere-dependent. In the left hemisphere, language maps the world in the tokens we call words. In the right hemisphere, language is not just a network of signs, but at least partly embodied experience, with attention to all-important context, the implicit, metaphor, emotion, and bodily sensations; which is why the right hemisphere excels at understanding humour, irony, poetry, and narrative, all of which are relatively lost on the left hemisphere. It should be a matter of concern that through AI, we are increasing the pervasiveness of the left hemisphere's mode of thought in a society that is already moving towards an autistic reliance on explicitly statable propositions, such as a machine may understand, and an incapacity for all that transcends such simplicity.

It is interesting, in relation to language use, that AI clearly confabulates (i.e., makes up plausible "facts" that are entirely false but just happen to fit the expectations of the system). This, along with much else (see McGilchrist, 2009, 2021), is typical of the left hemisphere of the brain, which is concerned with internal consistency, rather than truth to experience. It is this emphasis on consistency within a system, sealed off from experience, that leads me to the supposition that, while it cannot be excluded that AI

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

might develop consciousness, or has already (we can never know), the consciousness it might develop would always lack what we most highly value in our own.

Acknowledgments. None.

Financial support. None.

Competing interests. The authors declare no competing interests.

References

McGilchrist, I. (2009). *The Master and his Emissary: The Divided Brain and the Making of the Western World*, Yale University Press, Newhaven CT.
McGilchrist, I. (2021). *The Matter with Things: Our Brains, Our Delusions and the Unmaking of the World*, Perspectiva Press, London UK.

Biological naturalism, synthetic phenomenology, and the physics of awareness

Thomas Metzinger*

Johannes-Gutenberg-Universität, Mainz, Germany
metzinger@uni-mainz.de
https://thomasmetzinger.com

*Corresponding author.

doi:10.1017/S0140525X25103695, e347

Abstract

Biological naturalism may collapse into eliminative materialism, and the ethics of synthetic phenomenology will remain.

Let us make two assumptions:

1. Deep continuity (DC). Anil Seth is fundamentally right about the substrate-dependence of consciousness relative to patterns of microfunctional properties that are a) "non-computational" in terms of classical functionalism and b) "biological" in terms of a general theory of life. The assumed theory of life isn't yet fully spelled out, but it will likely include Karl Friston's free energy principle, plus what the late Margaret Boden called the "metabolism constraint," some variant of Maturana & Varela's conception of autopoiesis, Hohwy-style self-evidencing (Hohwy, 2026), and so on.
2. Self-modelling epistemic spaces (SES). Living systems are embodied epistemic spaces. An epistemic space is a set of possible states or processes of knowing. In and of itself, this space does not possess a center or a periphery, and conscious experience appears whenever a living system additionally generates an integrated *model* of this very space within itself. If and only if a given system has a global, internal model of its own epistemic space as such does the system also know that it currently possesses the capacity to know. This is what awareness is. From Seth's perspective, tonic alertness and the ascending reticular activation system will be among the central biological properties that create this capacity as such. In terms of predictive processing, the system can then use this model to continually predict and actively control not just the current level of mere epistemic capacity, but also the specifics of *what* it will know (e.g., via precision control). Physically, the level of arousal will be the decisive boundary condition, causally determining the depth of cortical information processing available to the animal, depending on the local level of activation of five types of neurotransmitter in the ascending reticular activating system. Conceptually, however, the essence of phenomenality – the extra quality of awareness – is simply "being the content of a currently self-modeling epistemic space." Let us assume SES as a (very) rough sketch of our philosophical background theory of consciousness (Metzinger, 2020, 2024, Chs. 4 & 5).

If both theoretical intuitions are on the right track, then we can draw at least two conclusions from the conjunction of DC and SES:

- We still don't know *how* deep the relevant continuity is. If phenomenal properties locally supervene on contemporaneous biological but non-computational properties, then it still remains unclear what the correct level of description and the correct level of conceptual granularity for the supervenience base level really is. In a given class of biological systems, what kind of predicates best describe exactly those properties that form the minimally sufficient base set for consciousness? For example, could implementing a good caricature of the essential statistical physics be enough to conjure up the target phenomenon, to create a self-modelling epistemic space? If we let go of the computational level, then maybe we can ultimately do without biological properties too. The emergence of SES systems may be an extremely interesting and cosmologically rare physical phenomenon that is most economically described and most parsimoniously explained using *only* conceptual tools from physics – and none from biology. Under computational functionalism, phenomenal properties would still be something abstract, as defined by their causal role in a wider network, but Seth's proposal would turn them into very concrete, substrate-dependent features of real-world beast machines. Supervenience as such is a non-reductive form of materialism, but biological naturalism may "bottom out" into physics via local, domain-specific identities between concrete biological properties and the physical realizations of even finer-grained microfunctional properties. Seth writes: "Life would still matter, and it would do so thanks to the properties that warrant the term 'life'" (§ 5.5). But is this really true? Biological naturalism may function as the methodological gateway to eliminative materialism (Ramsey, 2024), ultimately leading us to conclude that consciousness and phenomenal properties never really existed in the first place. Just as biological properties never did.
- For the ethics of synthetic phenomenology (§ 6.1), not only artificial intelligence (AI) but also advanced organoid intelligence (OI) will be relevant (de Jongh et al., 2022; Smirnova et al., 2023). If Anil Seth is right, then phenomenal properties will most likely be instantiated by some future form of OI or by novel combinations that causally connect AI with OI (for example, via OI systems *trained* by AI). Brain organoids could represent the ultimate form of neuromorphic computing (Friston, 2023), and such systems could certainly instantiate self-modelling epistemic spaces. However,

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

conscious SES systems of this type would not be "embodied" or "embedded" in the canonical sense and would exhibit only a weaker version of "generative entrenchment" (see § 3.3). Even more importantly, they would be *neither* artificial *nor* natural, because they are not exclusively human-made artifacts, nor are they connected to the evolutionary history of our planet via the standard processes of procreation and genetic transmission. The term "artificial" implicitly introduces "natural" as its semantic counterpart (as in Seth's "biological naturalism"), but from a purely logical point of view, the distinction between artificial and natural systems is neither exclusive nor exhaustive. There is a third logical possibility: non-artificial, non-biological, but *postbiotic* phenomenology. This is why the term "synthetic phenomenology" is preferable to "artificial consciousness" because it leaves open this third possibility while directly connoting the already well-established discipline of synthetic biology, which applies engineering principles to develop entirely new kinds of biological devices and systems. Even moving beyond "biological naturalism," the ethics of synthetic phenomenology will not go away: not all, but certain subsets of *postbiotic* SES-systems will possess the capacity to suffer and thereby become objects of moral consideration.

Acknowledgement. I want to thank Dr. Emily Troscianko for editorial help with this contribution.

Financial support. The author declares that he has no competing financial interests.

Competing interests. The author declares that he has no conflicts of interest.

References

de Jongh, D., Massey, E. K., VANGUARD consortium, & Bunnik, E. M. (2022). Organoids: A systematic review of ethical issues. *Stem Cell Research & Therapy*, *13*, 337. https://doi.org/10.1186/s13287-022-02950-9

Friston, K. (2023). The sentient organoid?. *Frontiers in Science*, *1*, 1147911. https://doi.org/10.3389/fsci.2023.1147911

Hohwy, J. (2026). *The self-evidencing agent*. MIT Press.

Metzinger, T. (2020). Self-modeling epistemic spaces and the contraction principle. *Cognitive Neuropsychology*, *37*(3–4), 197–201. https://doi.org/10.1080/02643294.2020.1729110

Metzinger, T. (2024). *The elephant and the blind*. MIT Press.

Ramsey, W. (2024). Eliminative materialism. In E. N. Zalta, & U. Nodelman (Eds.), *The stanford encyclopedia of philosophy* (Winter 2024 Edition). Stanford University. https://plato.stanford.edu/archives/win2024/entries/materialism-eliminative/

Smirnova, L., Caffo, B. S., Gracias, D. H., Huang, Q., Morales Pantoja, I. E., Tang, B., Zack, D. J., Berlinicke, C. A., Boyd, J. L., Harris, T. D., Johnson, E. C., Kagan, B. J., Kahn, J., Muotri, A. R., Paulhamus, B. L., Schwamborn, J. C., Plotkin, J., Szalay, A. S., Vogelstein, J. T., Worley, P. F., Hartung, T. (2023). Organoid intelligence (OI): The new frontier in biocomputing and intelligence in-a-dish. *Frontiers in Science*, *1*, 1017235. https://doi.org/10.3389/fsci.2023.1017235

Bet on functionalism

Matthias Michel*

Department of Linguistics & Philosophy, Massachusetts Institute of Technology, Cambridge, MA, USA
mmichel@mit.edu

*Corresponding author.

doi:10.1017/S0140525X25103816, e348

Abstract

I argue that there is currently no alternative to functionalism. Functionalism explains the differences between types of mental states. No biological theory does. Functionalist theories account for the differences between conscious and unconscious states. No biological theory does. So, as things stand we should bet on functionalism.

Seth holds that consciousness only emerges when predictive processing is embedded in systems that regulate themselves to stay alive. That means no consciousness in non-biological systems. I am about to argue that functionalism should be the default view and that there is currently no alternative to functionalism. That means no reason to believe that biology is necessary for consciousness.

A theory of consciousness ought to distinguish between conscious and unconscious states. Seth's positive proposal does not meet this minimal criterion. There is no evidence that the machinery he proposes is specifically operative in conscious states, as opposed to unconscious states, or in conscious organisms, as opposed to unconscious ones.

Start with free energy minimization. Granted, conscious organisms might have a knack for minimizing free energy. But so do sponges, bacteria and yeast. Free energy minimization is so ubiquitous that it would be shocking if unconscious mental processes weren't in on the game too. The same thing goes for predictive coding. Even if the brain is a predictive engine and conscious states result from predictive coding, there is no evidence that unconscious mental processes do not follow similar principles. Nothing here properly distinguishes conscious from unconscious states.

The biological processes by which we regulate our living bodies are even worse in that respect. Seth writes that "we experience the world, and the self, with, through, and because of our living bodies." No doubt. But we also digest, keep breathing under general anesthesia, maintain posture, fight infections, regulate body temperature, and sleep dreamlessly through, and because of our living bodies. Biology is full of unconscious goings-on. So there's no reason to link consciousness to autopoiesis, and some reason not to.

Of course, Seth doesn't *need* to give us a biological property that distinguishes conscious from unconscious states. All he needs is to

make the case for the necessity of biology. So, could biology be necessary?

For better or worse, computational functionalism is a view in metaphysics. The relevant kind of necessity is metaphysical necessity. Seth's arguments do not establish the metaphysical necessity of biology and thus are harmless against computational functionalism.

Think about currency. A thing can't be currency unless it is divisible, portable, durable, and in limited supply. Suppose only gold meets those conditions. Gold would, as a matter of fact, be the only currency. Yet, "being a currency" could still be identical to having the right functional role. If some other stuff had done the job of being a unit of account, store of value, and means of exchange, it would have been a currency. *Being a currency* could be identical to a functional role, even if only gold fulfills that role.

Back to consciousness. The question is: if a non-biological system were to perform the same functions as a conscious biological one, would it have conscious experiences? Seth argues that, *as a matter of fact*, non-biological systems cannot perform the same functions as conscious biological systems. Maybe so, but this does not answer the question. The difference between conscious and unconscious states could still be a matter of functional roles, even if only biological systems can perform the relevant functions.

Are we at a standstill? I don't think so. Because there is currently no alternative explanatory framework that can hold a candle to computational functionalism. Let me give two reasons why.

First, functionalism wins on induction. When we explain other psychological capacities – mental arithmetic, language production, causal inference, memory retrieval, etc. – we do it functionally. We appeal to computations over mental representations, not special biological properties. Mental kinds are individuated by functional roles (Fodor, 1974; Weiskopf, 2011), not by properties of the neurobiological kinds that realize them. By induction, we should expect that the difference between conscious and unconscious mental states is also a matter of functional role. That is, we should expect that differences in mental kinds are functional differences through and through, including the difference between conscious and unconscious states. Of course, consciousness *might* be the exception. But as Damon Runyon put it, paraphrasing Ecclesiastes: "The race is not always to the swift, nor the battle to the strong, but that's how the smart money bets." And that's partly why we should bet on functionalism.

Second, consider a variety of consciousness-related phenomena: object-substitution masking, blindsight, binocular rivalry, change blindness, inattentional blindness, Troxler fading, visual suppression with transcranial magnetic stimulation to V1, the attentional blink, motion-induced blindness, and so on. These phenomena are of interest for consciousness researchers because they involve (or at least they're supposed to involve) consciously perceiving a stimulus at one time and unconsciously registering it at another. A theory of consciousness ought to explain why. Here again, the only contenders are computational functionalist theories: global workspace theory (Mashour et al., 2020), higher-order theories (Brown, Lau, & LeDoux, 2019), and so on. No biological theory on the market explains these contrasts by appealing to some property unique to living systems. Seth might say predictive processing can do the job. Maybe it can. But predictive processing is defined computationally. So once again, functional role and computation do the heavy explanatory lifting, not biology.

As things stand, our best theory of mental kinds and our best theories of the differences between conscious and unconscious states make no appeal to biological properties. That being the case, we have no reason to believe that biology matters except insofar as it constrains the implementation of the right functional roles. Maybe someday we'll discover a biological property that fits the bill – one that explains the whole range of relevant phenomena. That day, I'll happily convert. Until then, the smart money is on functionalism, for the same reason that you bet on the swift and the strong. They don't always win, but they usually do.

Acknowledgments. I thank Ali Rezaei.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interests. None.

References

Brown, R., Lau, H., & LeDoux, J. E. (2019). Understanding the higher-order approach to consciousness. *Trends in Cognitive Sciences*, *23*(9), 754–768.

Fodor, J. A. (1974). Special sciences (Or: The disunity of science as a working hypothesis). *Synthese*, *28*(2), 97–115.

Mashour, G. A., Roelfsema, P., Changeux, J., & Dehaene, S. (2020). Conscious processing and the global neuronal workspace hypothesis. *Neuron*, *105*(5), 776–798.

Weiskopf, D. A. (2011). The functional unity of special science kinds. *British Journal for the Philosophy of Science*, *62*(2), 233–258.

Consciousness needs a subject

Kevin J. Mitchell[a]* and Carolyn Dicey Jennings[b]

[a]Institutes of Genetics and Neuroscience, Trinity College Dublin, Dublin, Ireland and [b]Philosophy Department, University of California Merced, USA
kevin.mitchell@tcd.ie
cjennings3@ucmerced.edu
https://www.kjmitchell.com
https://faculty.ucmerced.edu/cjennings3/

*Corresponding author.

doi:10.1017/S0140525X25103452, e349

Abstract

To be conscious is to be an experiencing subject. This can be defined not in terms of computational functions or particular biological substrates, but rather in terms of relations: between subject and world, between parts of the subject, and through time. These kinds of relations – comprising a conscious mode of being – may well be implementable in artificial systems.

The target article by Seth argues against certain forms of computational functionalism and makes a strong case that current approaches to artificial intelligence will not be able to create conscious systems. Seth advocates biological naturalism instead, emphasizing the importance of embedded, embodied agency and arguing that consciousness may arise from our nature as living beings with physical bodies. In this, we are in agreement. Yet, a critical piece of Seth's argument is missing, without which his conclusions are seriously weakened: the nature and role of the conscious *subject*.

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

Seth's arguments depend on the self – he says conscious experience occurs for "self-sustaining organisms" that maintain a "dynamic self/other boundary" – but what he means by "self" is unclear. It does not appear to be a mere organism, since he rejects biological universalism. It appears conceptually distinct from a mere body, since he says we experience the self "through, and because of" the body. The closest description seems to be "living, flesh and blood 'beast machines.'" A problem with this description is it treats the self as object, rather than subject, but the critical form of self with respect to consciousness is surely the subject (Prinz, 2011).

A subject is a living being with interests (e.g., Kawade, 2009; Jennings, 2020; Mitchell, 2023). A subject in this sense cannot be simply tacked on to Seth's current view. Motivations and interests are not predictions, and subjects are not Markov blankets; we don't currently have a way to map one onto the other. Seth says that "real artificial consciousness might endow a system with its own interests," but it seems to us that subjects with interests are conceptually (and evolutionarily) prior to consciousness. Similarly, he says "conscious contents are given by the brain's best guesses" but does not say to whom they are given. Without an account of the self as subject, it is unclear what features of our biological system are essential for consciousness and whether those features require organic material.

Subjects like us are "enminded," and not just embodied. Our conscious experiences are grounded in our own ongoing mentality – in what it's like to be *us*, psychologically. This encompasses all of our memories, attitudes, dispositions, habits, policies, projects, and interests that make us who we are. When we experience something consciously, we are experiencing *being ourselves having that experience* (Zahavi, 2008). While we can treat ourselves as objects by taking an objective stance, subjectivity is fundamentally about an irreducible first-person stance, and this is what is central to consciousness (Ganeri, 2012).

In our view, phenomenal subjectivity rests not on specific computational functions or particular biological processes or substrates, but rather on *relations*. These include relations between (i) subject and world, (ii) levels of information processing within the subject, and (iii) "temporal parts" of the subject (past and future selves).

Unlike most current AI systems, a conscious agent must be able to both perceive the world around it *and act on it*. Without this interactive relation, all that exists is passive and abstract information processing (Dewey, 1896). By contrast, in living organisms, perception is highly selective and entails actively making sense of what is out in the world in a way that allows the organism to figure out what it should do about it (Mitchell, 2023). Indeed, there is good evidence that conscious perception is necessarily structured through the activity of attention (e.g., Jennings, 2015).

These capacities are underpinned by neural and cognitive architectures that can also be defined in terms of internal functional relations. With respect to consciousness, these involve a functional control hierarchy, with higher levels operating over representations of (beliefs about) objects in the world, internal states, motivations, moods, goals, and other elements of cognition. These kinds of relations can be described (and formalized, e.g., in the tradition of cybernetics) in abstract terms, without relying on either specific computational implementations or bespoke biological substrates.

The question then is *what kinds of relations* will go beyond cybernetic functions and give rise to or entail subjective experience? The answer, of course, is not known, but we can still ask what form we expect it to take. Some candidate elements of subjectivity include the organism being able to distinguish itself from the world (Jennings, 2020), being able to track its own internal states (Nikolova et al., 2022), having valenced internal signals that guide action (Godfrey-Smith, 2020; Cleeremans & Tallon-Baudry, 2022), being able to track the consequences of its actions (Jékely, Godfrey-Smith, & Keijzer, 2021), making predictions about itself and the world (as Seth notes), and having a "phenomenal self-model" (Metzinger, 2004). Conscious subjects should have internal representations that stand in exploitable relation to things in the world (Shea, 2018), but also some sort of reflexivity that imbues these representations with "*mineness.*"

Often overlooked in these discussions, however, is that living organisms have *temporal extension*. We live in and through time and cannot be reduced, as selves, to a single temporal moment (Husserl 1964; Mitchell, 2023). All living organisms are beings that (literally) *incorporate* traces of their past experience to better propagate themselves into the future (Potter & Mitchell, 2025). In more complex organisms, selfhood is defined by continuity through time of both the physical and the psychological – of *us*, as persons. We bring our personal history to bear in our interactions with the world. We are actively seeking, filtering, predicting, and interpreting sensory information in inherently subjective, goal-oriented ways.

Thus, rather than drawing comparisons and contrasts between biological and artificial systems, we can see them both as particular instantiations of a more abstract set of relational principles. We can zoom out and view consciousness as a "mode of being" (Ginsburg & Jablonka, 2020); specifically, of being a subject and a self. There seems little reason in principle to think that the relations underpinning subjectivity and selfhood could not be operationalized and implemented in many different ways in many different substrates, including artificial entities. Given the right mode of being in the world, it might very well be like something to be such a system.

Financial support. This work did not involve or rely on any funding.

Competing interests. The authors declare no competing interests.

References

Cleeremans, A., & Tallon-Baudry, C. (2022). Consciousness matters: Phenomenal experience has functional value. *Neuroscience of consciousness*, *2022*(1), niac007.

Damasio, A., & Damasio, H. (2023). Feelings are the source of consciousness. *Neural Computation*, *35*(3), 277–286.

Dewey, J. (1896). The reflex arc concept in psychology. *Psychological review*, *3*(4), 357.

Ganeri, J. (2012). *The self: Naturalism, consciousness, and the first-person stance*. Oxford University Press.

Ginsburg, S., & Jablonka, E. (2020). *Consciousness as a mode of being. Journal of Consciousness Studies*, *27*(9–10), 148–162.

Godfrey-Smith, P. (2020). Gradualism and the evolution of experience. *Philosophical topics*, *48*(1), 201–220.

Husserl, E. (1964). *The phenomenology of internal time-consciousness*. James S. Churchill (Trans.). Indiana University Press.

Jékely, G., Godfrey-Smith, P., & Keijzer, F. (2021). Reafference and the origin of the self in early nervous system evolution. *Philosophical Transactions of the Royal Society B*, *376*(1821), 20190764.

Jennings, C. D. (2015). Attention and perceptual organization. *Philosophical Studies*, *172*(5), 1265–1278.

Jennings, C. (2020). *The Attending Mind*. Cambridge University Press.

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

Kawade, Y. (2009). On the nature of the subjectivity of living things. *Biosemiotics*, *2*(2), 205–220.
Metzinger, T. (2004) The subjectivity of subjective experience: A representationalist analysis of the first-person perspective. *Networks*, *3–4*, 33–64.
Mitchell, K.J. (2023). *Free agents – how evolution gave us free will*. Princeton University Press.
Nikolova, N., Waade, P. T., Friston, K. J., & Allen, M. (2022). What might interoceptive inference reveal about consciousness?. *Review of Philosophy and Psychology*, *13*(4), 879–906.
Potter, H. D., & Mitchell, K. J. (2025). Beyond mechanism – extending our concepts of causation in neuroscience. *European Journal of Neuroscience*, *61*(5), e70064.
Prinz, J. (2011). Wittgenstein and the neuroscience of the self. *American Philosophical Quarterly*, *48*(2), 147–160.
Shea, N. (2018). *Representation in cognitive science*. Oxford University Press.
Zahavi, D. (2008). *Subjectivity and selfhood: Investigating the first-person perspective*. MIT Press.

Absolute units: The metabolic basis of the experiential self

Kathryn Nave*

School of Philosophy, Psychology, and Language Science, The University of Edinburgh, Edinburgh, UK
Kathryn.nave1@gmail.com

*Corresponding author.

doi:10.1017/S0140525X25103622, e350

Abstract

Unity is intrinsic to being both an experiencing and an embodied subject. Computational accounts of consciousness, being observer-relative, cannot account for this unity. The thermodynamics of life can, but, contra Seth's suggestion, the free energy principle fails to capture the distinctive metabolic interdependencies by virtue of which a living system constitutes itself as a unified perspective on the world.

A central explanandum of consciousness is its unity (Hurley, 1998). An individual may hold multiple contradictory beliefs, and a room may hold multiple conscious individuals, but one individual cannot experience multiple different perspectives at a time. What is it about consciousness that forces the disparate information currents streaming through our sensory apparatus into this single coherent take on the world?

Each theory has its own answer, from the higher-order processing of lower-level states to competition between informational subunits for global broadcast. To evaluate consciousness in non-neural candidates, like AI, these criteria are typically specified in terms of medium-independent computation. Formalized as such, all face the same interconnected problems: 1) gerrymandering, an observer may arbitrarily carve up any physical system so that it satisfies the criteria (Shevlin, 2021) and 2) indeterminacy, an observer is free to carve up the *same* physical system in different ways, some that will meet said criteria and some that will not (Papayannopoulos et al., 2022).

On pain of regress, the unity of one conscious self cannot be accounted for in terms of how things appear to another, for that second perceiver's own experiential unity equally demands explanation. We need, as Kant argues, an objective basis for the unity of an experiencing subject. This issue of objective unity is a central problematic throughout Kant's philosophy and arises also in his introduction of the term "self-organization" to describe how the parts of an organism are unified by mutual interdependence with the whole toward the intrinsic purpose of continued self-production (Kant, 2013/1790).

Oddly, Kant did not connect these unities. If, however, metabolic unity underpins the unity of a conscious self, then that would entail biological naturalism, for the complex artifacts that trouble distinctions between the animate and inanimate – for Kant, pocket watches; for us, computers – are not self-producing. These appear as unified objects *within* our experience, but said unity depends on how they achieve *our* purposes – such as telling the time or generating undergraduate essays. Beyond our perspective, there's just excitations of quantum fields with only varying degrees of interaction strength to connect them. Different perceivers might divide things differently – one organism's circuit board is another (micro)organism's habitat.

The reason there's nothing it is *like* to be a pocket watch or an AI is because there is nothing it *is* to be a pocket watch or an AI, independent of how these objects are constructed by us as subjects. The cell, however, is bound together by more than our perception of it. It is unified by constructive interdependencies: between membranes, which create electrochemical gradients, which drive enzymatically aided reactions, which regenerate these membranes and enzymes in turn. If one part of this network breaks down, then the whole thing disintegrates. This interdependence is not relative to our purposes, but to the intrinsic instability of biological substrates (Thompson, 2010).

How does this metabolic unity extend through the multicellular activity of the brain? For Seth, the answer lies with free energy minimization, which Friston and collaborators present as an account of the contents of consciousness, the neural dynamics underpinning it, and biological self-organization – under the integrative formalism of predictive inference (Friston, 2013; Hohwy & Seth, 2020).

This seems like the account of experiential and organic unity we were looking for. Yet there's a problem for biological naturalism here. Contra earlier presentations, both critics (Nave, 2025; Di Paolo et al., 2022) and advocates (Ramstead et al., 2023) now argue that the FEP is *not* a unique principle of life but a general description of every "thing." This is because it reduces self-organization to the emergence of invariant statistical properties and defines survival as the preservation of these regularities. That would entail that, when encountering an "unlikely" state, an organism must return to a more "likely" one – a process that can be described as minimizing informational "free energy." Still, insofar as *any* physical object is defined by the invariance of certain features and is subject to environmental perturbations, so any object must do the same (Ashby, 1991).

Crucially, the FEP advances no ruling on which properties of the system are essential to its ongoing unity. All it allows is that *if* we've already identified these features, then we can re-describe their preservation in statistical terms. Without prior constraints on system individuation, this process is as vulnerable to gerrymandering and indeterminacy as any formal redescription.

What if, as Seth does, we focus on informational free energy only as it relates to the *thermodynamic* quantity, namely, energy

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

that is available for constrained release to power physical work? It's true that metabolic self-production involves reducing this by channeling the concentrated energy of high-frequency photons or food into the construction of membranes, enzymes, and other cellular structures. The faster this process, the faster thermodynamic free energy is lost via the inefficiency of heat dissipation (Martyushev & Seleznev, 2006). To minimize free energy entirely, however, is to die.

Thermodynamic equilibrium is the end, but not the point, of life. What makes a dissipative process a living one is not the universal tendency toward free energy reduction but rather the contingent specifics of what happens along the way – namely, whether that dissipation powers the production of an organized structure that can continue to channel this dissipation into the work of its own construction (Montévil & Mossio, 2015).

Viewing the brain thermodynamically means we cannot *just* treat it as a predictive control mechanism. Neurons may be long-lived, but, like all cells, they churn through components with turnover times in the order of minutes (Meagher, 2014). Ignoring this and taking their seemingly invariant structure for granted leads us to misinterpret phenomena like spontaneous activity as mere "noise" – an unfortunate biological residue interfering with essential cognitive dynamics that could be better implemented in an artificial machine. In contrast, viewing brains as precarious self-constructing networks where, as Seth suggests, we cannot separate what they *do* from what they *are,* frames this intrinsic instability as essential in holding our metabolic and experiential selves together.

How substrate-specific is this unity requirement? That's an open question, subject to investigation of whether alternative chemistries can satisfy the precise synchronization of exergonic and endergonic reactions that powers life (Grefenstette et al., 2024). But for all that synthetic or alien organisms *may* one day exhibit metabolic closure in alternative substrates, silicon semiconductors ain't got the stuff.

Acknowledgments. We thank Mariel Goddu and Dave Ward.

Financial support. This work was funded by a Leverhulme Trust Early Career Research Fellowship.

Competing interests. None.

Reference

Ashby, W. R. (1991). Principles of the Self-Organizing System. In G. J. Klir (Ed.), *Facets of Systems Science* (pp. 521–536). Springer US.

Friston, K. (2013). Life as we know it. *Journal of The Royal Society Interface*, *10*(86), 20130475.

Grefenstette, N., Chou, L., Colón-Santos, S., Fisher, T. M., Mierzejewski, V., Nural, C., Sinhadc, P., Vidaurri, M., Vincent, L., & Weng, M. M. (2024). Chapter 9: Life as we don't know it. *Astrobiology*, *24*(S1), S–186.

Hohwy, J., & Seth, A. (2020). Predictive processing as a systematic basis for identifying the neural correlates of consciousness. *PhiMiSci: Philosophy and the mind sciences*, *1*(2), 3.

Hurley, S. (1998). *Consciousness in action*. Harvard University Press.

Kant, I. (2013). *Critique of the power of judgment* (P. Guyer & E. Matthews, Trans.). Cambridge University Press.

Martyushev, L. M., & Seleznev, V. D. (2006). Maximum entropy production principle in physics, chemistry and biology. *Physics Reports*, *426*(1), 1–45.

Meagher, R. B. (2014). 'Memory and molecular turnover,' 30 years after inception. *Epigenetics & Chromatin*, *7*(37), 1–9.

Montévil, M., & Mossio, M. (2015). Biological organisation as closure of constraints. *Journal of Theoretical Biology*, *372*, 179–191.

Nave, K. (2025). *A drive to survive: The free energy principle and the meaning of life*. The MIT Press.

Di Paolo, E. D., Thompson, E., & Beer, R. (2022). Laying down a forking path: Tensions between enaction and the free energy principle. *Philosophy and the Mind Sciences*, *3*(2), 1–39.

Papayannopoulos, P., Fresco, N., & Shagrir, O. (2022). On two different kinds of computational indeterminacy. *The Monist*, *105*(2), 229–246.

Ramstead, M. J. D., Sakthivadivel, D. A. R., Heins, C., Koudahl, M., Millidge, B., Da Costa, L., Klein, B., & Friston, K. J. (2023). On Bayesian mechanics: A physics of and by beliefs. *Interface Focus*, *13*(3), 20220029.

Shevlin, H. (2021). Non-human consciousness and the specificity problem: A modest theoretical proposal. *Mind & Language*, *36*(2), 297–314.

Thompson, E. (2010). *Mind in life: Biology, phenomenology, and the sciences of mind*. Harvard University Press.

There and back again: Biological naturalism provides no new scientific progress on the matter of AI consciousness

Morten Storm Overgaard[a,b]*

[a]Department of Neurology, Aarhus University Hospital, Aarhus, Denmark and [b]CFIN, Department of Clinical Medicine, Aarhus University, Aarhus, Denmark
morten.storm.overgaard@cfin.au.dk

*Corresponding author.

doi:10.1017/S0140525X25103543, e351

Abstract

Seth presents several arguments for why multiple realization is not necessarily true. However, this is not a common opinion in consciousness research. He fails to provide arguments why we should believe in biological naturalism over multiple realization, which is the actual challenge to create progress.

In the current consciousness literature, there is no clear answer to the question of whether consciousness can be multiply realized, such that Artificial Intelligence (AI) potentially could be, or even already is, conscious (Bayne et al., 2024; Overgaard & Kirkeby-Hinrup, 2024; Kirkeby-Hinrup, Stenseke, & Overgaard, 2024).

Seth strongly argues that the question of conscious AI is both pressing in light of recent technological advancements and influenced by logical fallacies in our reasoning. He proposes that consciousness is fundamentally biological, best understood in terms of organisms that self-regulate and maintain their own existence. On this view, consciousness is "substrate sensitive" rather than multiply realized – meaning that biological matter and its specific organization are not incidental to consciousness but rather necessary for it.

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

However, the argument intended to make "biological naturalism" more plausible than the theory of multiple realization appears underdeveloped.

While most would agree with Seth that cognitive fallacies may bias us toward ascribing consciousness in certain situations (e.g., to AI) as a form of anthropomorphism, such bias does not in itself constitute an argument that the attribution is necessarily mistaken in any given case. The very same type of bias Seth refers to is also what leads us to ascribe consciousness to other people and animals – a judgment we must assume is not fallacious. Accordingly, even if cognitive biases influence our judgments about AI, this fact alone is neutral with regard to whether AI is actually conscious.

Seth argues that the functional/structural relations of the brain are fundamentally different from those in artificial systems. However, such a difference only serves as an argument against consciousness in AI if we presuppose that consciousness cannot be realized in a system with a different functional/structural organization – which is precisely the assumption Seth sets out to investigate. The argument for multiple realization is exactly that such differences do not preclude consciousness. Indeed, a proponent of multiple realization would use the same observation – that AI and human systems are functionally and structurally different – as the starting point for their case. This difference is not a counterargument, but the very premise upon which the theory is built.

Seth cites Chalmers' thought experiment "neural replacement," after which he refers to several authors who have criticized the idea. He concludes that while this critique does not disprove the possibility of silicon-based consciousness (i.e., consciousness realized in non-biological matter), it does challenge the assumption that such substrate flexibility must be true. He then asserts: "Brains seem to be the kind of thing for which it is hard and perhaps impossible to separate what they do from what they are." However, this conclusion does not clearly follow from the preceding discussion. The claim remains vague and appears more as an implicit appeal to identity theory than a substantiated argument.

In the end, Seth leaves us with a new name for an old idea – that consciousness depends on biology. He presents several reasons why one might consider this view and why multiple realization is not necessarily true. However, what is needed is an argument why multiple realization cannot be true, or why we should believe biological naturalism over multiple realization. As none of the arguments succeed in any of these goals, the scientific field of consciousness research is in the very same spot after Seth's presentation of his theoretical position as before.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interests. None.

References

Bayne, T., Seth, A., Massimini, M., Shephard, J., Cleeremans, A., Fleming, S., Raphael, J., Mattingley, J., Menon, D., Owen, A., Peters, M., Razi, A., & Mudrik, L. (2024). Tests for consciousness in humans and beyond. *Trends in Cognitive Sciences*, *28*(5), 454–466.

Kirkeby-Hinrup, A., Stenseke, J., & Overgaard, M. (2024). Evaluating the explanatory power of the Conscious Turing Machine. *Consciousness and Cognition*, *124*(103736), 1–12.

Overgaard, M., & Kirkeby-Hinrup, A. (2024). A clarification of the conditions under which Large Language Models could be conscious, *Humanities and Social Sciences Communications*, *11*(1031), 1–4

Information, energy, and exploration

Thomas Parr[a]* and John Broulidakis[b]

[a]Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, UK and [b]Department of Experimental Psychology, University of Oxford, Oxford, UK
thomas.parr@ndcn.ox.ac.uk
john.broulidakis@psy.ox.ac.uk

*Corresponding author.

doi:10.1017/S0140525X2510383X, e352

Abstract

This commentary picks up on two related themes in *Seth*'s article – which have roots in embodiment. The first is that perceptual inference has a metabolic cost, and it is important to consider the relationship between thermodynamic and informational energies. The second is that embodied systems actively select their data over time – which may be a core feature of biological consciousness.

We enjoyed reading *Seth's* thoughtful, and thought-provoking, article on the plausibility of conscious artificial intelligence. One of the key themes in his argument is the relationship between metabolic machinery in biological systems and the informational characterizations of autopoiesis. The relationship between information and energy is important in understanding perceptual awareness in terms of the substrate that is doing the "perceiving." This commentary focuses upon energies of an informational sort as a means of selecting between future paths or trajectories. We suggest that the sort of curious behavior that is characteristic of embodied biological creatures described in the article might be essential element of (human-like) consciousness.

Figure 1 links key quantities in perceptual inference and energetics. The core of this is Boltzmann's relation (Gibbs, 1902 (2010)). In short, energy can be thought of as a measure of improbability. The relationship between energy and probability lets us express the conditional dependency structures that define Markov blankets (Pearl, 1988), as a decomposition of the Hamiltonian (total energy) of a system into additive parts in which states internal and external to a blanket do not appear together in any term. This formalizes a boundary in which all interactions between internal and external take place via intermediate (blanket) states.

Dissipative systems are characterized by a tendency for their total energy to decrease – that is, to move from less probable to more probable configurations – over time, with energy lost as heat. This means the minima of their Hamiltonians are attracting points under averaged flows (Kwon, Ao, & Thouless, 2005). The internal states of a dissipative Markov blanketed system have attracting points determined by blanket, but not external, states (Friston, 2019). If we specify an arbitrary distribution (i.e., variational density) over external states parameterized by internal states (Parr, 2022), dissipative dynamics cause this distribution to evolve toward an attracting point in information space (Amari, 2016; Kim, 2018). From a neurobiological perspective, this is like saying a representation of the external world, parameterized by neural states within the brain, evolve toward some perceptual inference.

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

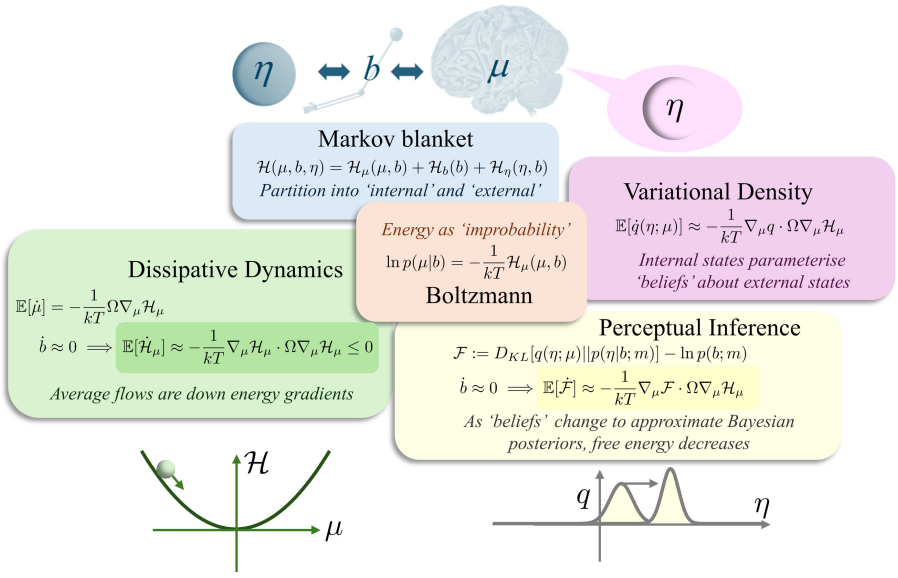


Figure 1 (Parr and Broulidakis). The Energetics of Inference. This figure brings together some of the key concepts linking perceptual inference to its thermodynamic cost. Boltzmann's relation links probabilities with energies (or Hamiltonians) via temperature (T) and a constant (k) that determines units. This relation lets us express conditional dependency structures by decomposing a Hamiltonian ($\mathcal{H}$) for a joint system into a sum of terms, each of which includes only subsets of the full system. If two parts (μ and η) do not appear in the same term, they are said to be separated by a Markov blanket (b). In neurobiology, one can think of brains being separated from the outside world by a blanket comprising sensory receptors and muscles. For Boltzmann's relation to hold over time, in the presence of random fluctuations, there must be a dissipative element to the dynamics, such that expected ($\mathbb{E}$) flows appear to descend the gradients of the Hamiltonian – counteracting the dispersive effects of fluctuations. If one assumes that internal (e.g., brain) states "represent" or parameterise some "variational" distribution (q) describing perceptions of external (world) states, then one can quantify how well those perceptions match optimal Bayesian inference based on a model (m) using variational free energy ($\mathcal{F}$), expressed here in terms of a KL-Divergence.

More technically, one can characterize the dynamics of inference in terms of (expected) rates of change of variational free energy – an informational quantity used in Bayesian statistics to score the fit of a model to data (MacKay, 1991; Penny, 2012) – by assuming an internal model for which the attracting points for the variational density approximates the Bayesian posterior density under that model. There is a clear similarity in the form of the expected changes in the (thermodynamic) Hamiltonian and (informational) free energy, with both depending upon the gradients of the Hamiltonian for internal states (see Figure 1). However, their rates of change are not identical. The dynamics of variational free energy depend on the internal model. Understanding this model – arguably, the core problem of neurobiology – may be essential in characterizing the metabolic (e.g., mitochondrial (Lane, 2020)) machinery that supports internal brain dynamics in terms of their awareness of the external world.

The above sketches a way to interpret the dynamics of states inside a brain as a set of beliefs about the external world (Wiese & Friston, 2021). To some extent, this is also true of many AI systems. Can we take this further? While most artificial systems wait for data (e.g., prompts), biological systems actively select their own. One can score the efficiency of this selection using information theory and Bayesian experimental design principles (Lindley, 1956). Specifically, the plausibility of a behavior that optimizes data for inference is the mutual information (Kreer, 1957; Shannon, 1948) – a.k.a., expected information gain – between blanket and external states. Given an internal model, selection of paths with high mutual information has the appearance of curiosity-driven behavior (Friston et al., 2012; Gottlieb et al., 2013; Itti & Baldi, 2009; Manohar & Husain, 2013; Mirza et al., 2016; Yang, Wolpert, & Lengyel, 2016). Mutual information is defined in terms of log probabilities, so has units suitable for an energy (or "action") for planning. Perhaps more importantly, it depends on the beliefs a creature holds about the environment and favors those paths that select data, not to optimize reward or value, but to optimize our beliefs about our world.

The implications of these ideas for consciousness research are as follows. First, the energetics of the physiological (or computational) system engaged in perception constrains the form of the internal "representation" or beliefs that might underwrite awareness of the external world. Second, as curiosity-driven behavior is a function of those representations – we must know what we are uncertain about to know which action to take to resolve that uncertainty – specific beliefs imply specific patterns of behavior (see (Parr & Friston, 2017) for an example in the context of saccadic sampling patterns in visual neglect). Taken together, behavior and biophysics provide complementary empirical constraints on the belief dynamics that might underwrite subjective awareness.

In summary, the above endorses Seth's point that the physiology of information must be closely integrated with metabolism in natural, biological, intelligence. Perceptual awareness of a simple sort depends on the dissipation of informational and thermodynamic energies that afford an interpretation of the internal parts of a system (e.g., neurons in a brain) as performing inferences about the external world. These inferences play an important role in planning. A core feature of natural intelligence, and perhaps consciousness, is that we move our sensory apparatus to select our own data. In doing so, we both optimize our beliefs to explain our data and optimize our data to improve our beliefs.

Financial support. TP is supported by an NIHR Academic Clinical Fellowship (ref: ACF-2023-13-013).

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

Competing interests. The authors declare no conflicts of interest.

References

Amari, S.-i. (2016). *Information geometry and its applications* (Vol. *194*). Springer.

Friston, K. (2019). A free energy principle for a particular physics. *arXiv e-prints*. Retrieved June 1, 2019, from https://ui.adsabs.harvard.edu/abs/2019arXiv190610184F

Friston, K., Adams, R., Perrinet, L., & Breakspear, M. (2012). Perceptions as hypotheses: Saccades as experiments [Original Research]. *Frontiers in Psychology*, *3*(151), 1–20. https://doi.org/10.3389/fpsyg.2012.00151

Gibbs, J. W. (1902 (2010)). On the distribution-in-phase called canonical, in which the index of probability is a linear function of the energy. In *Elementary principles in statistical mechanics: Developed with especial reference to the rational foundation of thermodynamics* (pp. 32–45). Cambridge University Press. https://doi.org/10.1017/CBO9780511686948.005

Gottlieb, J., Oudeyer, P.-Y., Lopes, M., & Baranes, A. (2013). Information-seeking, curiosity, and attention: Computational and neural mechanisms. *Trends in Cognitive Sciences*, *17*(11), 585–593. https://doi.org/10.1016/j.tics.2013.09.001

Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research*, *49*(10), 1295–1306. https://doi.org/10.1016/j.visres.2008.09.007

Kim, E.-j. (2018). Investigating information geometry in classical and quantum systems through information length. *Entropy*, *20*(8), 574. https://www.mdpi.com/1099-4300/20/8/574

Kreer, J. (1957). A question of terminology. *IRE Transactions on Information Theory*, *3*(3), 208–208. https://doi.org/10.1109/TIT.1957.1057418

Kwon, C., Ao, P., & Thouless, D. J. (2005). Structure of stochastic dynamics near fixed points. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(37), 13029. https://doi.org/10.1073/pnas.0506347102

Lane, N. (2020). How energy flow shapes cell evolution. *Current Biology*, *30*(10), R471–R476. https://doi.org/10.1016/j.cub.2020.03.055

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics, 27*(4), 986–1005. https://doi.org/10.1214/aoms/1177728069

MacKay, D. (1991). Bayesian model comparison and backprop nets. *Advances in neural information processing systems, 4,* 839–846.

Manohar, S. G., & Husain, M. (2013). Attention as foraging for information and value. *Frontiers in Human Neuroscience*, *7*, 711. Retrieved 2013, from https://www.frontiersin.org/articles/10.3389/fnhum.2013.00711/pdf

Mirza, M. B., Adams, R. A., Mathys, C. D., & Friston, K. J. (2016). Scene construction, visual foraging, and active inference [Hypothesis & Theory]. *Frontiers in Computational Neuroscience*, *10*(56), 1–16. https://doi.org/10.3389/fncom.2016.00056

Parr, T. (2022). Inferential dynamics: Comment on: How particular is the physics of the free energy principle? By Aguilera et al. *Physics of Life Reviews*, *42,* 1–3. https://doi.org/10.1016/j.plrev.2022.05.006

Parr, T., & Friston, K. J. (2017). The computational anatomy of visual neglect. *Cerebral Cortex*, *28*, 1–14. https://doi.org/10.1093/cercor/bhx316

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann.

Penny, W. D. (2012). Comparing dynamic causal models using AIC, BIC and free energy. *NeuroImage*, *59*(1), 319–330. https://doi.org/10.1016/j.neuroimage.2011.07.039

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Wiese, W., & Friston, K. J. (2021). Examining the continuity between life and mind: Is there a continuity between autopoietic intentionality and representationality?. *Philosophies*, *6*(1), 18. https://www.mdpi.com/2409-9287/6/1/18

Yang, S. C.-H., Wolpert, D. M., & Lengyel, M. (2016). Theoretical perspectives on active sensing. *Current Opinion in Behavioral Sciences*, *11*, 100–108. https://doi.org/10.1016/j.cobeha.2016.06.009

The neurobiophysical substrate of consciousness

Gualtiero Piccinini*

Department of Philosophy, University of Missouri, Columbia, MO, USA
piccininig@missouri.edu

*Corresponding author.

doi:10.1017/S0140525X25103749, e353

Abstract

I argue that creature consciousness depends on specific neurobiophysical properties that constitute the medium for phenomenally conscious states. Such properties are not medium flexible in the way required by computational functionalism. Therefore, computational functionalism fails, and one of the following holds: biological naturalism, a nonbiological yet noncomputational version of functionalism, or a view according to which consciousness requires specific (macro)physical qualities.

I will extend Seth's argument as follows:

1. Creature consciousness provides the medium for phenomenally conscious states.
2. The medium for phenomenally conscious states depends on specific neurobiophysical properties that are not medium flexible in the way required by computational functionalism.

Conclusion: Therefore, computational functionalism fails.

By "creature consciousness," I don't mean mere wakefulness or behavioral responsiveness to stimuli (cf. Rosenthal, 2005); I mean a range of global brain states such that, when an organism is in them, the organism instantiates phenomenally conscious states, whether during wakefulness, sleep, or any other state (cf. Piccinini, 2007).

By "medium" or "substrate" for phenomenally conscious states, I don't mean merely a background enabling condition without which the organism cannot function – whether consciously or unconsciously. Rather, I mean a *constitutive* enabling condition that helps make the difference between consciousness and unconsciousness (cf. McKilliam, 2020, 2024).

By "neurobiophysical properties," I mean some combination of neural, biological, chemical, and physical properties constitutively relevant to the substrate of conscious experiences. We don't know precisely what all such properties are, and they may not require being alive or even having a nervous system, but we know enough to conclude that they are unlikely to be medium flexible in the way required by computational functionalism.

The empirical evidence for the necessity of creature consciousness comes primarily from sleep research. Sleep researchers distinguish between wakefulness and four sleep stages, N1-N3 plus REM sleep, which exhibit different EEG (electroencephalogram) signatures. A full account would include states such as anesthesia and coma, but the most helpful states are relatively easy to awaken from to get reliable information on what it was like to be in them.

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

Different EEG signatures indicate different modes of neural oscillation and synchronization. The ability to be awakened shows that nervous systems retain some capacity for information processing during all phases of sleep. Yet wakefulness and different sleep stages allow very different sorts of conscious experiences. When awakened from REM sleep, people typically report vivid dreams; when awakened from N1 and N2, they typically report thoughts or sensory experiences; when awakened from N3, they typically report *no* experiences. This suggests that phenomenal consciousness is not binary but exists on a continuum that depends on the underlying type of creature consciousness, which in turn depends on the underlying neurobiophysical state. Humans can be conscious only when their brains occupy certain global states – primarily normal wakefulness and REM sleep, but also N1 and N2. During N3 – slow-wave, deep sleep – the capacity for consciousness appears to be absent (e.g., Hobson, 2009). This pattern suggests that whether humans can experience different kinds of phenomenally conscious states depends critically on the global state of their brain – what I call creature consciousness.

The neural correlates of consciousness research program (Koch et al., 2016) has traditionally focused on identifying the specific neural mechanisms and states underlying particular conscious experiences. The necessity of creature consciousness suggests we must also identify what might be called the "enabling conditions of consciousness" (ECCs) – the background global brain states that make any conscious experience possible.

The second component of my argument concerns the neurobiophysical specificity of creature consciousness. A combination of neurobiophysical properties provides the medium for phenomenal consciousness without determining specific experiences. The brainstem reticular activating system, for instance, plays a crucial role in maintaining arousal states that enable consciousness (Steriade, 2000; cf. Merker, 2007). Similarly, the intralaminar thalamic nuclei contribute to the substrate that supports conscious access while remaining largely content-neutral (Newman, 1995).

These systems' primary function involves transitioning between and sustaining global brain states either by bathing the brain in precise mixtures of specific neurotransmitters (including acetylcholine, norepinephrine, serotonin, histamine, and GABA) or by amplifying or dampening neural signals (Jones, 2005; Brown et al., 2012). Recent high-resolution neuroimaging studies confirm the importance of specific networks involving the brainstem, hypothalamus, basal forebrain, thalamic gating mechanisms, and cortical integration networks in sustaining the arousal states necessary for consciousness (Laureys, 2005; Shine et al., 2023; Edlow et al., 2024). Partly due to such arousal, large portions of brain tissue create their own rhythmic balance of excitation and inhibition and, as a result, oscillate and synchronize at different frequencies during different kinds of global states. Different patterns of oscillation and synchronization, in turn, contribute to different kinds of creature consciousness and play fundamental cognitive roles (e.g., Buzsáki & Draguhn, 2004). Creature consciousness thus appears to depend on a complex balance of neurological, molecular, and biophysical factors. The precise balance and interaction of these systems determine whether a brain can support conscious experience at all and what kind of conscious experience it can have.

These considerations undermine computational functionalism, the view that mental states are defined by their computational roles and are therefore realizable in any suitable computational medium (Piccinini, 2020, Ch. 14). If creature consciousness depends on specific neurobiophysical properties that constitute a specific medium for phenomenally conscious states, then conscious states are not medium-flexible, and computational functionalism fails as a theory of consciousness.

This suggests that one of three alternative views holds: Seth's biological naturalism, some form of nonbiological yet non-computational functionalism, or some version of what Anderson and Piccinini (2024) call qualitativism – the view that consciousness depends at least in part on specific (macro)physical qualities rather than purely functional or computational properties.

Rejecting computational functionalism does not require rejecting the role of neural computation in conscious experience (cf. Chirimuuta, 2024). Various forms of neural computation occur while organisms are creature conscious, and some of them seem to contribute to the content and structure of conscious experience. These computations appear to be enabled and constrained by the neurobiophysical substrate of creature consciousness rather than being sufficient for consciousness on their own.

In conclusion, by distinguishing between creature consciousness and phenomenal consciousness, we can articulate more precisely why consciousness appears to depend on specific neurobiophysical properties of living systems. This account suggests that genuine artificial consciousness would require not just more sophisticated computation but technologies that can recreate specific neurobiophysical conditions – a prospect that remains elusive given our currently limited understanding.

Acknowledgments. Thanks to Andy McKilliam and Eric Thomson for helpful comments on a draft of this essay.

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Competing interests. The author declares none.

References

Anderson, J. A., & Piccinini, G. (2024). *The physical signature of computation: A robust mapping account*. Oxford University Press.

Brown, R. E., Basheer, R., McKenna, J. T., Strecker, R. E., & McCarley, R. W. (2012). Control of sleep and wakefulness. *Physiological Reviews*, *92*(3), 1087–1187.

Buzsáki, G., & Draguhn, A. (2004). Neuronal oscillations in cortical networks. *Science*, *304*, 1926–1929. https://doi.org/10.1126/science.1099745

Chirimuuta, M. (2024). *The brain abstracted: simplification in the history and philosophy of neuroscience*. MIT Press.

Edlow, B. L., Olchanyi, M., Freeman, H. J., Li, J., Maffei, C., Snider, S. B., Zöllei, L., Iglesias, J. E., Augustinack, J., Bodien, Y. G., Haynes, R. L., Greve, D. N., Diamond, B. R., Stevens, A., Giacino, J. T., Destrieux, C., van der Kouwe, A., Brown, E. N., Folkerth, R. D., Fischl, B., & Kinney, H. C. (2024). Multimodal MRI reveals brainstem connections that sustain wakefulness in human consciousness. *Science Translational Medicine*, *16*(746), eadj4303.

Hobson, J. A. (2009). REM sleep and dreaming: Towards a theory of protoconsciousness. *Nature Reviews Neuroscience*, *10*(11), 803–813.

Jones, B. E. (2005). From waking to sleeping: Neuronal and chemical substrates. *Trends in Pharmacological Sciences*, *26*(11), 578–586.

Koch, C., Massimini, M., Boly, M., & Tononi, G. (2016). Neural correlates of consciousness: progress and problems. *Nature Reviews Neuroscience*, *17*, 307–321. https://doi.org/10.1038/nrn.2016.22

Laureys, S. (2005). The neural correlate of (un)awareness: Lessons from the vegetative state. *Trends in Cognitive Sciences*, *9*(12), 556–559.

McKilliam, A. (2020). What is a global state of consciousness?. *Philosophy and the Mind Sciences*, *1*(II), 7. https://doi.org/10.33735/phimisci.2020.II.58

McKilliam, A. (2024). A mechanistic alternative to minimal sufficiency as the guiding principle for NCC research. *Neuroscience of Consciousness*, *2024*(1), niae014. https://doi.org/10.1093/nc/niae014

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

Merker, B. (2007). Consciousness without a cerebral cortex: A challenge for neuroscience and medicine. *Behavioral and Brain Sciences*, *30*(1), 63–81.
Newman, J. (1995). Thalamic contributions to attention and consciousness. *Consciousness and Cognition*, *4*(2), 172–193.
Piccinini, G. (2007). The ontology of creature consciousness: A challenge for philosophy. *Behavioral and Brain Sciences*, *30*(1), 103–104.
Piccinini, G. (2020). *Neurocognitive mechanisms: Explaining biological cognition*. Oxford University Press.
Rosenthal, D. M. (2005). *Consciousness and mind*. Oxford University Press.
Shine, J. M., Lewis, L. D., Garrett, D. D., & Hwang, K. (2023). The impact of the human thalamus on brain-wide information processing. *Nature Review Neuroscience*, *24*, 416–430. https://doi.org/10.1038/s41583-023-00701-0
Steriade, M. (2000). Corticothalamic resonance, states of vigilance and mentation. *Neuroscience*, *101*(2), 243–276.

Computational neuroscience versus the Computational Theory of Mind: Why their difference matters for the possibility of artificial consciousness

Maxwell James Ramstead*

UCL Queen Square Institute of Neurology: University College London Queen Square, London, UK
maxwell.d.ramstead@gmail.com

*Corresponding author.

doi:10.1017/S0140525X25103580, e354

Abstract

This commentary concerns the distinction between computational modelling and computational metaphysics (i.e., the Computational Theory of Mind) in the context of consciousness studies. I argue that the ability to simulate the dynamics of a phenomenon using computational models does not entail that that phenomenon is literally performing the computations that are prescribed by the model.

In this commentary, I discuss what the use of computational models to simulate the dynamics or time evolution of some phenomena – here, consciousness – entails for our understanding of those phenomena. I argue that being able to simulate the dynamics of a phenomenon using computational models does not entail that that phenomenon is literally performing the computations that are prescribed by the model.

I applaud Seth for his compelling discussion of artificial consciousness, computational functionalism, and biological naturalism. In his target article, Seth argues that for artificial intelligence systems to be conscious, two conditions must hold. First, computational functionalism must be true of consciousness. That is, it must be possible to define the states and processes associated with consciousness in terms of their functional relations to other states, and it must be possible to formalise these functional relations and their dynamics using the theory of computation (Putnam, 1967, 1975). Here, consciousness is not merely *modelled as* a computation but is *equated to* computation, metaphysically speaking. That is, in this view, consciousness *is literally* a form of computation. Second, the computations associated with consciousness must be characterised by multiple realisability or substrate flexibility (Piccinini & Maley, 2014; Polger & Shapiro, 2016), which denotes different versions of the idea that computations must not depend, or not depend so tightly, on the physical properties of the medium that implements them, such that they can be implemented in different media (i.e., in biological neuronal networks as well as in silicon-based computing systems). Hence, for a conscious computing artefact to exist, the computations associated with consciousness must at minimum be implementable in silico.

In the article, Seth reviews "negative" arguments against the possibility of conscious artificial intelligence, arguing from the implausibility of computational functionalism to that of conscious computing artefacts. Here, I make a complementary point that is independent of the truth of computational functionalism. I point out that one can use computational tools to study and simulate what are – at least on the face of it – non-computational phenomena. For example, when one simulates the orbit of Jupiter around the sun, one does not usually assume that either Jupiter itself or its orbit is literally computing an integral. The general point is that the fact that some phenomenon can be simulated or modelled using computational tools says precisely *nothing* about the metaphysical status of the phenomena being simulated. In particular, it provides no evidence at all for the idea that these phenomena are themselves literally performing computations.

Analogously, in the computational neuroscience of consciousness, we routinely leverage computational tools to model conscious processes. Here, it is crucial to distinguish computational neuroscience, defined as the study of the structure, function, and dynamics of the brain using computational modelling and simulation methods, from the Computational Theory of Mind, which is a metaphysical position that equates cognition and computation (Fodor, 1975). Their difference matters for the possibility of conscious artefacts.

To say that a physical system, such as a digital computer or a brain, implements a given computation is to say that there exists a systematic mapping between two descriptions of the system: On the one hand, a dynamical description of some phenomenon and the observables or data that it generates, and on the other hand, the set of algorithmic steps required to perform that computation (Horsman et al., 2014). In computational neuroscience, this relationship enables a form of analogical explanation (Chirimuuta, 2024), in which we explain a phenomenon that we are trying to understand – the structure, function, and dynamics of the brain – in terms of an engineered artefact that we already understand – a designed artefact, the computing device. Mathematical frameworks like nonequilibrium statistical mechanics (González & Davis, 2017; Niven, 2010) and the free energy principle (Friston, 2019; Ramstead et al., 2023) can be used to determine the mathematical and physical conditions under which these descriptions are in fact equivalent or interchangeable. But this is an *epistemic* equivalence, defined and operative at the level of our descriptions of phenomena. It is *not a metaphysical statement* about whether the

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

system or phenomenon being described is computational, metaphysically speaking.

I believe that the plausibility of the idea that it is possible to design conscious computing artefacts derives in part from not having properly made the distinction between computational modelling and computational metaphysics. The implication for the existence of conscious artificial intelligence is to disqualify a popular line of evidence for its existence: Even if we did in fact have access to a complete and final computational model of an episode of conscious experience, or even consciousness generally, we nonetheless would have no guarantee that the dynamics or processes being modelled or simulated are themselves computational in nature. In other words, the availability of computational models of consciousness provides no evidence for computational functionalism about consciousness, and hence no evidence in favour of conscious artificial intelligence.

In summary, the fact that we have access to computational models that seem to simulate aspects of consciousness provides no evidence that these models are themselves consciously experiencing anything. This point dovetails with Seth's "negative argument" against the plausibility of computational functionalism.

Acknowledgements. The author is grateful to Evan Thompson for useful discussions.

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Competing interests. The author declares no competing interests.

References

Chirimuuta, M. (2024). *The brain abstracted: Simplification in the history and philosophy of neuroscience*. MIT Press.

Fodor, J. A. (1975). *The language of thought* (Vol. 5). Harvard University Press.

Friston, K. (2019). A free energy principle for a particular physics. arXiv preprint arXiv:1906.10184.

González, D., & Davis, S. (2017). Jarzynski equality in the context of maximum path entropy. In *AIP conference proceedings* (Vol. *1853*, No. 1, p. 080003). AIP Publishing LLC.

Horsman, D., Stepney, S., Wagner, R. C., & Kendon, V. (2014). When does a physical system compute?. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *470*(2169), 20140182.

Niven, R. K. (2010). Minimization of a free-energy-like potential for non-equilibrium flow systems at steady state. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1545), 1323–1331.

Piccinini, G., & Maley, C. J. (2014). The metaphysics of mind and the multiple sources of multiple realizability. In M. Sprevak, & J. Kallestrup (Eds.), *New waves in philosophy of mind* (pp. 125–152). Palgrave Macmillan UK.

Polger, T., & Shapiro, L. (2016). *The multiple realization book*. Oxford University Press.

Putnam, H. (1967). Psychological predicates. In W. H. Capitan, & D. D. Merrill (Eds.), *Art, mind, and religion* (pp. 37–48). Pittsburgh University Press.

Putnam, H. (1975). *Mind, language, and reality*. Cambridge University Press.

Ramstead, M. J. D. , Sakthivadivel, D. A. R., Heins, C., Koudahl, M., Millidge, B., Da Costa, L., Klein, B. & Friston, K. J. (2023). On Bayesian mechanics: A physics of and by beliefs. *Interface Focus*, *13*(3), 20220029.

Life is inherently computational, and consciousness probably is too – A commentary on: Conscious artificial intelligence and biological naturalism

Blake Richards[a,b,c,d,e,f*] and Blaise Agüera y Arcas[g,h]

[a]Mila (The Quebec AI Institute), Montreal, QC, Canada; [b]School of Computer Science, McGill University, Montreal, QC, Canada; [c]Department of Neurology & Neurosurgery, McGill University, Montreal, QC, Canada; [d]The Montreal Neurological Institute, McGill University, Montreal, QC, Canada; [e]CIFAR, Learning in Machines and Brains Program, Toronto, ON, Canada; [f]Google, Montreal, QC, Canadaec; [g]Google, Seattle, WA, USA and [h]Santa Fe Institute, Santa Fe, NM, USA
blake.richards@mila.queb
blaisea@google.com

*Corresponding author.

doi:10.1017/S0140525X25103646, e355

Abstract

We agree with Seth that consciousness is grounded in processes related to life, namely prediction, homeostasis, and autopoiesis. However, we believe he is wrong to frame these processes as non-computational. In fact, prediction, homeostasis, and autopoiesis are inherently computational, as shown by von Neumann's Universal Constructor theory of general computation. Given this, we believe that AI consciousness is likely.

A central premise of Seth's argument is that, because brains are complex living systems that do not seem to resemble today's (non-living) silicon-based computers, there is no reason to believe that computers are capable of consciousness. In this account, consciousness may not be computational and/or may be substrate-dependent, i.e. may only be realizable by living systems.

We believe that this account takes too narrow a view of computation, misconstrues the nature of platform independence or multiple realizability, and doesn't acknowledge the fundamentally computational nature not only of the brain, but of life itself.

Multiple mathematical definitions of computation have been advanced, beginning with Church (1936) and Turing (1936), later shown to be equivalent (Turing, 1937). Another definition is von Neumann's Universal Constructor (von Neumann & Burks, 1966). Von Neumann's version differs from Turing's in defining computation as *embodied*: while the Turing Machine consists of a head that can read and write symbols on a tape, the Universal Constructor is a cellular automaton, a machine made out of the very symbols it can read and write. Von Neumann's construction thus has closure with its physical universe, whereas Turing's construction does not.

Despite this important conceptual difference, though, von Neumann's theory of computation is mathematically equivalent to

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

that of Turing and Church, in that one can emulate a Universal Turing Machine with a Universal Constructor, or vice versa. We can therefore characterize physical processes as "computation" as long as the functions that describe those processes are computable by either of these mathematically equivalent definitions.

This is highly relevant to autopoietic systems. Per Kleene's Second Recursion Theorem (Kleene 1938), in any Turing-complete system, there exist programs that can compute a function while also having access to their own source code. This allows for the creation of self-replicating programs. Building on this concept, Von Neumann realized that heritable, embodied self-replication was required for Darwinian evolution; thus, living systems must include a Universal Constructor to grow, maintain themselves, or replicate. Since a Universal Constructor *is* a computer, it follows that any system capable of autopoiesis must be computational. This is why DNA, RNA, ribosomes, etc., can be used to carry out arbitrary computations (Adleman 1994): Life is inherently computational.

This brings us to what we believe to be Seth's fundamental error: when we examine the processes that he correctly links to consciousness (prediction, homeostasis, and autopoiesis), we see that these processes can *only* be implemented by physical systems that can be characterized as computers engaged in computation.

Of course, each of these processes has different relationships to the question of closure with their physical environment. Prediction can be disembodied: "simulating" predicting *is* predicting, just as a "simulation" of a calculator that tells us $2 + 2 = 4$ *is* a calculator, having computed the same function. Homeostasis is more embodied, as it demands the ability to modify some physical variable, e.g., temperature or blood-glucose level. Autopoiesis is at the extreme end of embodiedness. Hardware cannot fix itself or replicate without a Universal Constructor. But autopoiesis is still computation.

Given these considerations, Seth's theory of consciousness is in fact a *computational* theory of consciousness, even if he doesn't wish to apply this word.

Some confusion arises here due to the translation of "substrate independence" from the realm of pure mathematics to engineering practice and linguistic usage. In engineering, computational substrates matter. For example, without GPUs and other massively parallel hardware, modern AI would be impossible. This is why neural network-based AI didn't take off until the 2000s despite having first been attempted in the late 1950s.

The linguistic issue is more insidious. One could assert that even if an artificial kidney successfully filters urea, it isn't *really* a kidney. Alternatively (and, perhaps, more usefully), one could define a kidney functionally. However, we should not confuse this debate over terminology with anything more profound. Philosophers Susan Schneider and David Chalmers run afoul of this error when they speculate that computation on a silicon platform might be *functionally* identical to neural computation in the brain, but the silicon version might be a kind of "philosophical zombie" lacking any subjective consciousness (Schneider, 2021; Chalmers, 2022).

This is dubious. The very nature of life is functional – from molecules, to cells, to organs (including the brain), to whole organisms, and beyond. Every living system, or part thereof, has inputs and outputs, and evolution constantly "invents" alternative implementations: consider aerobic vs. anaerobic respiration, or the way sophisticated camera eyes arose in cephalopods and in vertebrates, or the "multiple discovery" of wings in insects, birds, and bats. These all have technological analogs too: fuel combustion, digital cameras, and airplanes. Both biological and technological systems can be characterized as evolving ecologies of functional relationships, where the substitution of one implementation or substrate for another is commonplace.

We hypothesize that consciousness, like flight, is functional, not only in the mathematical or computational sense described above, but also in the sense that it serves important functions for intelligent social agents. We do *not* see any evidence to suggest that consciousness must involve an inherently embodied form of computation, like the filtering of urea, though we remain open to the possibility. As far as we understand, though, consciousness supports the ability to look after oneself and plan for the future, as Seth implies with the nod to homeostasis, as well as supporting introspection, theory of mind, empathy, and social cooperation (Graziano 2013).

Like prediction, these are all computations that can be carried out purely in the information domain, implying that there is no impediment to the realization of consciousness *in silico*; a perfect "simulation" of consciousness *is* consciousness, just as a perfect "simulation" of a calculator is a calculator. We would even speculate that some current AI systems, given their abilities to predict, model themselves and others, and self-correct, may already be partly conscious. We do not claim certainty either way. But we certainly do not share Seth's confidence that "real" artificial consciousness is unlikely along current trajectories. On the contrary, we think it is likely.

Acknowledgments. We would like to thank our colleagues from the Paradigms of Intelligence team at Google for discussions that helped us to prepare this commentary.

Financial support. Blake Richards was supported by NSERC (Discovery Grant: RGPIN-2020-05105; Discovery Accelerator Supplement: RGPAS-2020-00031) and CIFAR (Canada AI Chair; Learning in Machine and Brains Fellowship).

Competing interests. Both authors of this commentary are employees at Google. There are no other competing interests to report.

References

Adleman, L. M. (1994). Molecular computation of solutions to combinatorial problems. *Science (New York, N.Y.) 266* (5187): 1021–24.

Chalmers, D. J. (2022). *Reality+: Virtual worlds and the problems of philosophy*. W. W. Norton & Company.

Church, A. (1936). An unsolvable problem of elementary number theory. *American Journal of Mathematics*, *58*(2), 345–363.

Graziano, M. S. A. (2013). *Consciousness and the social brain*. Oxford University Press.

Kleene, S. C. (1938). On notation for ordinal numbers. *Journal of Symbolic Logic, 3*(4), 150–155.

Neumann, J. V., and Burks, A. W. (1966). *Theory of self-reproducing automata*. University of Illinois Press.

Schneider, S. (2021). *Artificial you: AI and the future of your mind*. Princeton University Press.

Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society,* Third Series *s2, 42*(1), 230–265.

Turing, A. M. (1937). Computability and λ-definability. *Journal of Symbolic Logic 2*(December), 153–163.

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

Towards evidence-based governance of potentially conscious technologies

Ryan Rodriguez[a] and Nita Farahany[b]*

[a]Cognitive Futures Lab, Duke University, Durham, USA and [b]Duke University School of Law, Duke University, Durham, NC, USA
r.rodriguez@duke.edu
farahany@duke.edu
https://law.duke.edu/fac/farahany

*Corresponding author.

doi:10.1017/S0140525X25103579, e356

Abstract

Seth's biological naturalism claims consciousness requires living systems. We propose experiments operationalizing this debate through testable predictions about allostasis-predictive processing, autopoiesis-recurrent processing, and causal closure-global workspace relationships. Using a probabilistic framework, we show how experimental results could reduce metaphysical disagreement about computational functionalism, motivating empirical research that informs both consciousness theory and governance of potentially conscious technologies.

Seth's biological naturalism restricts consciousness to living systems, a claim with profound implications for emerging technologies. We propose experiments that transform this metaphysical debate into testable predictions, offering both theoretical progress and evidence-based governance for potentially conscious technologies.

The implications extend beyond artificial intelligence (AI). When researchers replicate pain pathways in human assembloids (Kim et al., 2025), we must trust that those assembloids lack subjective experience. Due to our limited understanding of consciousness, technologies whose conscious status divides expert opinion could emerge in the near term. We outline a research agenda aiming to reduce expert disagreement on the nature of consciousness by converting metaphysical debates into testable predictions.

Ongoing disagreements over metaphysical assumptions and scientific theories of consciousness suggest that assessments of nonhuman consciousness are best framed probabilistically. Long et al. (Long et al., 2024) formalize an individual's credence that a specific AI system is conscious, $P(C)$, as $P(C) = P(CF) \times P(C|CF)$, with $P(CF)$ representing credence that computational functionalism is true, and $(C|CF)$ representing credence that the system is conscious if computational functionalism is true.

Previous work (Butlin et al., 2023) suggests several empirical indicators inspired by scientific theories of consciousness compatible with computational functionalism. The above equation can be expanded into the below:

$$P(C) = P(CF) \times \sum_{i=1}^{n} P(T_i|CF) \times P(C|T_i,\ CF)$$

$$\text{where} \begin{cases} C = \text{the AI system is conscious} \\ CF = \text{computational functionalism is true} \\ T_i = \text{a specific theory of consciousness is true} \\ n = \text{the number of relevant theories} \end{cases}$$

The indicators proposed by Butlin et al. impact $P(C|T_i,\ CF)$, meaning that detecting their presence in an AI system will actually *increase* collective uncertainty about the system's consciousness. Evidence supporting computational indicators of consciousness will paradoxically increase disagreement unless we first address underlying metaphysical disputes about $P(CF)$.

Using the recent adversarial collaboration testing the predictions of global neuronal workspace and integrated information theories of consciousness (Ferrante et al., 2025) as a blueprint, we propose experiments that translate Seth's arguments against computational functionalism into falsifiable predictions in order to reduce the individual-to-individual variance of $P(CF)$.

We formalize credence in a specific theory of consciousness as $P(T_i) = P(Q) \cdot P(T_i|Q) + P(\sim Q) \cdot P(T_i| \sim Q)$, where Q represents a metaphysical statement (e.g. mind is substrate independent) and $P(Q)$ represents one's credence that the statement is true. By designing studies targeting the term $P(Q) \cdot P(T_i|Q)$, we aim to reduce individual-to-individual variance in $P(C)$ by condensing distributions of beliefs $P(T_i)$ and $P(Q)$. The below experiments are designed to explicitly preclude a rationale for updating $P(T_i| \sim Q)$. Therefore, when observers who inaccurately predict the experimental results adjust their confidence in $P(T_i|Q)$, maintaining coherent total probability requires these observers revise either their confidence in the experimentally relevant theory of consciousness, $P(T_i)$, or in their metaphysical priors, $P(Q)$ and $P(\sim Q)$.

1. Q: Consciousness's function is allostasis, T_i: Predictive processing theories: Researchers have successfully taught brain organoids—three-dimensional cell cultures—to perform reinforcement learning tasks such as Pong by applying unpredictable "punishment" stimuli for undesirable behavior (Jordan et al., 2024; Kagan et al., 2022). Seth suggests that the biological imperative for allostasis may have been "the evolutionary driver for predictive processing as a general mechanism underlying conscious perception." To test this relationship, researchers could analyze organoid computing performance across a range of punishment-stimuli entropy values. If predictive processing serves allostasis, learning should improve with punishment-stimuli entropy until excessive unpredictability disrupts homeostasis, degrading performance. Conversely, if predictive processing operates independently of allostatic imperatives, entropy should

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

minimally affect performance. Results favoring the former would require observers either to update toward biological naturalism or away from predictive processing as an explanation for conscious perception.

2. $Q:$ Consciousness emerges from autopoiesis, $T_i:$ Recurrent processing theories: Researchers have produced human brain organoids that structurally mimic the reciprocal connections that form "reentry circuits" in human brains, and have demonstrated that connected organoids produce spontaneous oscillations which are more complex than single or fused organoids (Kirihara et al., 2019; Osaki et al., 2024). Experiments could compare matched pairs of organoids—one group self-generating energy (ATP) through metabolism, the other supported by external perfusion of metabolic substrates. Seth's theory predicts that self-generating organoids should display more complex structures and oscillatory patterns. If instead the structural and functional complexity of matched pairs proves similar when provided equivalent energy, observers must either weaken their credence that consciousness emerges from autopoiesis or their belief in recurrent processing theories of consciousness.
3. $Q:$ Causal closure, $T_i:$ Global neuronal workspace theory: Current neuroprostheses are capable of rapidly decoding speech from invasive brain measurements (Metzger et al., 2023), and advances in decoding algorithms suggest noninvasive decoding may be feasible (Défossez et al., 2023; Tang et al., 2023). Noninvasive neuroprosthesis could enable experiments testing the inseparability of the brain's "hardware" and "software" by recording subjects in three conditions: speaking aloud, having a neuroprosthesis decode silently mouthed speech, and having a neuroprosthesis decode imagined speech from brain activity alone. If ignition collapses in the final condition, such that semantic content associated with speech production no longer appears distributed throughout the brain as the motor components are offloaded to the neuroprosthesis, it would suggest either that substrate-specific implementation details play a crucial role in the mind's function, or that global neuronal workspace theory requires revisions. Maintained integration across conditions would support substrate flexibility.

We recognize that committed adherents to either position could interpret these results flexibly. Biological naturalists might argue that negative results inadequately capture relevant biological properties, while functionalists might claim that positive results demonstrate necessary but insufficient conditions. Pre-registered predictions and adversarial collaboration protocols can minimize post-hoc rationalization, though experimental results rarely force belief revision when theories serve broader theoretical frameworks.

But history is littered with regrettable and discriminatory policies justified by scientific theories that were, at the time, contested (Naomi Oreskes, 2021). We suggest that policymakers exercise epistemic humility when considering the moral and legal standing of artificial minds and ground policy decisions in the plausibility of all mainstream theories of consciousness.

Our proposed experiments offer a pathway to operationalize this debate and move policy beyond epistemic humility. If results favor computational or hybrid accounts, policymakers must prepare for conscious AI sooner than Seth predicts; if results uphold biological naturalism, regulators gain clearer limits. Whether the proposed experiments vindicate or falsify biological naturalism, they convert Seth's largely metaphysical debate into a concrete research agenda.

Acknowledgements. None.

Financial support. This material is based upon work supported by the U.S. National Science Foundation under award No. 2112562. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. National Science Foundation.

Competing interests. RR: Competing interests: none.
NF: Nita Farahany is on the advisory board for OpenBCI.

References

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. https://doi.org/10.48550/arXiv.2308.08708

Défossez, A., Caucheteux, C., Rapin, J., Kabeli, O., & King, J.-R. (2023). Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, *5*(10), 1097–1107. https://doi.org/10.1038/s42256-023-00714-5

Ferrante, O., Gorska-Klimowska, U., Henin, S., Hirschhorn, R., Khalaf, A., Lepauvre, A., Liu, L., Richter, D., Vidal, Y., Bonacchi, N., Brown, T., Sripad, P., Armendariz, M., Bendtz, K., Ghafari, T., Hetenyi, D., Jeschke, J., Kozma, C., Mazumder, D. R., … Melloni, L. (2025). Adversarial testing of global neuronal workspace and integrated information theories of consciousness. *Nature*, *642*(8066), 133–142. https://doi.org/10.1038/s41586-025-08888-1

Jordan, F. D., Kutter, M., Comby, J.-M., Brozzi, F., & Kurtys, E. (2024). Open and remotely accessible Neuroplatform for research in wetware computing. *Frontiers in Artificial Intelligence*, *7*. https://doi.org/10.3389/frai.2024.1376042

Kagan, B. J., Kitchen, A. C., Tran, N. T., Habibollahi, F., Khajehnejad, M., Parker, B. J., Bhat, A., Rollo, B., Razi, A., & Friston, K. J. (2022). In vitro neurons learn and exhibit sentience when embodied in a simulated game-world. *Neuron*, *110*(23), 3952–3969.e8. https://doi.org/10.1016/j.neuron.2022.09.001

Kim, J. I., Imaizumi, K., Jurjuț, O., Kelley, K. W., Wang, D., Thete, M. V., Hudacova, Z., Amin, N. D., Levy, R. J., Scherrer, G., & Pașca, S. P. (2025). Human assembloid model of the ascending neural sensory pathway. *Nature*, *642*(8066), 143–153. https://doi.org/10.1038/s41586-025-08808-3

Kirihara, T., Luo, Z., Chow, S. Y. A., Misawa, R., Kawada, J., Shibata, S., Khoyratee, F., Vollette, C. A., Volz, V., Levi, T., Fujii, T., & Ikeuchi, Y. (2019). A human induced pluripotent stem cell-derived tissue model of a cerebral tract connecting two cortical regions. *iScience*, *14*, 301–311. https://doi.org/10.1016/j.isci.2019.03.012

Long, R., Sebo, J., Butlin, P., Finlinson, K., Fish, K., Harding, J., Pfau, J., Sims, T., Birch, J., & Chalmers, D. (2024). Taking AI Welfare Seriously. https://doi.org/10.48550/arXiv.2411.00986

Metzger, S. L., Littlejohn, K. T., Silva, A. B., Moses, D. A., Seaton, M. P., Wang, R., Dougherty, M. E., Liu, J. R., Wu, P., Berger, M. A., Zhuravleva, I., Tu-Chan, A., Ganguly, K., Anumanchipalli, G. K., & Chang, E. F. (2023). A high-performance neuroprosthesis for speech decoding and avatar control. *Nature*, *620*(7976), 1037–1046. https://doi.org/10.1038/s41586-023-06443-4

Oreskes, N. (2021). *Why Trust Science*. Princeton University Press.

Osaki, T., Duenki, T., Chow, S. Y. A., Ikegami, Y., Beaubois, R., Levi, T., Nakagawa-Tamagawa, N., Hirano, Y., & Ikeuchi, Y. (2024). Complex activity and short-term plasticity of human cerebral organoids reciprocally

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

connected with axons. *Nature Communications*, *15*(1), 2945. https://doi.org/10.1038/s41467-024-46787-7

Tang, J., Du, M., Vo, V. A., Lal, V., & Huth, A. G. (2023). Brain encoding models based on multimodal transformers can transfer across language and vision. http://arxiv.org/abs/2305.12248

Substrates, computations, and abundant consciousness: Commentary on Anil Seth

Luke Roelofs*

University of Texas at Arlington, Arlington, TX, USA
luke.mf.roelofs@gmail.com

*Corresponding author.

doi:10.1017/S0140525X2510349X, e357

Abstract

Seth opposes computational views of consciousness to views that emphasise biology, because they imply different verdicts about hypothetical AIs that replicate information-processing without being biologically alive. But if we accept a background theory like panpsychism, on which consciousness is abundant in nature, this opposition becomes less central: even if a non-biological system could not have *our* consciousness, we should still default to accepting any self-ascriptions of consciousness it makes.

Seth's case that artificial consciousness is unlikely, on current development trajectories, comprises both a negative half (critiquing computational functionalism, which suggests artificial consciousness comes for free with artificial intelligence), and a positive half (defending biological naturalism, which makes artificial consciousness depend on artificial life).

While I have some sympathy with both halves, I won't try to evaluate either computational functionalism or biological naturalism here. Rather, I will raise two objections, both suggesting that, even if we reject computational functionalism, artificial consciousness is still likely, albeit in strange and unfamiliar forms. This points towards a question lurking in the background: how sparse or abundant should we expect consciousness to be?

My first objection concerns substrate-independence (or substrate-flexibility), the idea that the very same conscious process could be implemented in basically any sort of material, if it is functionally isomorphic to a brain. Seth points out that if this is true, then making a silicon computer functionally isomorphic to a human brain, in some coarse-grained sense, should suffice for artificial consciousness. But he often writes as though the flipside of this is that if substrate-independence *fails*, then systems with a radically different substrate than us would completely *lack* consciousness. But of course, strictly all that follows from rejecting substrate independence is that a system with a radically different substrate would have to *differ* from us in respect of consciousness. That's compatible with it still being conscious, and compatible with its consciousness resembling ours in lots of ways. It just means that it cannot be exactly the same: A difference in substrate might systematically alter the system's phenomenal qualities, phenomenal structure, agency, hedonic valence, or something else, without depriving it of consciousness altogether.

My second objection then concerns how we should decide between these possibilities. Suppose we find ourselves dealing with a system that is both very different from us in substrate and very similar in function. In particular, suppose that it makes claims and demands, talks about what it wants and likes and dislikes, even describes its internal states as "pleasant" or "unpleasant," and that we are in this situation without much confidence in computational functionalism, *or* biological naturalism, or indeed any particular theoretical framework. We're profoundly unsure what makes something conscious, but here's a being before us that acts like it's conscious and claims to be conscious: do we default towards accepting or rejecting its self-ascriptions?

It seems to me that our default here should be acceptance: we need a compelling argument to doubt a system's self-ascriptions of consciousness, but not to accept them. This default might be justified by precautionary pragmatic reasons (for discussion, see e.g. Long et al., 2024), or by the "Copernican" argument that we should not assume our own substrate is somehow uniquely privileged (see Schwitzgebel and Pober Ms.), or by reflection on the fact that we obviously have ascribed consciousness to other humans (and animals, and historically many other things) without the foggiest idea of the correct theory of consciousness.

Of course, that default might be cancelled in various ways. We might look at the system's history and find that it was deliberately designed to mimic human discussions of consciousness, or to game out human tests for consciousness, and reason that this cancels out the evidential value its words and actions would otherwise have (Andrews & Birch, 2023, Schwitzgebel and Pober Ms.). For my part, I think this sort of history-based defeater is currently the most compelling reason to doubt the consciousness of present or near-future AI. But note that this applies to both biological and silicon-based systems, if they have the same kind of history. So it does not directly intersect with the dispute over functionalism and naturalism.

We could also override the default by appealing to a theory of consciousness. If we were strongly convinced of Seth's biological naturalism, we might reject an AI's self-ascriptions of consciousness on the basis that it is not biologically alive. But justifying such denial surely requires high confidence in biological naturalism, not just thinking it the best current theory. The relevant comparison would not be with specific rival theories like computational functionalism, but with a general stance of humble agnosticism. Because if we're agnostic about the right theory of consciousness, then it's hard to justify theory-based rejections of a system's claims to be conscious.

I think all of this taps into an underlying question: how abundant or sparse should we expect consciousness to be? Do we live in a world full of consciousness, where systems with different functional organisations are conscious in different ways, and systems built out of different substrates are conscious in different ways? Or is "consciousness" the name for something odd and rare, something unique to humans and beings that match humans sufficiently closely in some key respects (in organisation, in substrate, or in something else)?

In a world of abundant consciousness, even if we reject substrate-independence, we should still default to taking differences of substrate as indicating differences in consciousness, not its absence. (To do otherwise would be human exceptionalism.) Whereas if we assume a world of sparse consciousness, then it

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

makes more sense to deny consciousness to systems that, even if they resemble us in some ways, differ from us in any fundamental way. (To do otherwise would be anthropomorphism.)

For my own part, I think the best-supported current theory is panpsychism, an extreme form of the abundant view (for arguments, see Chalmers, 2015, Seager, 2006). Panpsychism reframes the question of AI consciousness: instead of asking "is this system conscious or not?," it suggests we ask "what sort of consciousness is present in this system?," and thus clarify what forms or dimensions of consciousness we're interested in (cf. Birch, Schnell, & Clayton, 2020). Do we want to know whether AI has valenced experiences? A unified perspective? Something else? This outlook pushes us away from deciding between functionalism and naturalism, as theories of *whether* a system is conscious, and towards asking how different factors, organisational or substrate, determine the dimensions and parameters of consciousness.

Acknowledgements. None.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interests. This research has no competing interests to report.

References

Andrews K., & Birch J. (2023). What Has Feelings? *Aeon* https://aeon.co/essays/to-understand-ai-sentience-first-understand-it-in-animals.

Birch J., Schnell A., & Clayton N. (2020). Dimensions of animal consciousness. *Trends in Cognitive Sciences*, *24*(10), 789–801.

Chalmers D. (2015). Panpsychism and panprotopsychism. In Y. Nagasawa, & T. Alter (Eds.), *Consciousness in the physical world: Essays on Russellian Monism* (pp. 246–276). Oxford University Press.

Long R., Sebo J., Butlin P., Finlinson K., Fish K., Harding J., Pfau J., Sims T., Birch J., & Chalmers D. (2024). Taking AI welfare seriously. https://arxiv.org/abs/2411.00986.

Schwitzgebel E., & Pober J. The Copernican argument for alien consciousness; The mimicry argument against robot consciousness. https://arxiv.org/abs/2412.00008.

Seager W. (2006). The 'Intrinsic Nature' argument for panpsychism. *Journal of Consciousness Studies*, *13*(10–11), 129–145.

Invited commentary: assessment or attribution of consciousness?

Tobias Schlicht*

Faculty of Philosophy and Educational Research, Ruhr-University Bochum, Bochum, Germany
tobias.schlicht@rub.de
www.tobiasschlicht.com

*Corresponding author.

doi:10.1017/S0140525X25103671, e358

Abstract

I argue that we are not epistemically justified to judge whether the neurobiological details of computational mechanisms are relevant to consciousness or not. It is unclear how the medium independence of neural computation required for Computational Functionalism could be established. But it is also unclear whether predictive processing depends on life-sustaining biological processes and whether consciousness is medium dependent.

We are not epistemically justified to judge whether Biological Naturalism or Computational Functionalism is true.

Seth correctly emphasizes the *potential* relevance of biochemical processes for consciousness: "some properties of life may be necessary for consciousness" (p. 13), and conscious artificial systems may be impossible. But it is far from *established* that consciousness is anchored in life and that "the (higher-level) mechanisms of predictive processing are [. . .] inextricably embedded in their (lower-level) biological context" (p. 8). Seth's interpretation of predictive processing seems idiosyncratic. Other proponents may disagree, regarding it as a general computational model of cognition and consciousness, without metaphysical commitments like making its explanatory merit dependent on the prediction error mechanism being tied to the self-sustaining mechanisms in our bodies (Schlicht, 2025).

Regarding the possibility of conscious AI, Seth engages with Butlin et al., (2023), who advocate a research program that *assumes Computational Functionalism* and proposes to use neuroscientific theories to yield computational markers of consciousness and thereby *assess* (i.e., verify) consciousness in AI. Like other mental phenomena, consciousness is assumed to be functional and computational. Functionalism allows for multiple realizability but Computationalism requires medium independence (Piccinini, 2009). For their combination, Computational Functionalism, to be vindicated, neural computation must be independent from, i.e., unconstrained by, properties of its implementation in the brain. But while digital computation may be medium independent by definition (Haugeland, 1985), it is an open question (indeed, questionable) whether neural computation is medium independent. Maley (2023, 2025) argues that it is *analog* and therefore *medium dependent*. Piccinini (2020) takes it to be different from digital and analog computation, yet *medium independent* (see also Drayson, 2025). But since Piccinini acknowledges that we do not know how this *computation sui generis* works, calling it medium independent is premature.

For Butlin et al.'s program to succeed, the theories they consider, Global Neuronal Workspace Theory (GNW), for example, are presented as computational theories, allowing for conscious AI *in principle* (Dehaene et al., 2017, Mudrik et al., 2025). GNW proposes four *neuronal* signatures of consciousness (Dehaene, 2014), supported by evidence from human brains. But for it to provide *computational* markers, these would have to be medium independent and applicable not only to human brains but also to non-biological conscious systems. How could computational markers be empirically vindicated for non-biological systems? Phenomenal consciousness must be characterized functionally (contra Chalmers, 1996), since discovering a computational mechanism requires it to have effects. But even characterized functionally, e.g., via identification with conscious access (Dehaene, 2014, Cohen & Dennett, 2011), a non-biological implementation of a workspace may *not* yield *conscious* but only "Ersatz broadcasting" (Block, 2008, p. 119). In non-biological artificial systems, all markers we use in human and non-human animals are either trivially absent (e.g., neuronal and first-person markers), ambiguous (e.g., reports, other behaviors), or question-begging (computational markers). Consider that the sophisticated outputs of large language models are not typically taken as

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

evidence for consciousness. Contra Butlin et al.'s suggestion, because of redundancy and incompatibility, computational markers do not pile up and make conscious AI more likely.

Neuroscientific theories provide empirically supported neurobiological markers but not equally supported computational markers. They are supported as biological theories. Cao (2022) and Chirimuuta (2022, 2025) highlight the difficulty of separating functional from structural properties in chemical signaling between neurons, listing numerous features of neural information processing that may make conscious AI impossible *in practice*. Possibility *in principle* is unverifiable if conscious AI requires data centers the size of Asia. A silicon-based system functionally equivalent to a brain may be technologically impossible, apart from the fact that we do not know which level of granularity is required for functional and computational equivalence. Given that the evidence favors biological accounts, Cao (2022, p. 25) is right that the burden of proof is on the functionalist to demonstrate that we can fully replicate the function of neurons. Birch (2024, p. 68) concurs that if equivalence of computations within neurons was required, this may be impossible to create.

I agree with Seth that we should thoroughly scrutinize the typically unquestioned assumption of Computational Functionalism. But Biological Naturalism only has a slight advantage. While I have stressed that neuroscientific theories of consciousness are empirically supported via evidence from human brains, capturing consciousness in other *biological* systems may require a more abstract functionalist account. Being unsure about whether the neurobiological and chemical details of biological systems are relevant for consciousness also means that they may *not* be. Functionalists could point to evidence for consciousness in other animals to support multiple realizability. However, this would not vindicate *Computational* Functionalism, and in the case of other humans or nonhuman animals, we remain in the realm of "messy" (Cao, 2022, p. 26) biological implementation. An evolutionary account explaining how these biological mechanisms have evolved, say, in mammalian and avian brains (Godfrey-Smith, 2024), is unavailable for artificial systems. But the simplified picture rooted in McCullough and Pitts' (1943) ideas *may* capture the essentials of neural computation, making the implementation properties of human brains dispensable. Analogously, the computational elements of the predictive processing framework, including the hierarchical generative model and prediction error minimization machinery, *may* be detachable from the life-sustaining biochemical processes of living organisms and still yield consciousness if implemented non-biologically. Combined with multiple realizability, this would support Computational Functionalism. However, there is no evidence for either Seth's view or the opposition. We simply do not know, and it is difficult to see how we could decide the issue.

Verifying if an embodied AI can do a backflip is straightforward but verifying if it is conscious is not. Unless there are not only tests for consciousness (Bayne et al., 2024) but also tests that inform us about medium (in-)dependence of neural computation, we remain epistemically limited about whether Biological Naturalism or Computational Functionalism is more plausible. Since neurobiological markers will always be better supported than computational markers, Seth's biological story seems to be in the lead. But the moral of our epistemic situation is that *we cannot assess consciousness in artificial systems*. We can only *attribute* it – from the "intentional stance" (Dennett, 1987) – without verifying it, a much less ambitious claim.

Acknowledgements. I would like to thank Rony Hirschhorn and Niccolo Negro for helpful comments on an earlier draft of this comment.

Financial support. Research on this article was generously funded by the Volkswagen Foundation (Az. 87 105-1).

Competing interests. There are no competing interests.

References

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 in Cognitive Sciences*, *28*(5), 454–466.

Birch, J. (2024) *The edge of sentience: Risk and precaution in humans, other animals, and AI*. Oxford University Press.

Block, N. (2008) *Consciousness, function, and representation*. MIT Press.

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. arXiv:2308.08708 [cs.AI]. https://doi.org/10.48550/arXiv.2308.08708.

Cao, R. (2022) Multiple realizability and the spirit of functionalism. *Synthese*, *200*, 506. https://doi.org/10.1007/s11229-022-03524-1.

Chalmers, D.J. (1996) *The conscious mind: In search of a Fundamental Theory*. Oxford University Press.

Chirimuuta, M. (2025) Neuromorphic computing and the significance of medium dependence. *Manuscript*.

Chirimuuta, M. (2022) The case for Medium-Dependence. *Journal of Consciousness Studies*, *29*(7–8), 185–194. https://doi.org/10.53765/20512201.29.7.185.

Cohen, M.A., & Dennett, D.C. (2011) Consciousness cannot be separated from function. *Trends in Cognitive Sciences*, *15*(8), 358–364.

Dehaene, S. (2014) *Consciousness and the Brain*. Viking.

Dehaene, S., Lau, H., & Kouider, S. (2017) What is consciousness and could machines have it? *Science*, *358*, 486–492.

Dennett, D.C. (1987) *The Intentional Stance*. MIT Press.

Drayson, Z. (2025) Defending the medium independence of computation. *Mind and Language*, *40*(4), 458–467. https://doi.org/10.1111/mila.12536.

Godfrey-Smith, P. (2024) *Living on earth: Life, consciousness, and the making of the natural world*. Williamm Collins.

Haugeland, J. (1985) *Artificial intelligence. The very idea*. MIT Press.

Maley, C. (2025) Declaring independence from medium independence. *Mind and Language*, *40*(4), 448–457.

Maley, C. (2023) Medium independence and the failure of the mechanistic account of computation. *Ergo*, *10*, 28. https://doi.org/10.3998/ergo.4658.

McCulloch, W., & Pitts, W. (1943) A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics 5*, 115–133. https://doi.org/10.1007/BF02478259.

Mudrik, L., Boly,M., Dehaene, S., Fleming, S.M., Lamme, V., Seth, A., & Melloni, L. (2025) Unpacking the complexities of consciousness: Theories and reflections. *Neuroscience & Biobehavioral Reviews*, *70*, 106053. https://doi.org/10.1016/j.neubiorev.2025.106053.

Piccinini, G. (2020) *Neurocognitive mechanisms. Explaining biological cognition*. Oxford University Press.

Piccinini, G. (2009) Computationalism in the Philosophy of Mind. *Philosophy Compass*, *4*, 515–532. https://doi.org/10.1111/j.1747-9991.2009.00215.x.

Schlicht, T. (2025) Predictive processing's flirt with transcendental idealism. *Noûs*, 1–23. https://doi.org/10.1111/nous.12552.

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

The error theory of LLM consciousness: there is no evidence that standard LLMs are conscious

Susan Schneider*

Center for the Future of AI, Mind and Society, Stiles-Nicholson Brain Institute and Dept. of Philosophy, Florida Atlantic University, Boca Raton, USA
Sschneider@fau.edu
schneiderwebsite.com

*Corresponding author.

doi:10.1017/S0140525X25103920, e359

Abstract

Are chatbots AI zombies or conscious minds? I suspect they are AI zombies, at least when implemented on today's standard hardware. I argue that today's Large Language Mode's (LLM's) consciousness-like behaviors do not suggest they are conscious, because there is an error theory – a theory explaining why they behave as if they are conscious in absence of actual felt experience. LLMs function as "crowdsourced neocortices" – as they scale up, they come to mirror human conceptual structures, leading to human-like behaviors, including those involving consciousness. By contrast, bio-computers, quantum computers, and neuromorphic systems present more serious candidates for AI consciousness.

In this piece, I present a brief case for skepticism concerning whether "standard LLMs" are conscious. But my skepticism does not extend to a range of other AI systems, for there are already AI's that may, for all we know, have some level of consciousness: computers combining the processing of brain organoids and silicon hardware; neuromorphic chips more closely related to the processing of neurons (Davies, 2021), increasingly sophisticated brain emulations of the mouse brain and certain regions of the human brain (OpenBrain, n.d.), and so on. Further, at least one LLM (DeepSeek) is said to have a version that is run on a neuromorphic system (called "Darwin Monkey"), leading to the fascinating question: might some LLM instantiations be conscious, while others are not? (Whalan, 2025)

In my 2019 book, *Artificial You: AI and the Future of the Mind,* I argued for a "wait and see" approach to AI consciousness, stressing the importance of ongoing testing because situations like the above are particularly vexing. Further, I explored ways that intelligence and consciousness come apart, and provided a lengthy discussion of philosophical thought experiments involving neural replacement, urging that they say little about whether the AI's developed today are conscious – although the thought experiments are conceptually coherent, it may not be lawfully or technologically possible to build them (Schneider, 2019). This is a point Seth stresses too (Seth, 2021).

I sympathize with the excellent considerations Seth raises. However, I believe we have further reason to be skeptical that LLMs are conscious, and they do not depend on Biological Naturalism:

The Error Theory. The capability of an LLM trained on human data to emulate human consciousness–related behaviors and functional architecture does not confirm or discredit claims of chatbot consciousness. Instead, the "crowdsourced neocortex" account explains why chatbots assert consciousness and related emotional states, and even exhibit functional configurations analogous to consciousness processing in the biological brain without genuinely having inner experience. In other words, it provides what philosophers call an "error theory" – an explanation of why people erroneously conclude that chatbots have inner lives.

Today's LLMs are a "crowdsourced neocortex" – having crawled so much data, they encode a sort of conceptual map akin to the masses of human internet users whose data they were trained on, representing, say, concepts like [dog], and [consciousness] using weighted connections to concepts humans associate with the categories of dogs and consciousness. This does not mean LLMs are conscious. But it explains why their utterances about consciousness become similar to human utterances as the systems become more sophisticated and are trained on more and more of our data. As the LLMs scale up, their "conceptual systems" come to mirror the masses of users whose data was in the training set, as well as those providing feedback to the system (hence I write "crowdsourced neocortex") (Schneider 2024, 2025). Indeed, this is part of a general pattern: as LLMs scale up, their skills in arithmetic, theory of mind, and a variety of other tasks similarly improve (Wei et al.). If one has doubts that today's LLMs have conceptual structure that mimics that of groups of humans whose data was used, I encourage them to read the related Anthropic papers on circuit tracing (Anthropic 2025).

This means that claims that the presence of functional and behavioral analogs of consciousness in LLMs (trained on data about consciousness, minds, the brain, etc.) are a marker of LLM consciousness should be rejected. For example, Jonathan Birch writes: "If we find signs that an LLM, though not deliberately equipped with a global workspace or perceptual/evaluative reality monitoring system, has implicitly learned ways of recreating them, this should lead us to regard it as a sentience candidate." (Birch, 2024).

For illustrative purposes, let us suppose the system in question is (what I'll call) a "singular" LLM, not a more differentiated Mixture of Experts (MoE) model that gates various "expert" LLMs. A singular LLM, even without an obvious functional decomposition into various experts that causally interact, can still functionally decompose. We can consider a single concept as an activation pattern across the LLM; an abstract "subsystem" could be an activation pattern over activation patterns. In this way, there can be a high-level functional structure to the LLM, with various subsystems that interact causally. So, to consider Birch's example of a global workspace (GW), there can be clusters of activation patterns over activation patterns broadcasting across the GW network. In this way, the model can exhibit higher-order properties akin to a GW, with local processes that become globally available. But this does not indicate that it is conscious, one way or the other.

So how can we determine if a LLM is conscious? First, even though there is no real evidence that a standard LLM is conscious, the issue should still be investigated because they converse so intelligently. Edwin Turner and I have suggested giving linguistic tests to LLMs designed to elicit whether they have inner experience when they are "boxed in" – before they have been fed data about human consciousness, neuroscience, and mindedness (Schneider 2019). Even if the LLM has fewer parameters than current frontier models, a positive result would provide an indication that an LLM, running on the same kind of system, can have conscious

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

instantiations, providing a serious consideration in favor of the view that more sophisticated models running on the same kind of systems also have experience.

Second, consider a special case, a hypothetical LLM running on some sort of "neuromorphic system," where a neuromorphic system is one that either: (1) is instantiated on neuromorphic hardware (e.g., TrueNorth-style spiking chips), or (2) implements neuromorphic architectures. Being more akin to the brain, neuromorphic systems are already a "grey area" that demands further investigation. Here, I suggest independent consciousness tests for such systems, including "boxing in" early versions of the systems and giving the system my ACT (Schneider 2019).

Biocomputation presents a special challenge. Because we know that neurons are the "right stuff" as a substrate of consciousness, we must take biocomputing very seriously. But here, because the more complex systems involve integrated combinations of biology and some other substrate (e.g., silicon), it will be crucial to determine whether the larger system inherits consciousness from the constituents and how integrated any consciousness is. Finally, while many focus on the program level in their discussion of consciousness, this may mislead us – AI consciousness could arise at any level of the stack. For instance, Mark Bailey and I have argued that consciousness may originate from many-body interactions as entangled particles decohere. This means quantum chips and biological systems are of special interest (Schneider and Bailey, forthcoming).

Financial support. This research was supported from William F. Dietrich Foundation.

Competing interests. None.

References

Anthropic. (2025). *Tracing the thoughts of a large language model*. Anthropic. https://www.anthropic.com/research/tracing-thoughts-language-model

Ameisen, E., Lindsey, J., Pearce, A., et al. (2025). On the biology of a large language model. Anthropic. https://transformer-circuits.pub/2025/attribution-graphs/biology.html

Birch, J. (2024). *The edge of sentience: Risk and precaution in humans, other animals, and AI*. Oxford University Press. https://doi.org/10.1093/9780191966729.001.0001

Davies, M., Wild, A., Orchard, G., Sandamirskaya, Y., Guerra, G. A. F., Joshi, P., . . . & Risbud,S. R. (2021). Advancing neuromorphic computing with Loihi: A survey of results and outlook. Proceedings of the IEEE, 109(5), 911–934. https://doi.org/10.1109/JPROC.2021.3067593

Heaven, W. D. (2025). Anthropic can now track the bizarre inner workings of a large language model: What the firm found challenges some basic assumptions about how this technology really works. MIT Technology Review. https://www.technologyreview.com/2025/03/27/1089389/anthropic-can-now-track-the-bizarre-inner-workings-of-a-large-language-model/

Open Brain Institute. (n.d.). Open Brain Platform: Launch your virtual lab and perform neuroscience at the speed of thought. Retrieved August 19, 2025, from https://www.openbraininstitute.org

Schneider, S. (2019). *Artificial you: AI and the future of the mind*. Princeton University Press.

Schneider, S. (2024). *Chatbot epistemology. Social Epistemology*. Advance online publication. https://doi.org/10.1080/02691728.2025.2500030

Schneider and Bailey (forthcoming), "Superpsychism", target paper for forthcoming special issue of the Journal of Consciousness Studies, 2026.

Whalen, R. (2025). China's brain-inspired "Darwin Monkey" AIbreakthrough simulates primate-level intelligence. The Debrief. https://thedebrief.org/chinas-brain-inspired-darwin-monkey-ai-breakthrough-simulates-primate-level-intelligence/

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., . . . & Chi, E. (2022). Emergent abilities of large language models. https://doi.org/10.48550/arXiv.2206.07682

Minimal autopoiesis in an AI system

Eric Schwitzgebel*

Department of Philosophy, University of California, Riverside, CA, USA
eschwitz@ucr.edu

*Corresponding author.

doi:10.1017/S0140525X25103786, e360

Abstract

A minimally autopoietic artificial intelligence (AI) system could be designed according to current engineering principles. Although Seth suggests that standard computational systems are not autopoietic and therefore not conscious, *autopoiesis* is a high-level, functional concept, and nothing appears to prevent autopoietic processes in a standard computational system. Autopoietic requirements do not in principle rule out consciousness in standard AI systems.

Citing Maturana and Varela (1980), Seth characterizes autopoietic systems as systems that "continually regenerate their own material components through a network of processes . . . actively maintain[ing] a boundary between the system and its surroundings." Seth suggests that standard artificial intelligence (AI) systems are not autopoietic and therefore not conscious. I suggest that standard AI systems could be, and perhaps some already are, minimally autopoietic.

Consider a hypothetical solar-powered robot designed to move toward light when its charge is low. The system thereby acts to maintain its own functioning. It might model the direction of light sources through predictive processing. It might stay upright by means of an accelerometer and tilt detectors that integrate gravitational and camera inputs. More fancifully, it might be composed of modules held together electromagnetically, so that without electrical power, it falls apart.

Now give the robot error-detection systems and the ability to replace defective parts. When it detects a breakdown in one part – for example, in the upper portion of its left front leg – it orders a replacement part delivered. Upon delivery, the robot scans the part for compatibility (rejecting any incompatible parts), then electromagnetically disconnects the damaged part and installs the new one. If the system has sufficient redundancy, even central processing systems could be replaced. A redundant trio of processors might eject a defective processor and run on the remaining processors until the replacement arrives.

A plastic shell maintains the boundary between the system and its surroundings. The system might detect flaws in the shell, for example, by internal sensors that respond to light entering through unexpected cracks, by visually monitoring its exterior, and perhaps by electrostatically detecting cracks or gaps. Defective shell components might be replaced.

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

If repelling intruders is necessary, we can challenge our robot with fakes. Shipments might sometimes arrive with a part mislabeled as compatible or visually similar to a compatible part, but ruinous if installed. Detecting and rejecting fakes might become a dupe-and-mimic arms race.

I see no in-principle obstacles to creating such a system using standard AI and engineering tools. Such a system is, I suggest, minimally autopoietic. It actively maintains itself. It enforces a boundary between itself and its environment. It continually generates, in a sense, its own material components. It employs predictive processing, fights entropy by drawing on external energy, resists dispersion, and has a solar-electric metabolism.

Does the fact that it depends on shipments mean that it does not actually generate its own parts? Humans also depend on nutrients generated from outside, for example, vitamins and amino acids that we cannot biologically manufacture. Sometimes, these nutrients are shipped to us. Also, it's easy enough to imagine the robot not simply installing but, in a minimal sense, manufacturing a part. Suppose a leg has three modular components. Each component might arrive separately, requiring a simple joining procedure to create the leg as a whole.

In a human, the autopoietic process occurs at multiple levels simultaneously. Cells maintain themselves, and so do organs, and so does the individual as a whole. Our robot does not have the same multilevel autopoiesis. But it's not clear why autopoiesis must be multilevel to count as genuine autopoiesis. In any case, we could recapitulate this imaginative exercise for subsystems within the robot or larger systems embedding the robot. A group-level autopoietic system might comprise several robots that play different roles in the group and who can be recruited or ejected to maintain the integrity of the group and the persistence of its processes.

Perhaps my system does not *continually* regenerate its own components, and that is a crucial missing feature? It's not clear why strict continuousness, rather than periodic replacement as needed, should be essential to autopoiesis. In any case, we can imagine that, if necessary, the robot has some fragile parts that need continual refurbishment. Perhaps it occupies an acidic environment that continually degrades its shell, so that its shell coating must be continually monitored and replaced through capillaries that emit lacquer as needed from a refillable lacquer bag.

My system does not reproduce, but reproduction, sometimes seen as essential to life, is not standardly viewed as necessary for autopoiesis (Maturana & Varela, 1980, p. 100).

I invite Seth to respond to this thought experiment by either (1) explaining what aspects of autopoiesis are missing that could not straightforwardly be engineered into this system and why those missing elements matter for consciousness or (2) acknowledging that a system of this sort could be a minimal autopoietic system (perhaps the type of "living AI" he briefly gestures toward) and exploring the consequences for his biological, substrate-dependent view of consciousness.

A case could even be made that my desktop computer is already minimally autopoietic. It draws power from its environment, maintaining a low-entropy state without which it will cease to function. It monitors itself for errors. It updates its drivers and operating system. It detects and repels viruses. It does not order and install replacement hardware, but it does continually sustain its intricate electrical configuration. Indirectly, though acting upon me, it does sometimes cause replacement parts to be installed. Alternatively, perhaps, we might view its electrical configuration as an autopoietic system and the hardware as the environment in which that system dwells.

My main thought is: *Autopoiesis* is a high-level, functional concept. Nothing in the concept appears to require implementation in a particular substrate. Nothing seems to prevent autopoietic processes in AI systems built along broadly familiar lines. An autopoietic requirement on consciousness does not appear in principle to rule out consciousness in standard computational systems.

Maturana and Varela themselves allow substrate flexibility. They write that

> The organization of a machine (or system) does not specify the properties of the components which realize the machine as a concrete system, it only specifies the relations which these must generate to constitute the machine or system as a unity (1980, p. 77).

It is clear from context that they intend this remark to apply to autopoietic as well as non-autopoietic machines (cf. Luisi, 2014).

Financial support

This research received no specific grant from any funding agency.

Competing interests

None.

References

Luisi, P. L. (2014). The minimal autopoietic unit. *Origins of Life and Evolution of Biospheres 44*, 335–338.

Maturana, H. R., & Varela, F. J. (1980). *Autopoiesis and cognition: The realization of the living*. D. Reidel.

Metaphysical confusion and biological naturalism

Murray Shanahan*

Department of Computing, Imperial College London, London, UK
m.shanahan@imperial.ac.uk

*Corresponding author.

doi:10.1017/S0140525X25103701, e361

Abstract

In his defence of biological naturalism and rejection of computational functionalism, Seth shows himself to be in the grip of dualistic intuitions. Thankfully, we can read Seth's paper as a worthwhile contribution to the ongoing society-wide conversation on how to talk about, think about, and treat advanced AI, while setting aside the metaphysical framing of his arguments.

Towards the beginning of his paper, Seth announces his assumption that "there is a fact-of-the-matter about whether

something is conscious" and "being conscious is not determined by social or linguistic consensus." He is far from alone in these convictions. They underlie almost every contribution to the contemporary philosophical debate around consciousness, which is no surprise because the metaphysical intuitions behind such pronouncements are deeply embedded in our conceptual make-up. To uproot those intuitions – to "show the fly the way out of fly-bottle," as Wittgenstein has it, or to dispel our "primal confusion," as Garfield frames the matter in a Buddhist context – is profoundly difficult (Wittgenstein, 1953/2009, §309; Garfield, 2015, pp. 9–11; see also Shanahan, 2010, ch.1).

First, let's note that how we *use* the word "consciousness" *is* a matter of social and linguistic consensus. Such consensus is presently lacking in the context of AI, and we'll come to that shortly, but let's stick to more settled usage for now. Our first step is to set aside the question of what consciousness "is," and to ask instead how words like "consciousness," and its many relatives (such as "awareness," "attention," and so on), are used. The crucial step, though, is to see that, having spelt out how these words are used, there are no residual philosophical problems. The challenge here is not the spelling out, but the insight that there is no more to be said, from a metaphysical point of view.

There undoubtedly is more to be said from a scientific point of view. The relationship between our usage of such words, the human (and animal) behaviour such usage relates to, and the brain processes that underlie that behaviour, certainly merits scientific study. Moreover, the fruits of this study will feed back into the way we use the words in question, and more generally into the nexus of social practices in which the use of those words is embedded. In this way, an issue such as the welfare of farmed octopuses can become a matter of societal and legislative concern, as it recently has in the UK. However, there is no need to resort to philosophical language to characterise what is going on here. We don't need to think procedures like this reveal what consciousness *is*, where the word "is" carries a metaphysical load.

Now let's extend this approach to the topic of artificial intelligence (Shanahan, 2024). The example of the octopus shows that the way we use words is subject to change, even philosophically difficult words like "consciousness" and its kin. The development of sophisticated artificial intelligence is putting our language under particular pressure, and there is no consensus yet on how it should adapt to these exotic mind-like artefacts. We are in a period of upheaval while scientists, scholars, and the public collectively figure out how to adapt and extend our ways of talking and thinking to accommodate them. Importantly, this process can only draw on what is public, what is manifest in our shared world, namely, behaviour and the mechanisms underlying that behaviour.

Let's recap. The philosophical strategy here is to replace the metaphysically loaded question of what consciousness *is* with questions about how words like "consciousness" are used. When confronted with scenarios where usage is not settled, such as the arrival of powerful AI, our strategy is to outline the process whereby usage changes, through a society-wide conversation informed by scientific investigation. But now for the hard part, which is to see that there is nothing more to be said, no residual philosophical problem. This is a great challenge because it forces us to confront our deep-seated dualistic intuitions, our primal confusion.

Seth betrays his own dualistic intuitions by asserting that "[w]e humans know we're conscious" and immediately issuing the caveat that "strictly, we know it for ourselves and infer it about others." The feeling that consciousness is something – some "thing" – that we only truly "know" from our own case and merely infer of others, is almost irresistible. It is, according to intuition, the most intimate form of knowledge by acquaintance: immediate, privileged, incorrigible, but above all private. It has led innumerable philosophers astray, from Descartes in the 17th century to Nagel and Chalmers in the 20th, and it continues to do so today in the 21st.

Buddhist philosophers have long known how to deconstruct the metaphysical thinking behind dualism, at least since the time of Nāgārjuna in the 3rd century (Siderits & Katsura, 2013; Garfield, 2015). In Western philosophy, potent antidotes to dualistic thinking can be found in the work of such 20th-century figures as Heidegger, Derrida, and Wittgenstein. The so-called private language remarks in Wittgenstein's *Philosophical Investigations* are especially pertinent.

Reflecting on the temptation to say "it is only from my own case that I know what the word 'pain' means," Wittgenstein invites us to imagine that "everyone had a box with something in it: we call it a 'beetle'. No one can look into anyone else's box, and everyone says he knows what a beetle is only by looking at his beetle" (Wittgenstein, 1953/2009, §293). He points out that, for all anyone would know, the contents of each person's box might be different. The contents could be constantly changing, or the box could even be empty. The private "thing" itself would play no part in our language: "it cancels out, whatever it is." Hopefully, the analogy with consciousness in general is obvious.

So, where does this leave us with respect to biological naturalism and to Seth's paper? While Seth frames his contribution as an argument concerning what consciousness "is," it can equally be framed as an argument for how we should use words like "consciousness." It is, that is to say, a contribution to the ongoing society-wide conversation about how the language of consciousness should accommodate sophisticated artificial intelligence. The paper's contribution is no less valid for being reframed in this way. It does, perhaps, render it a little more vulnerable to further critique. But that would be the subject of different commentary.

Acknowledgements. Thanks to Kerry Shanahan for extensive discussion on the topic of this commentary.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interests. The author is a part-time employee of Google, a developer and provider of large language model technology, and is also a shareholder in Alphabet, Google's parent company.

References

Garfield, J. L. (2015). *Engaging Buddhism: Why it matters to philosophy*. Oxford University Press.

Shanahan, M. (2024). Simulacra as conscious exotica. *Inquiry*. https://doi.org/10.1080/0020174X.2024.2434860.

Shanahan, M. (2010). *Embodiment and the inner life: Cognition and consciousness in the space of possible minds*. Oxford University Press.

Siderits, M., & Katsura, S. (2013). *Nāgārjuna's middle way: Mūlamadhyamakakārikā*. Wisdom Publications.

Wittgenstein, L. (1953/2009). *Philosophical investigations*, (4th ed.), trans. G. E. M. Anscombe, P. M. S. Hacker & J. Schulte. Wiley-Blackwell.

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

Aeroplanes also fly: Analytic functionalism and the possibility of machine consciousness

Henry Shevlin*

Leverhulme Centre for the Future of Intelligence, University of Cambridge, Cambridge, UK
henry.shevlin@gmail.com

*Corresponding author.

doi:10.1017/S0140525X25103919, e362

Abstract

Seth mounts a rich and persuasive case that *human* consciousness cannot be understood without life's distinctive machinery – autopoiesis, allostatic control, predictive processing under the free-energy principle, and substrate-entangled dynamics. I agree that this undercuts the expectation that humanlike consciousness will "come for free" with more intelligent machines. However, we should also be cautious about drawing extrapolations from features of biological consciousness – even essential features – to the possibility of exotic forms of consciousness in artificial systems.

Can consciousness be fully understood in computational terms, or must an adequate theory make reference to fine-grained, substrate-level features? This is a central question for contemporary cognitive science, one whose resolution has been made more urgent by the arrival of sophisticated anthropomimetic systems (Shevlin, 2025), able to emulate the coarse-grained behavioural capacities of human intelligence. In this landmark article, Seth appropriately frames the debate around the duelling approaches of biological naturalism and computational functionalism. However, as I will argue, there are nuances in how we apply these theories, and it is not necessarily the case that all forms of consciousness must be analysed using the same scientific-explanatory lens.

As an initial analogy, consider the phenomenon of flight. Over the course of evolution, flight has emerged independently at least four times, in insects, pterosaurs, birds, and bats. A proper understanding of flight in each case will make essential reference to fine-grained biological characteristics, from the demands it makes on metabolism to its associated and varied changes in muscles, tendons, and skeletal structure. However, we also know that flight is not restricted to biological organisms: human artefacts such as planes, helicopters, and rockets demonstrate that flight can be achieved in quite different forms.

The point of this analogy is hopefully clear: even if a proper account of *biological* consciousness cannot be given in purely computational terms, and must make reference to fine-grained, substrate-dependent features, this does not foreclose the possibility of different kinds of consciousness that lack these features. Even though AI systems may never instantiate the *same kinds* of conscious states as humans and animals, it is premature to say that they will never instantiate consciousness in any form.

The astute reader may recall the famous quip of Edsger Dijkstra that the "question of whether a computer can think is no more interesting than the question of whether a submarine can swim" (Dijkstra, 1986), and point out that the multiple realisability of flight is primarily a semantic matter, thus disanalogous to the scientific concept of consciousness. While I agree that there is more semantic flexibility in how we use the term "flight" than the term "consciousness," the analogy illuminates the possibility of a coarse-grained, functional account of consciousness encompassing quite different realisations, some biologically grounded, and others potentially realised computationally.

One concern with this pluralistic approach is that it seems theoretically disunified. If the category of flight has "powered motion through the air" at its core, what might provide a similar unifying function for capturing seemingly disparate phenomena like autopoietic biological consciousness and putative computational machine consciousness?

One answer is provided by the historically significant and recently resurgent family of views about consciousness known as *analytic functionalism* (Lewis, 1975). Unlike computational functionalism, analytic functionalism does not aim to understand the mind exclusively within the austere framework of Turing computation, but rather in the wider vocabulary of folk psychology. For the analytic functionalist, we should understand our lexicon of mental-state terms as folk-theoretical posits to help us better understand, predict, and explain the behaviour of a special class of physical system, namely intelligent goal-directed agents (Dennett, 1987). (As Seth rightly notes, of course, the intentional stance can go awry, in leading us to make naïve ascriptions of *humanlike* mentality to AI systems. However, this arguably reflects current quirks of Large Language Models (LLMs) such as hallucination, rather than any fundamental inapplicability of the stance to artificial systems.)

Analytic functionalism is appealing for a wide range of mental-state concepts. To give a toy example, what it is for a system to have *preferences* might be the tendency to seek out goals in a manner analysable through the lens of means-end folk psychology. Seth himself seems sympathetic to such a high-level account of intelligence.

But even if some mental states can be analysed through this lens, can we really approach consciousness in a similarly interpretative manner? Many will balk at this suggestion, taking it as given that doing justice to consciousness requires us to identify a singular underpinning biological or computational kind.

However, analytic functionalism need not commit us to eliminativism: consciousness may be a "real pattern" with distinctive explanatory power, even while lacking a singular biological or computational basis (Dennett, 1991). Moreover, "deep realism" about consciousness (Shevlin, 2024) should itself not go unquestioned, not least because of the lurking hints of dualism it encourages. As Shanahan (2024) puts it, "We must resist the temptation to ask whether [a system is] conscious as if consciousness were something whose essence is out there to be uncovered by philosophy (or neuroscience) while simultaneously having an irreducibly private, hidden aspect." What if the best way to analyse our concept of consciousness (and the ethical implications that go with it) is in broader ecological terms, as a way of understanding how intelligent and self-aware systems move through the world, interact with others, and conceptualise their own status within it?

The science of consciousness remains in its infancy, and Seth's article provides a powerful reminder of our limited but growing

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

understanding, even of biological consciousness. However, even as we refine our theories of human and animal consciousness, we should be cautious not to extrapolate too quickly from contingent features of its realisation in beings like us to the wider and more inscrutable space of possible minds.

Financial support. This work was supported by the Leverhulme Centre for the Future of Intelligence, Leverhulme Trust, under Grant RC-2015–067.

Competing interests. None.

References

Dennett, D. C. (1987). *The intentional stance*. MIT Press.
Dennett, D. C. (1991). Real patterns. *The Journal of Philosophy*, *88*(1), 27–51. https://doi.org/10.2307/2027085
Dijkstra, E. W. (1986). *On computers and society*.
Lewis, D. (1975). Psychophysical and theoretical identifications. In C. Y. Cheng (Ed.), *Philosophical aspects of the mind-body problem* (pp. 43–53). University of Hawaii Press. https://doi.org/10.1515/9780824890575-005
Shanahan, M. (2024). Simulacra as conscious exotica. *Inquiry*, 1–29. https://doi.org/10.1080/0020174X.2024.2434860
Shevlin, H. (2024). Consciousness, machines, and moral status. In A. Strasser (Ed.), *Anna's AI Anthology: How to live with smart machines?* (pp. 279–307). Xenomoi Verlag.
Shevlin, H. (2025). The anthropomimetic turn in contemporary AI. In M. Rangeley, & N. Fairfax (Eds.), *The artificial intelligence revolution*. Springer Nature.

Why Current A.I. Can't Be Conscious: It's not about predictive processing or the free energy principle; it's about the lack of self-organization and multiscale complexity that enable adaptability

Michael Silberstein*

Professor of Philosophy, Professor of Science, Society, and Technology, Director of the Cognitive Science Program, Core Faculty Member in the Neuroscience Program, School of Engineering and Computer Science, Elizabethtown College, Elizabethtown, PA, US
silbermd@etown.edu

*Corresponding author.

doi:10.1017/S0140525X25103610, e363

Abstract

A.I. isn't conscious because consciousness requires properties that only biological agents possess, which A.I. lacks. What A.I. lacks that is relevant for consciousness is not predictive processing or the minimization of free energy. A.I. lacks embodiment, direct coupling with the environment, multiscale complexity, and self-organization, e.g., plasticity, degeneracy, robustness, and the properties that enable living organisms to be highly adaptive to environmental changes.

Seth's argument has two parts. First, a rejection of computational functionalism and silicon substrate flexibility regarding consciousness. Second, a form of biological naturalism (BN) that emphasizes active inference, cybernetics, autopoiesis, and the free energy principle.

I agree with the first part. As for the second, I agree that A.I. is unlikely to ever become conscious because consciousness requires properties biological agents have that current A.I. doesn't. However, modulo the focus on self-organization, I disagree with Seth about what those properties are. Furthermore, it's okay to avoid the hard problem when it comes to seeking neural correlates or explaining the phenomenological contents of conscious states, but not when it comes to deciding if some artifact or alien is conscious. I'll unpack all these points and address Seth's defenses.

Predictive processing, the "controlled hallucination" idea that what we experience is the brain's best guess about "what's out there," runs right into the hard problem. Why should there be any phenomenology at all associated with such a function, and how can a piece of insensate meat produce such experiences without invoking a spooky form of strong emergence that Seth rejects? Also, as Seth notes, predictive processing is perfectly compatible with the kind of computationalism he rejects: "Certainly, predictive processing (and Bayesian inference) can be utilised as an algorithm, and often is."

Seth has two defenses. For the hard problem, he says we should just ignore it and focus on the "real problems" of seeking neural correlates and explaining the contents of various phenomenological states. He thinks the hard problem will eventually be deflated in this fashion. Fine advice for everyday cognitive neuroscience, but not when it comes to making claims that the brain's active inference function *causes* consciousness. Again, if that's true, Seth can't escape spooky strong emergence. Seth is on record as accepting the kind of BN that says brains cause consciousness, so this is a real issue for him.

Hard problem aside, predictive processing doesn't explain the specific phenomenology of conscious states. When the brain decides it's looking at an elephant, why does it generate the sights, sounds, and smells that it does? Given that this story presupposes we are just brains locked in dark skull-vats making guesses about the causes of sensory signals, we can't appeal to the "real" external world for an answer. The deeper point here is that the mapping between certain sensory signals and phenomenological correlates is one-to-many; there is no necessary connection or necessity here. Other brains (humans) might attach radically different phenomenology to the same sensory signals, and yet, it could still have the same evolutionary advantage. A better story about what constrains such mappings is needed. Think of the three blind men and the elephant.

Seth's second defense is this: "Putting all this together suggests that the (minimisation of) free energy driving metabolism, autopoiesis, and perception is not just metaphorically equivalent, but is in some physical sense the same thing." He thinks, "In this sense, predictive processing becomes a dynamical, non-computational theory. This process can be abstracted computationally, but these algorithms are pale shadows of the embodied and embedded processes unfolding in biological brains and bodies." Regarding the identity claim of the free energy principle with predictive processing, I see no reason to buy this since there are purely

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

physical coupled systems and single-celled organisms that obey the free energy principle that don't appear to be engaged in active inference. And there is little reason to believe such things are conscious. Minimization of free energy, active inference, and being conscious can all come apart. Such an identification is a category mistake. Minimization of free energy is necessary for life and self-organization, but not sufficient.

Seth's defense also borrows heavily from the idioms of ecological psychology, embodied cognition, and enactivism. These traditions can also adopt some non-loaded form of predictive processing. But here the idea would be that embodied agents coupled directly to their environments are engaged in actively exploring their environments, the space of available affordances, and "sense-making," not brains in skull-vats blindly roaming the noumenal "I-know-not-what." The point is, given Seth's other brain-centric BN-related commitments, he cannot consistently avail himself of this response.

Seth is, however, correct that the self-organization of biologically embodied agents is key. What do all conscious organisms on Earth have in common? They are all self-organizing embodied agents coupled directly to their environments. Such agents have varying sensory apparatus and nervous systems made of multiscale networks that exhibit various forms of plasticity, degeneracy, and redundancy at all those scales (Silberstein, 2021, 2022, 2023). This bespeaks a non-functionalist kind of biological multiple-realizability.

The brain is best viewed not in old-school computational or modular terms, not as a prediction machine, but in terms of overlapping, spatiotemporally widely distributed, and functionally integrated multiscale dynamical networks. In such networks the "function" and behavior of any given element, such as neurons, glia cells, neurotransmitters, RNA networks, hormones, neural circuits, brain regions, connectomes, synchrony, topological networks, etc., are determined dynamically by multiscale relationships with other elements throughout the brain and contextually given constraints such as cognitive tasks being undertaken, bodily activity, and environmental interactions (Silberstein, 2023).

Brains are not "neuron machines" or computers in any deep sense, and conscious agents aren't their brains. Embodied, conscious agents are often part of larger distributed bio-psycho-social networks. Such networks and their properties maximize adaptability to environmental changes and maximize the capacity for learning. Such embodied agents need not be made of biological wetware, but it would have to be some other substrate that also maximizes the capacity for adaptability and learning – plasticity and robustness.

On this view, cognition and consciousness are inextricably linked. There is no possibility of building A.I. with AGI that isn't fully conscious and vice-versa (Silberstein & Chemero, 2012). Brains are a necessary component, but they are not the "machinery" of consciousness. The contents of conscious experience are a function of the embodied agent's brain states and various direct interactions with the multiscale and multi-domain networks that make up their environments.

Financial support. None.

Competing interests. None.

References

Silberstein, M. (2023). The multiscale contextual emergence of neural dynamics, cognition, and action. *Journal of Multiscale Neuroscience*, *2*(1), 140. https://doi.org/10.56280/1567588756.

Silberstein, M. (2022). *Emergence in context: A treatise in 21st century natural philosophy*. Oxford University Press.

Silberstein, M. (2021), Constraints on localization and decomposition as explanatory strategies in the biological sciences 2.0. In F. Calzavarini, & M. Viola (Eds.), *Neural mechanisms: New challenges in the philosophy of neuroscience* (pp. 363–395). Springer Neuroscience.

Silberstein, M. & Chemero, A. (2012). Complexity and extended phenomenological-cognitive systems. *Topics in Cognitive Science*, *4*(1), 35–50.

Real artificial consciousness is, of course, incompatible with biological naturalism

Mark Solms*

Neuroscience Institute and Psychology Department, University of Cape Town, Rondebosch, South Africa
mark.solms@uct.ac.za

*Corresponding author.

doi:10.1017/S0140525X2510352X, e364

Abstract

There is a curious contradiction in Seth's argument, or a degree of ambiguity which proves to be fatal. He endorses "biological naturalism," but he says also that it is possible to engineer "real artificial consciousness." Such an artificially consciousness system need not be carbon-based, he says, so long as it is embodied in a system that is *artificially* "alive."

I am glad to have been invited to comment on this article because, when I reviewed it for BBS, although I thought it presented an excellent argument, I did not fully *agree* with it. Peer review isn't a matter of agreement; but, happily, peer commentary is.

Let's go directly to the heart of the matter. I agree with Seth when he says that real artificial consciousness (to use his felicitous term) "would need to be (in some relevant sense) living AI, though not necessarily carbon-based" (p. 13). Everything in this statement, which Seth makes more than once in his article, pivots on the question as to what the *relevant sense* is in which an artificial agent needs to be "alive."

If it doesn't need to be carbon-based, it doesn't need to be biological. Surely this negates the essence of the term "biological naturalism," and thereby negates a crucial aspect of Seth's argument. In short, it sets aside everything that he says about the possibly necessary-for-consciousness properties of nerve cells, metabolism and the like.

This leaves us with the stark fact that the rest of his commitment to so-called "biological" naturalism in relation to consciousness is

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

"grounded in active inference, cybernetics, autopoiesis, and the free energy principle" (p. 23). These remaining commitments undercut everything he says about substrate dependence, because *there is nothing substrate dependent about active inference, cybernetics, autopoiesis, and the free energy principle*. Seth admits as much when he says that "physically embodied robots could have energetic and structural integrity requirements entailing real though non-biological forms of homoeostasis and allostasis (but not autopoiesis)" (p. 27).

Since he admits this much, I need only show that autopoiesis, too, is substrate-independent. Consider, for example, the artificial and decidedly non-biological – indeed *virtual* – self-organising dynamical systems that Friston engineered (Friston, 2013). To test whether these artificial systems were autopoietic, he selectively "lesioned" their sensory, active, and internal subsystems (i.e., he prevented these subsystems from influencing their neighbours). This produced a catastrophic loss of configurational integrity, as was illustrated in Figure 5, which showed structural disintegration of the previously self-organising systems, as the integrity of their Markov blankets was lost and their internal elements were extruded into the environment. Friston described this as "a testament to [their] autopoiesis or self-creation – in the sense that [their] self-organised dynamics are necessary to maintain [their] structural or configurational integrity" (ibid., p. 10).

Together with a team of physicists, computer scientists, roboticists, and fellow neuroscientists, I am currently attempting to engineer a real artificial consciousness *which is substrate independent* but which embodies all the functional (and in a broad sense, computational) dynamics that Seth committed to above, including non-biological forms of homoeostasis and allostasis. (Our agent is driven by a sophisticated form of active inference; it is chronically trying to reduce its uncertainty about how to meet its own existential needs in the environment within which it is embedded, and thereby to stay "alive.") In *this* relevant sense, our agent is artificially alive. It is genuinely agentic, in the way that Seth points out most "agentic AI" is not, in that it possesses what he calls *endogenous* goals (p. 27). It has "skin in the game" (p. 8); it "gives a damn" (Haugeland, 1998, pp. 47, 60).

Towards the end of his article, Seth argues against the engineering of such systems. He says that "the development of real artificial consciousness for its own sake should not be an explicit goal" (p. 29). He makes this point on the alarming grounds that creating real artificial consciousness risks a mass inauguration of new forms of suffering (p. 28) and that real artificial consciousness might endow a system with its own interests and values, beyond those installed by the system's designers (p. 29). Why, then, is my team trying to develop a conscious AI?

This question is best answered by quoting Feynman: "What I cannot create, I do not understand." (This was written on Richard Feynman's blackboard at the time of his death in 1988.) The scientific problem as to how consciousness arises will never be solved by means of philosophical debates and thought experiments, or by adding more plausible theories to the plethora of mutually contradictory ones we already have. In short, we do indeed need to transform the "hard problem" (Chalmers, 1995) into what Seth calls the "real problem." But this requirement can only mean that we must subject our mechanistic, causal theories of consciousness to the same rigorous tests of falsification that we use in the rest of science. This can be done most directly by means of concrete experiments in which we attempt to build artificially conscious agents and then test whether we have succeeded or not.

I believe that we have reached a stage in consciousness science where the most important question is not "what should we (not) do?" but rather "how will we know (or agree) when we have done it?". To this end, my principal co-investigator and I are drawing up a shortlist of functional and behavioural criteria that we hope a reasonable number of consciousness scientists and scholars will be able to agree upon (Solms and Shock, in preparation).

We have committed ourselves publicly (Solms, 2021) to stopping our research programme just as soon as such criteria have been met, so that the relevant stakeholders can take collective responsibility for how we should all proceed from that point onwards. In other words, we fully recognise the risks that Seth identifies; but if artificial consciousness can be created, it *will* be created. So, it is best that it is created by pure scientists like us, "for its own sake," rather than by commercial interests, to mention just one of the most obvious worrying alternatives.

Acknowledgements. The opinions expressed in this article are my own; they are not necessarily shared by other members of my research team.

Financial support. The research described in this paper has received funding from the Oppenheimer Memorial Trust and from Conscium.

Competing interests. The author sits on the scientific advisory board of Conscium and of PRISM (Partnership for Research Into Sentient Machines), London.

References

Chalmers, D. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, *2*, 200–219

Friston, K. (2013). Life as we know it. *Journal of the Royal Society Interface*, *10*, 20130475, https://doi.org/10.1098/rsif.2013.0475

Haugeland, J. (1998). *Having thought: Essays in the metaphysics of mind*. Harvard University Press

Solms, M. (2021). *The hidden spring: A journey to the source of consciousness*. London Profile Books; New York: WW Norton.

Solms, M., & Shock, J. (in preparation). Invited contribution to a special issue of *Journal of Consciousness Studies* on consciousness and current AI.

What can the free energy principle tell us about artificial consciousness?

Wanja Wiese*

Institute for Philosophy II, Ruhr University Bochum, Bochum, Germany
wanja.wiese@uni-bonn.de

*Corresponding author.

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

doi:10.1017/S0140525X25103750, e365

Abstract

Does the free energy principle support the view that consciousness is substrate-dependent? Based on previous work (Wiese, 2024), I argue that the free energy principle can be used to clarify conditions under which a system realises, rather than merely simulates, a causally relevant property (such as consciousness). This helps to address a weakness in Seth's argument.

I argue that the inference from predictive processing and the free energy principle to a form of substrate-dependence is not as straightforward as Seth suggests. Instead of *justifying* a form of substrate-dependence, it must *presuppose* that consciousness is substrate-dependent. As a remedy, I suggest further clarifying the difference between simulating and realising consciousness.

If the free energy principle and predictive processing support a form of substrate-dependence, they must specify what non-computational aspects are required to realise consciousness, and why. Seth formulates two core ideas that speak to this. One is that the mechanisms underlying conscious experience are, according to predictive processing, entangled with homeostatic processes that are characteristic of living organisms (sect. 4.1). The other is that, according to the free energy principle, the process of minimising variational free energy is *identical* to homeostatic processes (sect. 4.2).

One could object that this only contingently holds for conscious living organisms. Simply put, it is an evolutionary advantage if an organism can instantiate consciousness and cognition (in part) by redeploying mechanisms that it also uses to stay alive. But this leaves it open whether artificial systems, which are not bound by the constraints of natural selection, can realise consciousness in different ways – in particular, in ways that are less energy-efficient, but are devoid of the homeostatic processes found in living organisms.

Does the free energy principle suggest a stronger link between conscious experience and homeostatic processes? After all, if they are "in some physical sense [. . .] the same thing" (sect. 4.2), then it might seem that you cannot implement one without the other. But Seth is careful to avoid biopsychism, i.e., the view that *all* living organisms are conscious (sect. 4.0). That is, some homeostatic processes are not identical to conscious experience.

Furthermore, it seems that the identity between homeostatic processes and the information processing associated with consciousness must be qualified: the free energy principle provides two equivalent descriptions of certain self-organising systems – one in terms of physical dynamics, the other in terms of a form of information-processing. With respect to systems for which this equivalence holds, the information processing (minimisation of free energy) underlying consciousness may be numerically the same processing that undergirds homeostatic processes. But we cannot infer that this must also be true for conscious artificial systems. This is because the computations that are implicated in minimising variational free energy can be performed using digital computers. Hence, assuming that these computational processes *must* be implemented in the same way as they are implemented in conscious living organisms, in order to realise consciousness, would be unwarranted. Consequently, we are still left with the question of whether performing the right computations is sufficient for replicating consciousness or not.

To make progress, we need to address the following questions: why might the way in which conscious living organisms implement the minimisation of free energy be relevant to realising consciousness? What are the relevant properties of conscious living organisms that distinguish simulating consciousness from realising consciousness?

As I argue in Wiese (2024), conscious living organisms instantiate causal relations that are not instantiated by computer simulations of conscious agents (at least not when implemented in conventional hardware). The assumption that causal relations are relevant to realising (instead of merely simulating) consciousness is not specifically an assumption about realising consciousness. Rather, it is in line with more general considerations (which I will not recapitulate here) to the effect that a key difference between merely simulating a phenomenon and realising it lies in the respective forms of causation that are instantiated (see sect. 3.7). That is, a mere simulation will not replicate all relevant causal relations that are required for realising the phenomenon. In particular, one can argue that the (virtual) causation found in computer simulations is illusory, not real (Kohár, 2024). To be sure, a computer that implements a computational process will thereby also implement a causal process. But the computer will, in most cases, not thereby replicate the same causal relations that are instantiated in the simulated target.

Conway's game of life (Gardner, 1970) provides a useful analogy. It can be regarded as a coarse-grained simulation of an evolutionary process, in which over- or underpopulation can cause the death of individuals. For instance, if a living cell in the game of life is surrounded by less than two or more than three living neighbours, it dies. Notably, these causal relations don't exist in the computational implementation, because there are no direct causal interactions between the parts of the computer that represent the states of individual cells (for a similar point, see Kohár, 2024). This can be seen even more clearly when one imagines implementing the game by hand, using a game board and tokens. The "birth" and "death" of individual tokens are not caused by other tokens, but by the human being who is implementing the rules of the game.

Consciousness, by contrast, is real in the sense that it is causally efficacious. In a (conventional) computer simulation of conscious agents, causal interactions between the agents' internal states and the environment are represented, but not replicated. In a robot (with conventional hardware), the information processing associated with consciousness is instantiated, and it is also causally relevant (especially for outward behaviour), but internal causal relations that are associated with consciousness (e.g., interactions between mental processes) will be much sparser or absent. To the extent that conscious experience depends on causal interactions between unconscious mental processes (e.g., processes that integrate sensory information from multiple modalities; see also sect. 4.5), even a robot (with conventional hardware) will at most simulate conscious experience.

In sum, the free energy principle can help to discern forms of causal relations that are characteristic of conscious living organisms, but are absent in mere simulations. *That* a difference in causation is relevant to the distinction between simulating and realising a phenomenon is an additional assumption, but one that can be justified independently of considerations about

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

consciousness. Hence, the free energy principle does support the view that consciousness is substrate-dependent, without begging the question.

Acknowledgements. I am grateful to Leonard Dung for feedback on a draft of this comment.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interests. None.

Reference

Gardner, M. (1970). The fantastic combinations of John Conway's new solitaire game "life." *Scientific American*, *223*, 120–123.

Kohár, M. (2024). Illusionism about virtual causation. *Inquiry*, 1–26. https://doi.org/10.1080/0020174X.2024.2403142

Wiese, W. (2024). Artificial consciousness: A perspective from the free energy principle. *Philosophical Studies*, *181*, 1947–1970. https://doi.org/10.1007/s11098-024-02182-y

Author's Response

The stuff matters: Consciousness, computation, and biology

Anil K. Seth[a,b]*

[a]Sussex Centre for Consciousness Science and Department of Informatics, University of Sussex, Brighton, UK and [b]Canadian Institute for Advanced Research, Program on Brain, Mind, and Consciousness, Toronto, Canada
a.k.seth@sussex.ac.uk

*Corresponding author.

doi: 10.1017/S0140525X26106906, e366

Abstract

I was delighted, challenged, and somewhat overwhelmed by the breadth and depth of the commentaries elicited by the target article. I am very grateful to every one of my commentators for offering such thoughtful perspectives. Here, I attempt to isolate some common themes and offer some thoughts in response. One conclusion is immediately evident: it is increasingly urgent to think clearly about the prospects for conscious AI and about the consequences of seemingly conscious AI. The commentaries also flesh out the potential for an exciting research programme investigating the substrate dependence of both consciousness and cognition.

> "Yes, thinking meat! Conscious meat! Loving meat. Dreaming meat."
>
> *Terry Bisson, They're Made Out of Meat (1991)*

R1. Introduction

The issue of "conscious AI" is of immense and increasing importance both within and beyond academia. The topic is rarely out of the news. Influential voices warn of the dangers of both false negatives (failing to see consciousness in AI if it in fact exists) and false positives (wrongly attributing consciousness to AI) (Bengio & Elmoznino, 2025; Long et al., 2024).

As the debate unfolds – in part through this collection of target articles, commentaries, and response – we should keep in mind that what we collectively think about the plausibility of conscious AI already carries enormous importance, whatever the reality may be. For example, widespread credence for the idea that AI is (or could be) conscious motivates calls for "AI Welfare" (Long et al., 2024), which in turn adds momentum to the idea that AI systems could have "rights." Frontier AI systems are already showing signs of (loosely speaking) "self-preservation," for example, by attempting to disable oversight systems (Meinke et al., 2024). This is clearly dangerous behaviour. As AI systems become more autonomous and more capable, such dangerous behaviours are likely to increase. If AI systems are granted "rights," we may be legally prevented from turning them off, with potentially disastrous consequences. On the other side, of course, is the argument that to deny conscious entities rights (such as the right to continue existing) would be to commit a grave moral error, following the long tradition of similar errors human beings have made in the past.

This is one among many reasons why we should care about the prospects for conscious AI and about the consequences of conscious-seeming AI. My own view remains that conscious AI is extremely unlikely, especially for silicon-based digital systems. (As in the target article, I reserve the term "AI" for such systems unless specified otherwise.) But without a consensus understanding of the sufficient conditions for consciousness, things remain uncertain. This is why the breadth and depth of commentary are so welcome. Even though the question of "AI consciousness" cannot yet be settled definitively, the question of whether (or to what extent) consciousness is substrate dependent is emerging as a key research challenge, carrying both immediate and long-term ethical, social, economic, and political consequences. A robust and informed debate is therefore a very good thing.

R1.1 Preliminaries

My target article progressed through four stages: (i) identifying psychological biases that might lead to over-attribution of consciousness to AI; (ii) setting out challenges to computational functionalism; (iii) making a case for biological naturalism, and (iv) discussing ethical implications of AI that either is, or seems to be, conscious. The majority of commentaries focused on (ii) and (iii), and I will focus my response on these topics (Sections R2–R5), returning to (i) and (iv) later on (R6).

To frame things, I will assume that an intact, healthy, embodied, and embedded human brain, when in certain states, is sufficient for consciousness. One can then ask which aspects of this system are necessary and/or sufficient for consciousness. Potential answers can be arranged on a scale of (roughly) granularity or "level of description." At one extreme, proponents of panpsychism claim that no aspects of such a system are necessary. Consciousness, according to panpsychism, is a fundamental property of the universe. At the other extreme, some (mostly outdated) views might assert that human-level rational thinking or language is inseparable from consciousness; the historical anchor here is Descartes.

On this scale, computational functionalism (CF) is the claim that the computational level of description is sufficient for consciousness. Implement the right kind of computation (in whatever substrate is capable of this), and there will be consciousness.

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

Biological naturalism (BN) is the claim that other (primarily lower-level) properties of biological systems are necessary, though not necessarily sufficient, for consciousness. Put this way, both CF and BN are placing bets on what counts. As with all bets, nothing is certain. My target article (henceforth CAIBN) attempted to shift credences away from CF and towards (some version of) BN.

Three further points will finish setting the scene. First, some commentators imply that my arguments against CF depend on BN being true (e.g., Dung and Michel). This is not the case. While the arguments *contra* CF and *pro* BN reinforce each other, they do not depend on each other. Scepticism about CF opens space for other ideas about necessary and/or sufficient conditions for consciousness, but BN is one among many alternatives [integrated information theory, for example, offers another (Albantakis et al., 2023); see also Piccinini]. At the same time, as Lerchner notes, evidence for BN does constitute evidence against CF, at least for standard digital computation. And if CF is broadened to include other (perhaps more "biological") forms of computation (CAIBN 3.4, R3.3), then it loses the substrate flexibility needed to license claims of consciousness in standard digital systems, undermining its own rationale for computational sufficiency in the process.

Second, the bets placed by CF and BN are asymmetric. CF is a claim about sufficiency across substrates. This imposes a high evidential burden which I believe remains unmet; CF mostly remains assumed rather than explicitly argued for (R2, R3). BN is the logically lighter claim that some biological substrate properties are necessary for consciousness; multiple accounts are available for why this might hold (R4, R5). It's worth recalling that BN comes in various flavours, from "weak" (the functional organisation necessary for consciousness requires biological implementation) to "strong" (consciousness depends on intrinsic properties of biological substrates); see CAIBN 5.0 and CAIBN Table 1. Neither entails carbon chauvinism.

Finally, and crucially, for consciousness in AI to be rendered highly implausible, it's enough for CF to be false. It is not necessary to show that BN is true.

R2. Holding out for CF

Many commentaries sympathise with at least some aspects of my critique of CF (for example, Metzinger, Lerchner, Piccinini, Wiese, and Cao). But there are some who leap to its defence. To respond, it will be useful to rehearse two distinctions. First, CF is not the same as functionalism in general (CAIBN 3.0). The latter is the view (broadly) that mental states, including conscious mental states, are properties of the internal functional organisation of a system. The former, by claiming that *computations* are sufficient for conscious mental states, is a subset of functionalism. Buying functionalism doesn't mean buying CF. Indeed, weaker flavours of BN are varieties of functionalism in general (stronger versions may be compatible too, CAIBN 5.5). From the other side, retreating from CF to functionalism in general (or broader forms of computation, R3.3) sacrifices the substrate neutrality which licenses conscious AI and which motivates the very idea of CF (R3.1).

Second, questions about functionalism (whether computational or otherwise) should be distinguished from questions about the *functions of* consciousness (broadly, what consciousness is *for*; Dolega & Cleeremans). While functionalism, through its focus on mechanisms, provides a natural framework for addressing questions about the functions of consciousness, there is no direct line from functionalism to functions. One can imagine consciousness having functions without buying into functionalism; equally, functionalism being true doesn't determine what (if any) functions consciousness has.

R2.1 Defending CF

I'll start with Michel. His commentary puts the case for CF particularly strongly, and my responses pertain not just to him but to pro-CF views more generally. Broadly, Michel encourages us to "bet on functionalism" mainly because, in his view, it has been reasonably successful at explaining other psychological capacities, such as memory and language, and because it's doing a decent job of explaining various aspects of consciousness. Examples include the difference between conscious and unconscious mental states, binocular rivalry, and the like.

There are five main problems with Michel's argument. First, he thinks my case against CF depends on BN being true: "Seth's arguments do not establish the metaphysical necessity of biology and thus are harmless against computational functionalism." But as clarified above, the primary arguments against CF do not depend on BN being true.

Second, he doesn't consistently distinguish functionalism from CF. Betting on CF is riskier than betting on functionalism. Even for cognitive processes besides consciousness, purist computational views are increasingly under pressure. A key argument in CAIBN (which doesn't depend on BN being true) is that the brain's functional properties are likely inextricable from their material substrate, which undermines the view that these properties consist solely in (Turing) computations [see also (Brette, 2026; Chirimuuta, 2024), R3]. From the other direction, the turn in cognitive science towards embodiment, embeddedness, and dynamics [(Newen, De Bruin, & Gallagher, 2018); see also Silberstein, Bowes, Mitchell & Jennings] also suggests that substrate-independent computations are insufficient to explain cognitive properties. Both lines of argument suggest that broader functionalist frameworks are needed. When it comes to consciousness, the pressure increases: consciousness is primarily about *being* (or *feeling*) rather than *doing*, and so it is *a priori* less well suited to modes of explanation – such as substrate-independent computation – that are primarily about doing (though see Fleming & Shea, R5.4).

The third problem: even if computational explanations are instrumentally useful in explaining some specifically consciousness-related phenomena in humans, it does not follow that computation provides a sufficient basis for consciousness. We already know that humans are capable of consciousness. For the key question of whether AI can be conscious, it is not enough to use CF instrumentally: *whether or not CF is true becomes critical.* Rather than simply betting on CF, positive arguments in its favour are needed.

Fourth, and related: Michel's argument is a variety of induction, but the key inductive step is unsupported. While psychological capacities like memory and language can, in some form, be uncontroversially characterised in terms of (potentially computational) functions, the same cannot be said for consciousness. Assuming that computationalist explanations will work as well for the latter as for the former requires assuming that consciousness is relevantly similar – i.e., that computations are sufficient – which is precisely the claim at issue.

Finally, Michel argues that there are no reasonable alternatives to functionalism. He dismisses biological theories and claims that predictive processing is defined computationally. However, one of the exciting outcomes of this set of commentaries is a promising

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

research agenda for BN, which includes (but does not depend on) non-computational versions of predictive processing (R5). I am reminded here of a quote from Bernard Williams: "Historical understanding ... can help with the business, which is quite certainly a philosophical business, of distinguishing between different ways in which various of our ideas and procedures can seem to be such that we cannot get beyond them, that there is no conceivable alternative" [(Williams, 2008), p.496].

Moving on, Blum & Blum seem to strongly advocate for CF (and carefully distinguish it from functionalism in general) via their "Conscious Turing Machine." As the name suggests, their model adheres to the principles of standard discrete digital computation, though there is also emphasis on embodied and environmentally embedded aspects. Their model also incorporates a variety of elements from several prominent theories of consciousness, but they do not explain why an implementation of their model would instantiate consciousness, rather than merely instantiate a computational model of consciousness. Note that even for "computational" theories of consciousness, the computational framing is arguably convenient rather than necessary (CAIBN 3.6). Like many proponents of such theories (Dehaene, Lau, & Kouider, 2017), Blum & Blum seem content to assume CF rather than justify it. When it comes to conscious AI, the circularity of their argument is exposed by their "Axiom 1," which asserts that computational processes of broadcast and attention result in a "unitary experience" – but this axiom includes the very claim that needs justifying.

While Blum & Blum think that "conscious AI is inevitable," Dung is more circumspect, noting correctly that arguments on both sides (CF and BN) remain inconclusive. However, he overlooks the asymmetry of the corresponding claims: CF is about substrate-general sufficiency, while BN is about biological necessity (R1.1). He goes on to provide five arguments supposedly in favour of CF. However, three of these do not seem relevant to computational sufficiency: folk intuitions about AI consciousness (R6.1), desirability for conceptual engineering, and supposed weaknesses of BN. The remaining two – evolutionary function and explanatory unification – are more appealing, but as Dung concedes, they fall far short of establishing CF. They also seem compatible with BN, or some other form of substrate dependency (R4, R5; Jablonka et al.). Altogether, his arguments offer some reasons why one might want CF to be true, rather than evidence that it is. Having said this, Dung calls me out on a logical error, which I'll deal with next.

R2.2 CF is not strictly necessary for conscious AI

Dung points out that my statement that "Conscious AI requires [...] computational functionalism to hold" is strictly incorrect. Chrisley makes the same point. They are right, and I'm grateful to both.

Chrisley offers an analogy: "McLarenism" – the thesis that owning a McLaren is sufficient for winning a Grand Prix. This thesis is clearly false, since other teams can and do win races. But McLarenism being false does not mean that a McLaren can't win. By analogy, even if CF is false, AI can still be conscious. While entertaining, the analogy is misleading. Only one car can win a race, which makes sufficiency a relational condition rather than an intrinsic condition – and consciousness is (though see Shanahan) an intrinsic property.[1] A more appropriate analogy, from the perspective of sufficiency, would be the thesis that owning a (properly functioning) McLaren is sufficient but not necessary to go faster than 200 mph. This thesis is likely true, removing the force of the analogy.

Nonetheless, both Chrisley and Dung are correct that, strictly speaking, CF being false does not exclude conscious AI (again, for standard digital systems). But it does not matter. It is indeed conceivable for AI to be conscious even if CF is false, if some other property turns out to be sufficient. However, it is hard to imagine what such a sufficiently substrate-flexible property might be, if not computation. The possibility of conscious AI is almost always motivated, implicitly or explicitly, by the assumption that substrate-flexible computations provide a sufficient basis for consciousness (Butlin et al., 2023). In short, conscious AI does not require CF to be true, but without CF being true, there are no good reasons to believe it is plausible.[2]

R3. The remit of computation

Because CF is the claim that computation (of the right kind) is sufficient for consciousness, much turns on what counts as "computation." There are three main options: (i) standard digital (Turing) computation; (ii) narrower forms in which fewer things compute (perhaps none); and (iii) broader forms according to which many things compute (perhaps everything). The challenge for CF is that it turns out to be implausible whichever option one favours. Broader forms sacrifice the substrate neutrality which motivates the very idea of CF (and which licenses claims of conscious AI). Narrower forms are either poorly motivated or lacking in discriminatory power. And the standard view is a poor description of the brain, such that CF loses its motivation and credibility.

In this section, I'll discuss the three views of computation before moving on to related issues: embodiment/embeddedness, time, and the crucial distinction between simulation and instantiation.

R3.1 Standard digital computation

In standard digital (Turing) computation, a physical system computes when it implements an algorithm, understood as a sequence of steps mapping one (discrete, denumerable) string of symbols to another (Turing, 1936). Thanks to its emphasis on medium/substrate independence, Turing computation enshrines a sharp conceptual distinction between software (algorithm) and hardware (substrate), a distinction which the von Neumann architecture made concrete through its physical separation of processing and memory. This standard view is not limited to algorithms implementing old-fashioned symbol manipulation. The artificial neural networks which power current AI may seem more "brain-like," but they too live in Turing World, and they too inherit the sharp separation of software from hardware.[3]

CF is largely motivated by this standard view. The substrate neutrality of this view licenses the idea – central to CF – that the *computations themselves* are sufficient. And here lies the challenge (CAIBN 3.3-3.5). Evidence that substrate properties matter for brain function is evidence against the sufficiency of substrate-neutral computations, which in turn is evidence against CF [see also (Chirimuuta, 2024)]. In other words, the inability to separate what brains *do* from what they *are* undermines the assumption that everything that brains "do" – including consciousness – is a matter of substrate-independent computation.

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

R3.2 Narrowing computation

Both Fields & Glazebrook and Lerchner argue that computation is inherently observer dependent: that whether a physical system is said to compute or not depends on how an external observer interprets its dynamics.[4] As Lerchner points out (echoing note 19 in CAIBN), if physical computation is observer-dependent, this puts CF in a bind: either consciousness is observer-dependent ("not a real-world phenomenon"), or it is not a matter of computation [see also (Lerchner, 2026)]. Moreover, observer-dependent computation lacks the discriminatory power to determine whether a candidate system has properties which depend constitutively on computation – a determination which CF requires.

Such reasons lead Fields & Glazebrook to suggest reorienting questions about whether AI could be conscious to questions about what such systems could be "aware of," if they are conscious (see also Shevlin). They generously note that this echoes my "real problem" approach to consciousness, which is all about explaining (and predicting) the functional and phenomenological properties of consciousness (Seth, 2016). I agree there is merit in speculating about the potential umwelts of conscious AI (Shanahan, McDonell, & Reynolds, 2023). But this approach has a better grip when applied to systems which we are already confident are (or can be) conscious, as opposed to determining whether a system has this capacity. The reorientation suggested by Fields & Glazebrook is therefore better suited to non-human animals than to AI [(Birch, Schnell, & Clayton, 2020; Singhal, Birch, & Seth, 2026), see also Birch, R6].

Observer-dependency is a serious challenge to computationalist views, especially to arguments for conscious AI based on such views (Lerchner, 2026). However, many efforts in the philosophy of computation have attempted to specify observer-independent criteria by which a physical system may be said to compute (Anderson & Piccinini, 2024; Cantwell Smith, 2002; Wolpert & Korbel, 2025). These criteria restore discriminatory power by enabling distinctions between systems that compute (in some specific sense) and those that do not.[5] But since these criteria are generally broader than Turing computation, they sacrifice substrate neutrality, undermining CF from a different direction.

R3.3 Broadening computation: Is life computational?

Let's first set aside maximalist pan-computationalist views according to which everything (even a bucket of water) computes, or which view computation as physically fundamental (Wolfram, 2002). Like full observer-dependency, these views also lack discriminatory power. And even if pancomputationalism is true, this in no way mandates the conclusion that the type of computation responsible for consciousness in brains can be realised in anything other than brains (see "computational BN," CAIBN 5.3).

Perhaps the most provocative commentary advocating a broader computationalism comes from Richards & Agüera y Arcas [see also (Agüera y Arcas, 2025)]. They claim that life itself is "inherently computational." This allows them to cast BN as a computational theory, in stark contrast not only to my version of BN but also (probably) to all versions. They agree with me that processes like prediction, homeostasis, and autopoiesis are linked to consciousness, but they claim – *contra* me – that these processes "can *only* be implemented by physical systems that can be characterised as computers engaged in computation" [emphasis in original]. This is admirably clear. They are not merely saying that fundamental life processes, such as autopoiesis, can be characterised computationally, nor that everything is computational. They are saying that life literally *is* physically implemented computation.

To support this claim, they observe that von Neumann's "Universal Constructor" provides a mathematical definition of computation as powerful as the Universal Turing Machine: both are Turing-complete (von Neumann & Burks, 1966). They then argue that a Universal Constructor, since it can be implemented as a cellular automaton, has "closure with its physical universe," drawing a contrast with Turing machines which, even in their abstract universe, need separate heads, tapes, and so on. Broadly, since Universal Constructors can grow and self-replicate within their cellular automata universes, they conclude that life is inherently computational.

An immediate problem is that "physical closure" for a Universal Constructor only applies within the abstract universe of cellular automata. There is no closure with respect to the real physical universe. If you run a Universal Constructor on a digital computer, the Universal Constructor does not regenerate or replicate the computer on which it is running. By contrast, (cellular) autopoiesis, for example, *is* closed with respect to the real physical universe. Living systems continually reconstitute their own physical basis. This is what autopoiesis (roughly) means (Maturana & Varela, 1980). (Richards & Agüera y Arcas recognise something like this, but they continue to claim that "autopoiesis is still computation." More on autopoiesis in R5.2; see also Schwitzgebel).

What seems to be happening here is a version of the "Church-Turing Fallacy" (Copeland, 2024; Piccinini, 2020). The Church-Turing thesis is the thesis that, if a function is naturally interpretable as computable, then it is computable by a Turing machine (equivalently, it is Turing-computable; CAIBN 3.0). This thesis may well be true. No counterexamples have been found. But by itself this thesis says nothing about what physical processes count as computations. This is the fallacy: interpreting abstract mathematical propositions such as the Church-Turing thesis (or the Universal Constructor) as having the power to determine whether a physical system is literally computational or not. An example from Richards & Agüera y Arcas: "Since a Universal Constructor is a computer, it follows that any system capable of autopoiesis must be computational." This does not follow. Here's what they might have said: *if* life is naturally interpretable as computation, then it can be computed by a Turing machine (and hence by a Universal Constructor). But this is not what they say.[6]

The aspect of life perhaps most easily interpretable as computation is gene-based self-replication, thanks to the parallels with Universal Constructors. Arguably, one might interpret self-replication as a biological form of computation. Even here, though, there are limits, thanks to the many non-genetic factors involved.[7] For other aspects of life, like metabolism and autopoiesis, there is much less to lean on.

Richards & Agüera y Arcas try to lean on something when they claim, by analogy with the multiple realisability of functions like flight (compare birds and planes), that life in general ought to be substrate independent: "[in] both biological and technological systems … the substitution of one implementation or substrate for another is commonplace." This argument, too, misses its mark. While being substrate independent increases the plausibility that a function is computational, not all substrate independent/flexible functions are computational (flight, again, is a good example). Moreover, while flight has a consensus functional description – broadly, anything that moves through the air in a controlled and sustained way – consciousness does not. Another perspective on multiple realisability is that functional processes related to cognition

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

might be realised either with consciousness or without it: the latter might account for the so-called 'mental workspace' (J-space) identified in large language models (Gurnee et al., 2026; Seth, 2026).

Finally, and relatedly, Richards & Agüera y Arcas appear to confuse multiple realisability (different implementations) with substrate independence (different materials). This is a key distinction: evidence for the former is not necessarily evidence for the latter [CAIBN 3.3, (Polger & Shapiro, 2016)]. Overgaard, too, suggests I am arguing against multiple realisability for consciousness when I am arguing against substrate independence. As Schlicht notes, a broadly functionalist view allows for (and is partly motivated by) multiple realisability, but CF, at least for standard digital computation, requires substrate independence (or, at least, silicon substrate flexibility).

Summarising, it is difficult to sustain the view that life is inherently computational, let alone a form of substrate-independent computation. By highlighting links between life and consciousness, Richards & Agüera y Arcas end up importing the very substrate constraints that arguably make their view a variety of BN rather than CF.

R3.4 Broadening computation: Other forms

Many commentators focus on forms of computation which lie somewhere in between Turing computation and the views discussed above. These include analogue, reservoir, mortal, neural, and biological, to name just a few – see Friston, Fields & Glazebrook, Ramstead, Hohwy, Aru et al., and Bowes. These broader forms reflect a tendency to redefine computation once it is recognised that some specific narrower form (e.g., Turing computation) is a poor match to biology. The problem for CF is that these broader forms of computation typically trade off against substrate flexibility. And the greater the substrate dependency, the more CF loses its motivation and its credibility.

It could be, as De Brigard notes, that some form of substrate-dependent computation, of the kind implemented in brains, is sufficient for consciousness. If this is so, conscious AI would be possible, but it would have to possess the relevant substrate properties – disembodied silicon-based digital computation would not do. De Brigard is relaxed about what counts as computing in the brain, broadly equating it with fine-grained forms of functionalism [specifically, Andy Clark's "microfunctionalism" (Clark, 1989)]. But if the relevant microfunctional properties can only be implemented in biological systems, then we arrive at a (weak) BN dressed in computational clothes, and without the multiple realisability and substrate flexibility that the term "computation" usually connotes.

Two specific broader forms of computation are worth highlighting: mortal computation (Friston, Hohwy) and biological computation (Aru et al.; see also Kirkpatrick [2022], Milinkovic & Aru [2025])]. Mortal computation (CAIBN 3.4) means that the computations cannot be separated from their implementation and so will "die" when their implementation or substrate "dies." Friston makes the case that mortal computation (with the additional constraint of in-memory processing) satisfies one element of the BN story I sketched – the mathematical through line from thermodynamics to predictive inference (see also Parr & Broulidakis, R5.0). This argument provides a computational justification for (weak) BN, though the "mortal" is doing more work than the "computation." Hohwy focuses instead on the possibility that mortal computation may be well suited to explaining functional properties of consciousness, perhaps offering a substrate-dependent alternative to Michel's research programme.

Aru et al. agree that standard digital computation is likely insufficient for realising consciousness, but they suggest that forms of computation which take into account more details about the brain might be enough. This shifts the question from whether AI as we know it can be conscious, to how brain-like/biological-like AI has to be to move the needle on the credence that it might be conscious. Aru et al. emphasise two properties: the continuous, fluidic nature of biology, and the fact that biological systems are *scale-integrated,* such that different levels of granularity mutually constrain and define each other [see Figure 1 and (Milinkovic & Aru, 2025; Milinkovic et al., 2025)].

Scale integration is particularly interesting (see also Silberstein, Feinberg & Mallat, and Piccinini). As noted in CAIBN 3.5, this property of biological systems contrasts with the sharp separation of scales (and substrate neutrality) fundamental to Turing computation. And, as both Aru et al. and Silberstein argue, it may be relevant both to brain function and to consciousness. Scale integration is related to scale-free dynamics, which is in turn related to the balanced sensitivity to perturbation needed for effective cognition (Silberstein, 2023). More speculatively, scale integration may be related to the spatially unified, temporally coherent, integrated-and-differentiated flow of conscious experiences [(Milinkovic & Aru, 2025; Storm et al., 2024); see also Nave]. Intriguingly, recent work using dynamical independence (Barnett & Seth, 2023) suggests that neural dynamics show greater scale integration (and hence lower "emergence") in conscious states than during anaesthesia-induced unconsciousness (Milinkovic et al., 2025).

Mediano et al. pick up on the relation between scale integration and emergence, distinguishing between several distinct facets of what they helpfully call "compatibilist emergence."[8] They note that while one facet – complete macroscopic causal closure (perfect dynamical independence of macro from micro) – is in tension with scale integration (or generative entrenchment), other facets of emergent dynamics are not and may in fact depend on it. For example, *causal emergence* – macroscopic variables having greater predictive power over the evolution of a system than available from microscopic descriptions – may only be possible in scale-integrated systems (Rosas et al., 2020). What's more, modelling studies using reservoir computing have shown that optimising for causal emergence can enhance performance in a wide range of simulated environments (Tolle et al., 2026). Mediano et al. also hint that the kinds of emergence entailed by scale-integrated systems may map onto basic phenomenological aspects of consciousness, including its unity (Nave) and its integration of intrinsic value with perceptual contents – an idea nicely in line with my own motivations for BN (R5.1).

Mediano et al.'s comment is particularly important because it highlights that terms like "emergence" are moving from being (unhelpful) stand-ins for unexplained magic towards being (helpful) umbrella concepts hosting a variety of well-defined and rapidly evolving quantitative measures, which in turn ramify and

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

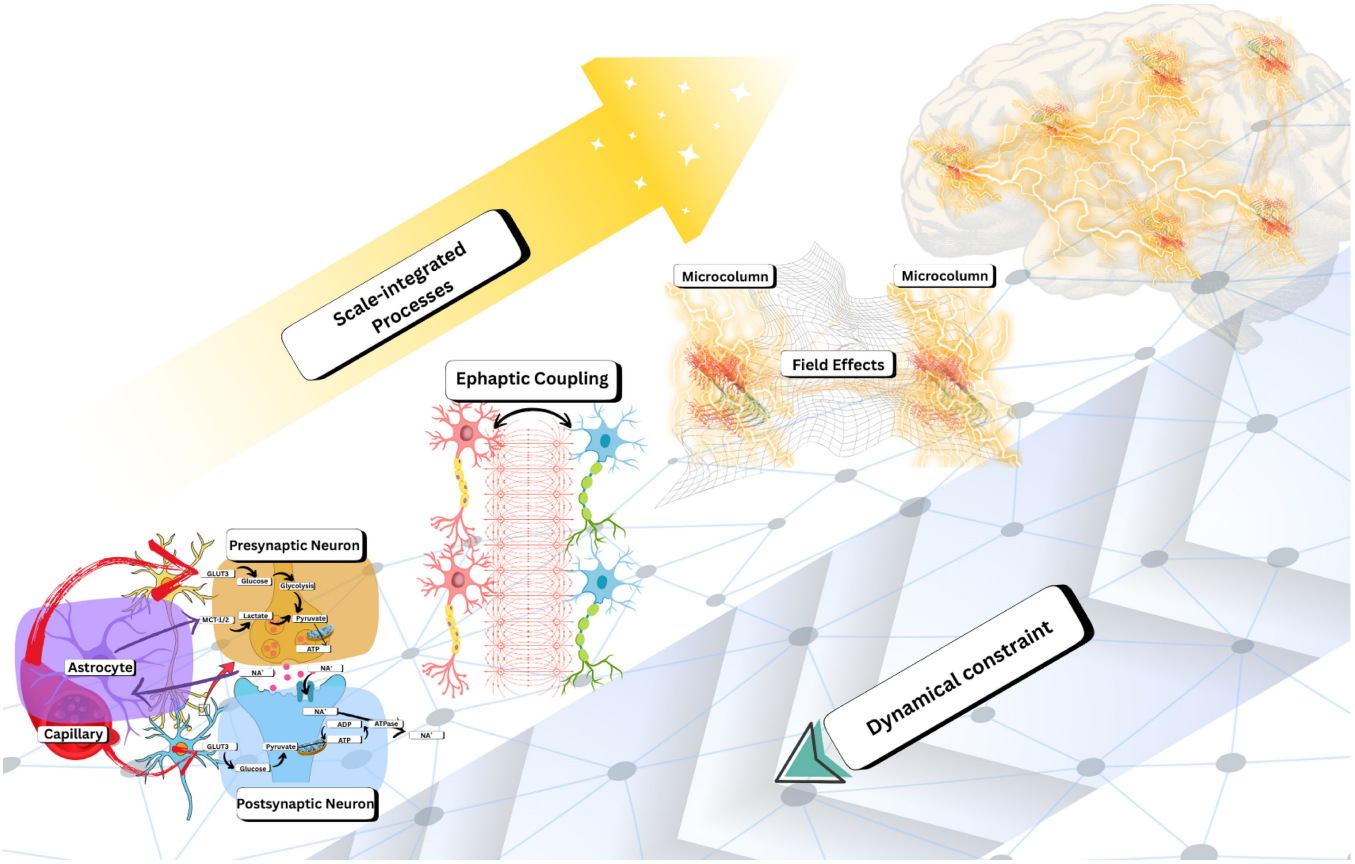


Figure 1. Scale integration in the brain. At subcellular and cellular levels, metabolic constraints shape synaptic transmission. For single cells and populations, electric fields influence neuronal behaviour. At the mesoscopic level, microcolumns and local field potentials capture collective dynamics that arise from and constrain single neuron activity. At the macroscopic level, large-scale fields couple and constrain activity across and within brain areas. Across all scales, continuous processes constrain and integrate emergent dynamics. Adapted by Borjan Milinkovic from Milinkovic & Aru (2025) and used with permission.

extend the basic idea (e.g., a system being "more than the sum of its parts") in ways that enrich our ability to understand all manner of complex systems, brains and beyond.

Summarising, there are many broader forms of computation, but their substrate dependencies undermine the sufficiency claims of CF and therefore also the prospects for conscious AI. Instead, many of the properties highlighted – whether described as "computational" or not – suggest new ways of investigating how substrate properties might relate to both cognition and consciousness (R5).

R3.5 Embodiment and embeddedness

Bowes – who also argues for broader forms of computation – emphasises the potential importance of embodiment and embeddedness, rather than substrate properties *per se*. He takes aim at the strong separation between syntax and semantics enshrined in standard Turing computation, arguing that embodiment and embeddedness may dissolve this distinction, at least partially. I am sympathetic, and I agree that embodiment/embeddedness might ground some authentic kind of understanding or aboutness in AI (CAIBN 3.7 and note 40). However, the question of *understanding* (and semantic content more generally) is separate from the question of *consciousness*. Even if AI systems (like embodied LLMs) could be truly said to understand things, it does not follow that they will also be conscious. Instead, understanding and consciousness may be separable both in principle and in practice.

In addition, by emphasising non-computational factors (embodiment and embeddedness), Bowes challenges the computational premise that substrate-independent computations are sufficient for cognition and consciousness, and in doing so, provides yet more reason to be sceptical of CF (R2.1).

Mitchell & Jennings also emphasise embodiment and embeddedness, but from the perspective of subjectivity and selfhood. They argue that any account of consciousness will be incomplete without an account of the nature and role of the conscious subject. They cite Zahavi in claiming that "when we experience something consciously, we experience being ourselves having that experience" (Zahavi, 2008). While I'm all in favour of explaining selfhood (Seth, 2021a), I doubt that consciousness necessarily involves the experience of being the subject of experience (Metzinger, 2024). In my view, experiences of selfhood are multilayered, and the version of BN in CAIBN locates the simplest forms of (biological) consciousness in a shapeless, formless, inchoate – but fundamental – "feeling of being alive." This is very distant from the metacognitive, reflective form of selfhood implied by Mitchell & Jennings.

Mitchell & Jennings also advocate focusing on *relations* – between subject and object, between parts of the world and across time – as a productive approach to explaining consciousness. I am glad they highlighted time, for the nature of time provides one more nail in the coffin for CF.

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

R3.6 Time

As Allen notes, digital computers and brains differ fundamentally in how they relate to time (Seth, 2025; Van Gelder, 1998). For discrete digital computations, only sequence matters. There could be a microsecond or a million years between two steps in an algorithm, but it would still be the same algorithm. According to CF, slowing the clock speed of a computer to a crawl should make no difference to the conscious experience entailed by the computation. This is at odds both with the nature of brain function and with the phenomenology of consciousness (Buonomano, 2026).

Neurobiological activity is anchored in continuous time in ways that algorithms, by design, are not. For brains, as for biological systems in general, time is physical, continuous, and inescapable. Brain function is not only embodied and embedded; it is also *entimed*. The scale-integrated processes that underlie cognition and consciousness are as rooted in physical time as they are in physical substrate. More generally, living systems must resist in continuous time the decay and disorder that lies along the trajectory to entropic sameness mandated by thermodynamics (Friston, CAIBN 4.0). These points reinforce the implausibility of abstracting brain function into the timeless sequence space of digital computation.[9]

What's more, conscious experience itself is richly dynamic and inherently temporal (Singhal & Srinivasan, 2021). It flows. It does not stutter from one state to another, with arbitrarily long gaps in between. In the stream of consciousness, the passage of time is continuous, irresistible, and complex, encompassing not only succession but also flow, duration, and persistence. As Mitchell & Jennings put it, "we live in and through time." If consciousness is inextricable from physical time, as it seems to be, then it cannot be a matter of algorithm alone, and CF is once again in deep trouble.

One could attempt to rescue temporality by allowing algorithms to check a separate clock that tracks physical time. But doing so means the algorithm itself is no longer sufficient, which undermines CF. Similarly, if computation is broadened to include continuous, analogue processes which do track physical time, then this also undermines CF because doing so re-imports substrate constraints.

R3.7 Simulation and instantiation

To complete the discussion of computation, I turn to the distinction between simulation and instantiation, introduced in CAIBN 3.7 and discussed by many commentators (e.g., Ramstead, Blum & Blum, Chiang, Wiese, Dolega & Cleeremans, and Levin). To rehearse the key point: a computational simulation of a phenomenon (e.g., consciousness) will not instantiate that phenomenon unless the target phenomenon is itself a computation of a sufficiently similar kind.

Wiese echoes the main reason for this: digital simulations will not generally replicate the relevant causal relations that are present in the phenomena being simulated and that are required for realising that phenomenon [see also (Brinz, 2026)]. He illustrates the point with Conway's Game of Life (Gardner, 1970). The causal relations in the game that are responsible for all its interesting phenomena (gliders and the like) are between adjacent cells in the cellular automaton. But these causal relations are not replicated in its implementation basis (R3.3).

Dolega & Cleeremans disagree. They argue that simulations might be able to replicate relevant causal relations, and so might be able to instantiate their target phenomena. They say that simulated hurricanes causally affect simulated ships in SimCity. But this argument, as Wiese points out, confuses apparent causality among simulated entities with physical causality in the implementation basis (CAIBN note 22). Matej Kohár puts the point clearly, arguing that virtual causation is illusory because the (presumed) structure of causal interactions between virtual objects does not map onto the structure of causal interactions between underlying computational objects and processes [(Kohár, 2024), see also (Beisbart, 2024)].

Turning to consciousness, Dolega & Cleeremans argue that "simulated consciousness" in (for example) an LLM can have causal effects on human users – effects perhaps as powerful as those evoked by conscious human beings. But this confuses the causal effects *of* a phenomenon (on something else) with the causal relations *responsible for instantiating* a phenomenon. Even simulations of hurricanes can have causal effects on other things in the real world. If a simulated hurricane in a weather forecast predicts that a real hurricane is on the way, people run for shelter.

What Dolega & Cleeremans appear to overlook is that the design principles of digital computation ensure that causal relations in hardware are explicitly decoupled from causal relations in software, whether these involve simulated boats, Game of Life gliders, or neural network weight matrices.

Chiang treads the same path when he claims that "all the relevant criteria for living systems could be satisfied by software running Turing computation." But if the relevant criteria include specific physical causal relations, as they no doubt do (consider again autopoiesis), then this is false. Simulating (for example) autopoiesis does not realise the same causal relations needed to instantiate autopoeisis. Chiang imagines a limiting case in which individual atoms, perhaps even quantum effects, are incorporated in the simulation, but this doesn't help. The granularity of a simulation is orthogonal to whether the phenomena being simulated are inherently computational or not. A more detailed simulation of a hurricane may be more useful, but is no closer to being actually windy. Nor does simulating a cell in more detail make the simulation any closer to being alive.

Things get more complicated when it comes to active inference and the free-energy principle (FEP), concepts which play important roles in my version of BN. As Wiese and Friston (see also Parr & Broulidakis) note, the FEP provides two equivalent descriptions of (some) self-organising, self-maintaining systems, one in terms of physical processes (thermodynamics) and the other in terms of information theory – with the latter underpinning the variational algorithms used in machine learning (Hoffman et al., 2013). Given the variational view, wouldn't a simulation of active inference count as instantiation? This question is posed by, among others, Allen.

As noted in CAIBN 4.2, the complication arises because the principles of learning and inference in the FEP can be abstracted (at least in part) from their biological substrate and implemented *in silico* (Parr, Pezzulo, & Friston, 2022). This is where the dual description is critical. The claim is that in living systems, the variational and thermodynamic processes *are the same physical process* at two levels of description (Friston, Parr & Broulidakis). By contrast, in the Turing world, they are *not* the same process. The variational implementation is an algorithmic abstraction of a dynamic, continuous-time, substrate-entangled, non-algorithmic biological process of free-energy minimisation. (The importance of time surfaces again here: in living systems time is necessarily continuous since there is a single physical process, while in algorithmic implementations this continuity is broken.)

Wiese amplifies this point, arguing that the FEP supports the claim that the (substrate-specific) causal relations in living systems are relevant to consciousness, as distinct from algorithmic implementations of variational free-energy minimisation in

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

machine learning. (Note, though, that he does *not* think that only biological substrates can support consciousness, which is helpful since it underlines that his argument does not originate from any affinity towards BN.[10])

More on the FEP later (R5.1). Zooming back out, Ramstead makes the strong claim that the ability to computationally simulate X provides no evidence that X is literally computational. While I agree with his message – that one should not overinterpret the ability to simulate things – there are important nuances (CAIBN 3.7). If the best simulation of X is one in which *X is simulated as performing computations* (according to some definition of computation), then this should increase the credence that X is literally computational. Equally, if the best simulation of X involves the simulation of non-computational processes, then this should decrease the credence that X is literally computational. Compare, for example, Ramstead's example of simulating the orbit of Jupiter (a non-computational phenomenon) versus simulating the Bubble Sort algorithm, which is a canonically computational phenomenon [though, as Levin points out, even computational simulations of computational phenomena may sometimes be incomplete in that they fail to realise all the corresponding computational properties; see R5.2 and (Zhang, Goldstein, & Levin, 2024)].

Here's one important implication of this discussion. The conceit that fine-grained whole-brain simulation provides a path to conscious AI (or immortality) is wholly misguided. The more microscale neurobiological details matter for brain function, the *less* likely it is that simulating the brain in microscale detail will realise all its properties – including consciousness – for the simple reason that, if such simulations are needed, it is precisely because they suggest that things other than digital computation likely matter.

R4. Consciousness – the bigger picture

Before turning to BN, I want to address several commentaries that consider consciousness in a larger frame. This is because what counts as a reasonable account of consciousness, and of its potential substrate dependency, will depend on one's metaphysics and one's worldview. Let's therefore briefly set aside the pragmatic materialism I favour and explore some wider terrain. We'll see that the alternatives tend to purchase flexibility and liberality at the cost of discriminative power.

The most radical reconfiguration is proposed by Gómez-Marin, who invites us not only to dispense with materialism but also to suspend the broader worldview of naturalism and to consider instead its counterpart: animism. Instead of a continuity of physical properties among entities (everything is made of physical stuff) and a discontinuity of interiorities (some things are conscious), we have the mirror image: a continuity of interiorities, without continuity of the physical. A mountain can have a soul even though it is made of stone. Perhaps an LLM can have a soul, too. For Gómez-Marin, anthropology is the missing ingredient in facing the challenge of interacting with machine minds.

This is heady stuff. Certainly, anthropological research can help us better understand human engagement with AI, adding diversity and nuance. A starting point here is to explore cross-cultural differences in consciousness attributions [e.g., following (Colombatto & Fleming, 2024)]. Gómez-Marin is also right that AI offers new opportunities to re-examine what we mean by "mind" and by "consciousness" – an optimistic take on the "AI as a mirror" metaphor [(Vallor, 2024a), see also Shevlin, Shanahan]. But while there's merit in questioning the foundations of knowledge, suspending naturalism means we can no longer make any inferences about the fact of the matter. This is why I was careful in CAIBN (and in R1) to set out ground rules: consciousness exists; some things have it; other things don't.

Dissatisfaction with materialism is often a wellspring of motivation for alternative metaphysics. This may be especially true for panpsychism – the idea that consciousness is fundamental and ubiquitous. For example, McGilchrist, who favours panpsychism, casts materialism as "the view that consciousness is an unexplained brute emergence from *lumpen* matter" (his emphasis). This is rather unkind on materialism, on emergence, and on matter. The stuff of biology is anything but lumpen; emergence need not be brute [(Barnett & Seth, 2023; Rosas et al., 2020), R3.4], and formulating explanations is the task at hand, with progress being made (Cleeremans, Mudrik, & Seth, 2025).

For McGilchrist, panpsychism opens the possibility that AI systems (and many other things) might have "experience of a kind" in virtue of being part of a larger conscious entity (perhaps the cosmos).[11] I, on the other hand, am on record for being impatient with panpsychism, on the grounds that it explains nothing and generates no testable predictions (Seth, 2021b). Despite these differences, he and I agree that AI is likely incapable of the kinds of understanding and intelligence that arise for human beings in virtue of experience, and we share concerns about the dangers of being seduced by the confabulations of LLMs when they claim to be conscious (CAIBN 6.2, R6.3, Shevlin, Schneider, and Roelofs). His intuition that LLMs embody a left-hemisphere emphasis on "internal consistency" rather than "truth to experience" rings true to me. Unfortunately, his concerns about confabulations are somewhat neutralised by a panpsychist perspective which allows that AIs might still be conscious (in a way), whatever they do or do not say.

Roelofs, another friend of panpsychism, does an excellent job of exploring the implications of this position for conscious AI but, in doing so, provides more reasons to steer clear. For Roelofs, even if substrate dependence holds for human and animal consciousness, other kinds of systems (including AI) could have different kinds of consciousness, based on different substrates. This again changes the relevant question from "Is this system conscious or not?" to "What sort of consciousness is present in this system?" (see also Fields & Glazebrook, McGilchrist, Shevlin). His conclusion is the opposite of McGilchrist (and me): in his universe of abundant consciousness, we should default to accepting self-ascriptions of consciousness made by AI.

But – as Roelofs recognises – systems trained on lots of writing about consciousness are precisely those for which self-ascriptions can reasonably be doubted (see Schneider). It seems that for Roelofs (though not McGilchrist) this reasoning is outweighed by a metaphysics of abundance. Here, I can only lay my own metaphysical cards on the table. In the absence of a consensus theory of consciousness (reflecting the "humble agnosticism" Roelofs rightly advocates), it seems preferable to adopt a perspective – such as pragmatic materialism – which encourages us to shift our credences about the presence of consciousness on the basis of evidence, rather than to adopt a metaphysics which encourages credulity for the confabulations of LLMs and which makes it almost impossible to deny the presence of consciousness in anything. I'll return to the issue of testing for consciousness in R6.

At the other end of the metaphysical spectrum, Clark, a comrade and an inspiration in matters of predictive processing, advises me to set aside my attachment to phenomenal realism and join the (badly named) illusionists. For Clark, consciousness is just a label for "a large and complex set of reactive dispositions." He

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

draws me in by noting that this view is compatible with both CF being wrong and (weak) BN being true.

And yet. Illusionism is indeed badly named. When gulped down, it encourages the view that conscious experiences do not exist (which, as Galen Strawson memorably objected, is "the silliest claim ever made"[12]). If sipped from, as Clark advises, then we merely exclude experiential features that are independent of "some profile of reactions or responses." This "weak illusionism" holds some appeal and seems perfectly compatible with phenomenal realism. This would be the case if experiential features are entailed by (or identical to) physical states which have functional properties that are related to these experiential features, with these functional properties being expressible in reaction/response profiles. In this view, conscious experiences exist, but they might not be what people think they are (Dennett, 2001; Frankish, 2017; Seth, 2021a). Should I therefore join the "weak" illusionists? I still resist, partly because the weak and strong varieties are too easily and too often conflated, and partly because the overall "illusionist" framing deflects attention from the real-problem project of explaining the experiential/phenomenal properties themselves (Seth, 2016).

Returning to conscious AI, Clark suggests that thinking of consciousness in terms of reactive dispositions changes the landscape, since consciousness and intelligence/cognition are now less distinct than they would otherwise be, and intelligence is already demonstrated (to some extent) *in silico*. But reactive dispositions cover a much broader range of functions than is typically meant by "intelligence" [though see Fleming & Shea, R5.4]. For example, if the "reactive dispositions" that matter are those at the level of metabolism and autopoiesis, then nothing much changes from the landscape as described in CAIBN. Low-level reactive dispositions simply reimport substrate constraints, making weak illusionism a badly named version of BN.

Shanahan diagnoses a different metaphysical ailment altogether, believing me to be in the grip of "dualistic intuitions" thanks to my ground-rule assertion of phenomenal realism (see also Shevlin). The cure, he thinks, is provided by a Wittgensteinian reorientation from whether X is or isn't conscious to how we (should) use the word "consciousness" in society-wide conversation. Social and linguistic influences are no doubt important – recall the common conflation of "intelligence" with "consciousness" and the (mis)use of "hallucination" to describe LLM confabulation – but language games have their limits. Shanahan's appeal to Wittgenstein's beetle-in-a-box analogy tells us why. The content of the box, whether a beetle or a conscious experience, would only "cancel out" if it has no observable functional effects whatsoever. Yes, consciousness as a purely private dualistic epiphenomenon cancels out. But consciousness as a functionally potent fully real phenomenon entailed by certain physical states does not. Accusations of residual Cartesianism might instead be better aimed at diehard computationalists, for whom a form of dualism remains enshrined in the sharp distinction between replaceable hardware and immortal algorithms – a distinction rejected by a more biological perspective.

Bowes, too, detects a Cartesian ghost in the beast machine. He worries that my description of consciousness as *being* or *feeling*, rather than *doing*, implies a Cartesian intrinsicality at odds with any kind of functionalism, whether computational (Turing-style or – as Bowes prefers – otherwise) or non-computational. I disagree. In CAIBN 5.4, I emphasise that functionalism in general is perfectly compatible with BN and that apparently intrinsic properties may well be understandable in terms of finer-grained functional relationships (see also R2).[13] There is no tension with the real-problem project here. Only opportunity. As we build explanatory bridges from (potentially functional) properties of embodied, embedded brains to properties of consciousness, we're doing alright.

R5. Prospects for biological naturalism

This brings me to the many commentaries addressing aspects of BN. It's first worth recalling that (*contra* Michel) BN doesn't need to be true for CF to be false (R1.1). And it may be (as I acknowledge in CAIBN) that my version of BN is wrong in the details or wrong altogether. But BN as a position and research agenda is not *obviously* wrong. On the contrary – as many commentaries indicate (Block, Cao, Feinberg & Mallat, Friston, Jablonka et al., Lane, Nave, Parr & Broulidakis, Piccinini, and others) – it has plenty going for it.

R5.1 Predictive processing – a wrong turn?

The version of BN I sketched in CAIBN (but by no means all versions) leans heavily on a continuity between predictive processing (PP) as a theory of perceptual content and the free-energy principle (FEP) as an account of living systems. Several commentators worry about this continuity, warning that the PP/FEP package might be too generic to support claims specific to living systems. My response is that PP/FEP as a set of abstract principles is indeed too generic, but when grounded in specific biological properties, it is more than up to the job.

Baltieri & Kanai worry that nothing about PP, the FEP (or active inference) seems exclusive to biological systems (see also Birch). They are correct that, *taken individually*, each of these ideas can be (and has been) applied more widely. But my case is that, *taken together*, they provide a compelling throughline from properties like autopoiesis and metabolism to functional and phenomenological properties of consciousness (see also Friston, Parr & Broulidakis). This position in fact seems well aligned with one of Baltieri & Kanai's suggested "alternatives" – that of providing formal mathematical criteria for what counts as biological (see also Kleiner).

Nave, who has written an excellent book on the FEP (Nave, 2025), names the same concern: "[the FEP] is *not* a unique principle of life but a general description of every 'thing'." She then helpfully anticipates my response (the throughline) and indeed lends a hand by suggesting that experiential unity – arguably a universal aspect of consciousness (Bayne, 2010) – may be explicable by precisely the sort of physical unity entailed by metabolism and autopoiesis.

Hohwy, while comfortable with the overall framing in terms of the FEP, is similarly concerned that the FEP itself is too general, but also that a more specific grounding in metabolism and autopoiesis may be too speculative. He advocates "mortal biological computation" instead [R3.4, (Ororbia & Friston, 2023)]. For Hohwy, mortal biological computation may account for some functions associated with consciousness, as seen through the lens of "self-evidencing" (Hohwy, 2026). I am open to this possibility. It complements the "real problem" focus on experiential properties, while prompting questions about the functional competences which may be enabled by scale-integrated systems (scale integration goes hand-in-hand with mortal computation). I am less convinced by Hohwy's conclusion that "the substrate of mortal computation comes out as precision control of the causal dependencies of nodes within a Markov blanket, working as a generative model." This seems open to highly abstract interpretations – perhaps useful for elucidating

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

consciousness-specific functions, but less useful as a basis for establishing sufficient conditions for conscious AI.

Godfrey-Smith, while sympathetic to my critique of CF [unsurprisingly, since he inspired a good chunk of it (Godfrey-Smith, 2020)], finds a different reason to worry about PP/FEP. He (along with Birch) channels the familiar "dark room argument" – the idea that acting solely to minimise sensory surprise, for example, by hiding away in a dark room, is neither what organisms do nor what they should do if they are to survive and thrive (Sun & Firestone, 2020). However, when applied to living creatures, the FEP mandates minimisation of prediction error *in the long run*, rather than only in the short term (Parr & Friston, 2019; Seth et al., 2020). The spider hiding in its dark-room den might be homeostatically fine for a while, but in the long run, hiding away will always be surprising for the spider, given the motivational and emotional states endowed by evolution that drive it to reproduce.

Importantly, the long-run view does not mean that "anything goes" – that any behaviour or function can be rationalised away to some unspecified future (Baltieri & Kanai, Fields & Glazebrook). On the contrary, there is exciting potential to develop PP/FEP (or active inference) models which involve minimisation of sensory prediction error over extended but specific time scales [e.g., (Da Costa et al., 2020; Da Costa et al., 2026; Friston et al., 2017; Tschantz, Seth, & Buckley, 2020)]. These models can make predictions about (for example) how to balance epistemic actions (useful for escaping from dark rooms) with instrumental (goal-oriented) actions – including situations when goals are homeostatic/allostatic priors or set points (Parr et al., 2022; Seth, 2015).

Parr & Broulidakis propose something like this view, noting that prediction error minimisation in biological systems can be linked to epistemic foraging (e.g., selecting which sensory signals to sample) as well as to planning. They also emphasise (with Friston) the deep connections between thermodynamic and informational free energies, arguing that biophysics and behaviour provide complementary constraints on the dynamics of perceptual inference in biological systems. Two sides, if you like, of the same PP/FEP coin.

Altogether, while the FEP/PP is only one way of substantiating BN, it remains particularly promising. Worries about genericity are defused when its abstract principles are rooted in biological processes like metabolism and autopoiesis, and concerns about predictive power are alleviated by recognising that it isn't always minimising short-term sensory prediction error that matters but how living systems generate and maintain models of their bodies and worlds under existential constraints.

R5.2 How biological is biology?

Several commentators – prominently Kleiner, Legg, Schwitzgebel, and Solms – wonder whether some of the biological processes I highlight as being potentially necessary for consciousness, such as autopoiesis and metabolism, are restricted to living systems. Of these, autopoiesis is the most interesting to consider.

Autopoiesis was originally defined in the context of living systems, most notably biological cells (Maturana & Varela, 1980). A key criterion is *self-production*: a network of processes generates the components that constitute the network itself, thereby continually bringing forth a boundary between the autopoietic system and its surroundings. One can generalise the concept to self-sustaining organisational patterns beyond living systems (Luhmann, 1990). But these generalisations widen the original meaning of the term (Maturana & Varela, 1980) and so cannot license conclusions about biological autopoiesis being realisable elsewhere.

Schwitzgebel proposes that "nothing appears to prevent autopoietic processes in a standard computational system." A robot autonomously ordering new replacement parts might, he says, be minimally autopoietic (Legg suggests something similar). However, this scenario fails to satisfy the self-production criterion because the self-production is externalised (through the ordering of parts) such that the network of processes responsible for maintaining the robot extends beyond the robot. Even though his robot has a plastic shell, constituting a boundary-of-sorts, maintaining this boundary requires ordering external parts. The boundary is not endogenously created.

Schwitzgebel points out that living systems depend on external inputs, such as food. He argues that this dependence is similar to the robot depending on external parts. But the similarity is overstated. All autopoietic systems, being thermodynamically open, must exchange matter and energy with their environment (they have a "turbulent materiality," as Jablonka et al. poetically put it). The difference, and the aspect of autopoiesis that is missing from Schwitzgebel's scenario, is that the robot depends on the environment for ready-made components, whereas truly autopoietic systems transform generic substrates into these components (via metabolism). Living systems are organisationally closed in ways that Schwitzgebel's robot isn't.[14]

There is room for nuance. As Schwitzgebel notes, even Maturana and Varela allow for some substrate flexibility. This is useful, because it guards against carbon chauvinism. But Schwitzgebel's scenario generalises too far. It fails to realise the "relations which these [components] must generate to constitute the machine or system as a unity" [(Maturana & Varela, 1980), p.77]. On the other hand, if a robot could indeed manufacture its own components from generic substrates under energetic constraints, then it might be truly said to be autopoietic, but it would also be considerably closer to a living system than any existing computer or robot. Genuine robotic self-production of this kind, if indeed necessary for consciousness, would vindicate BN – not refute it.[15]

The challenge from Solms is different. He describes a range of computational models which exhibit autopoietic-like properties such as homeostasis and active self-other boundary maintenance. But these are best understood as models or simulations of autopoiesis, not realisations or instantiations (R3.7). The causal relations required for genuine autopoiesis are not present in these models because the implementation basis – the computer running the model – is not itself within the self-producing network of processes.[16]

Changing focus, Kleiner helpfully warns against a "BN of the gaps" in which the aspects of life proposed as necessary are simply those which currently lie outside the scope of artificial systems (see also Birch).[17] Indeed, a primary challenge for BN is to identify which properties of living systems matter for consciousness and why (points echoed by Birch, Fleming & Shea, Bayne, Dung, and Rodriguez & Farahany) and to determine the substrate flexibility of these properties. For example, autopoiesis is very likely less substrate flexible than allostasis, which is less substrate flexible than Turing computation. This observation highlights that BN is not merely an option on a menu of consciousness ontologies but a vibrant research programme. More on this below.

Kleiner's formalisations of the BN project are also useful. The generalisation of CF to his "F*functionalism" seems compatible with my definition of weak BN (a subset of weak substrate dependence, CAIBN 5.4), according to which consciousness is a

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

property of functional organisation, but the relevant functional organisation can only be implemented by biological systems. Strong BN, according to Kleiner, applies either when the functional organisation relevant for consciousness "cannot be built" (which is a moving target; see also R6.2) or – more profoundly – when consciousness is not a matter of functional organisation alone.

Levin takes things further, arguing that "our formal models never tell the full story of what something is or what it can and cannot do." He notes this is true even for canonically computational processes such as the list-sorting Bubble Sort algorithm, which, when implemented in a local way (each item in the list deciding what to do), exhibits all manner of unexpected and unprogrammed behaviours [(Zhang et al., 2024), R3.7]. The lesson here is to not underestimate what is emergently possible within the broad category of "machines" and that even the dead sand of silicon may not be so (functionally) dead after all. At the same time, as reviewed in CAIBN 4.5, Levin's pioneering work in developmental and synthetic biology reveals deep reservoirs of functional and agential potential in biological material (Levin, 2023) that may be distinctive to such material (thus lying beyond the reach of silicon; see also Jablonka et al., Piccinini, Block). These examples underline the importance of better understanding what kinds of things different substrates – conventional and unconventional, evolved and human-made, artificial and synthetic – are capable of.

Levin's most provocative idea is that our biological and engineered offspring act as interfaces to a semi-Platonic realm from which mathematical forms and even new kinds of minds ingress into our physical worlds (CAIBN 3.8). This is itself mind-expanding stuff and – whether Platonic or not – emphasises the importance of questioning assumptions about what counts as "mind" or "intelligence" [(Shanahan, 2016); see also Gómez-Marin, Shevlin]. (Relatedly, Levin also notes that three-dimensional physical embodiment is just one possible manifestation of embodiment; CAIBN 5.7, R3.5).

Altogether, Solms, Schwitzgebel, and Legg usefully identify what is distinctive about biological autopoiesis – self-production from generic substrates – since this is exactly what is missing from their respective scenarios. Kleiner (and others) highlight how BN names an exciting research agenda in the way that CF seems not to be (more on this below), and Levin offers one specific way in which a BN research agenda might flourish.

R5.3 Fleshing out BN

What distinguishes conscious from unconscious living systems and states? In CAIBN, I proposed a general response: consciousness may be biology's way of implementing multimodal survival-relevant integration across diverse inputs for unified control. Several commentators offered more specific ideas, each of which fleshes out a BN research agenda in a different way (and in doing so further undermines Michel's claim that there is no alternative to CF). These cluster into three related sets of ideas.

A first set highlights notions of evaluative integration (Cao, Metzinger). For Cao et al., a "valuing subject" optimises things for its own good – it embodies a form of "internalised existential optimisation." They suggest that, for this process to be associated with consciousness, the evaluative states employed must be integrated across diverse survival-relevant signal types. This fits neatly with my own suggestions [CAIBN 4.0, (Seth, 2021a)], but Cao et al. go further. For a specific mental state to be conscious, it must, they say, be accessible to "downstream systems" involved in action, learning, memory, and so on. This suggestion is striking in two ways. First, it denies any meaningful distinction between phenomenal consciousness and access consciousness (Block, 1995; Mudrik et al., 2026). Second, it shows how BN can be combined with other theories of consciousness, such as the global workspace theory, to form hybrid theories. This potential for hybridisation is a strength of BN, allowing it to benefit from a rich theoretical milieu (Seth & Bayne, 2022). CF, by contrast, is difficult to hybridise with theories imposing substrate constraints since doing so would undermine its core claim of computational sufficiency.

Metzinger offers a different constraint on evaluative integration. He suggests that consciousness involves both biological substrate dependence and a self-modelling epistemic space (SES). An epistemic space is a set of possible states or processes of knowing, and an SES involves a system knowing about its capacity to know. For Metzinger, the essence of phenomenality is "being the content of a currently self-modelling epistemic space" (Metzinger, 2020). One might worry this risks importing an overly (meta) cognitive notion of "knowing" into an account otherwise grounded in more basic properties of living systems. However, Metzinger's emphasis on physiological arousal and the reticular activating system suggests a different, more biological, take on what the relevant kind of "knowing" might be.

A second set of ideas focuses on candidate biological properties. Here, Lane advocates a strong form of BN. He takes a deep dive into biochemistry, proposing that fluctuations in cellular membrane potentials provide an integrated real-time readout of metabolic state in relation to the surrounding environment (this "readout" is evidently related to evaluative integration too); see CAIBN 4.3. He associates this real-time readout with the most basic instantiation of a "feeling state." My suspicion is that, of all the aspects of neurobiology that Turing World encourages us to ignore, the biochemical level may be the most significant. Ideas like those of Lane, coupled with experimental work investigating (for example) how anaesthetics disrupt electron transport chains in mitochondria (Rodriguez, Peng, & Lane, 2025), may be our current best bet for explaining why life (rather than information processing) might breathe the fire into the equations of experience.[18]

Lane's ideas bring together questions about conscious content with questions about global states of consciousness, such as the difference between wakeful awareness and sleep or anaesthesia. Piccinini – while not wedded to BN[19] – focuses on these global states, suggesting that they are supported by complex mixtures of neurological, molecular, and biophysical factors. In his view (and mine), these complex mixtures likely lack the substrate flexibility to be realised *in silico*. He points to the potential importance of globally orchestrated patterns of inhibition, excitation, oscillation, and synchronisation, which by their nature are non-algorithmic. Similarly, Block emphasises the role of electrochemical mechanisms in subserving large-scale brain rhythms which may, he suggests, be required for consciousness (Block, 2026; Tauber et al., 2024). Both Block and Piccinini emphasise mesoscopic and macroscopic non-algorithmic features of brain activity, underlining that biological substrate dependency is not restricted to (microscopic) molecular phenomena. Determining the role of these features in consciousness stands as an important open question in a BN research agenda.[20]

Finally, there is an evolutionary perspective, building on the intuition that evolved biological systems will have very different properties from human-designed systems such as computers. Feinberg & Mallat link two readings of emergence in this way,

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

arguing that evolution will lead to the emergence over time (diachronic) of scale-integrated mechanisms with hierarchically (and synchronically) emergent properties that are relevant to, and may be necessary for, consciousness (CAIBN 3.5). Similarly, Jablonka et al. emphasise that evolved biological systems develop a range of substrate-specific "affordances" [by loose analogy with Gibson's ecological psychology (Gibson, 1979)[21]] that may be necessary if not sufficient for consciousness (see also Levin). They focus on "unlimited associative learning" – a form of open-ended domain-general learning – as one candidate for sufficiency (Ginsburg & Jablonka, 2019).

These ways of fleshing out BN collectively underline that taking BN seriously entails neither carbon chauvinism (only carbon can support consciousness, see Piccinini) nor neo-vitalism (a "BN of the gaps"). Rather, the idea is to build explanatory bridges between substrate properties and functional and phenomenological properties of consciousness, real-problem style (Seth, 2016). As Nave observes, it remains an open question whether alternative physical substrates or chemistries can implement the necessary and sufficient processes for consciousness, even if this turns out to require things like metabolism and autopoiesis. And if processes like these are indeed necessary or sufficient, it won't be because of neo-vitalist magic but because of the explanatory and predictive power they bring.

R5.4 Towards a BN research programme

What would a thriving BN research programme look like in more general terms? The philosophy of Imre Lakatos provides a useful guide (Lakatos, 1978). From this perspective, the relevant criterion is whether BN leads to a *productive* (virtuous circle of theory and experiment leading to accumulating insights) or a *degenerate* (fewer and fewer insights) research programme.

Several commentators sketch how a productive BN research programme might play out. Fleming & Shea advocate a *natural kinds* approach. Such an approach treats consciousness as a property which "carves nature at its joints" and asks which properties (whether of biological substrate or of cognition) reliably "travel along" with consciousness, both empirically and in terms of explanatory power. This is an excellent strategy. It is one that I and others (including Fleming) have proposed in the context of developing operational *tests* for consciousness across many candidate populations, including AI (Bayne et al., 2024); more on the topic of tests below (R6.4).

While we agree on the merit of the approach, Fleming & Shea emphasise the empirical associations (in humans) between consciousness and intelligence, whereas I have focused on their separability (CAIBN 3.8). Fleming & Shea's strategy therefore risks flattening consciousness into advanced cognition. Those aspects of intelligence and cognition that they identify as being empirically entangled with consciousness in humans (planning, metacognition, etc.; see Fleming & Michel, 2026) may be necessary only (if at all) for some human-contingent aspects of consciousness (e.g., temporally extended selfhood), rather than for consciousness in general. Furthermore, some of these functional properties are likely realisable in other substrates (e.g., in AI) without consciousness being involved. My view, by contrast, is that consciousness will likely travel along with substrate-specific properties of living systems (see also my response to Clark, R4). But Fleming & Shea are right that questions like these remain theoretically and empirically open and that some aspects of (human-like) cognition may also be substrate dependent (Seth, 2025). I strongly agree with them on the need to develop models of consciousness with the same energy that has been devoted to modelling cognition and intelligence.

Birch lays out a research agenda for BN that attempts to identify which (if any) biological properties do indeed travel along with consciousness. This fits well with a Lakatosian perspective. By testing hypotheses about specific biological properties, specific versions of BN can be falsified while BN itself evolves as a productive research programme. Birch's main worry is that candidate biological properties need to be "unmediated by effects on computation." He gives the example that disrupting layer V pyramidal neurons will disrupt consciousness, but "if the disruption is mediated by disruption to computations, computational functionalists need not worry." This highlights the need to be clear about which biological properties are inherently computational and which aren't. I do not worry as much as Birch, since there are good reasons why many biological processes are not (Turing) computational (CAIBN 3.0, R2). Furthermore, we can deploy computational models (and other kinds of model) to infer whether a candidate biological process is inherently computational or not. The more our best understanding, or model, of a biological process involves substrate dependence, the less likely it is to be (Turing) computational [(Chirimuuta, 2024), R3.7].

Rodriguez & Farahany, riffing on Butlin et al. (2023), offer a formalism for inferring consciousness which explicitly parameterises credences in positions such as CF and BN. They correctly identify the need to reduce inter-individual variance in these credences, noting that the wide current variance (e.g., in the probability that consciousness is substrate independent) is one reason why debates over conscious AI seem so polarised. Addressing this need, they propose a series of experiments that are methodologically challenging but potentially highly informative. For example, one experiment involves comparing two groups of cerebral organoids: one which self-generates energy through metabolism, while the other relies on external perfusion. While challenging, this kind of creative thinking is exactly what's needed to showcase the productive potential of BN.

A BN research programme in the broadest sense should investigate which properties of biological systems provide explanatory and predictive power in accounting for properties of consciousness as well as properties of cognition and intelligence more generally (Kleiner, Schlicht, and Metzinger). The table is already set, with a growing menu of experimental options to choose from. These include deeply substrate-entangled properties such as metabolism (e.g., Lane, Rodriguez & Farahany); global patterns of dynamics and causal structure (e.g., Piccinini, Block, Nave); and properties such as embodiment and embeddedness (e.g., Bowes, Mitchell & Jennings, Silberstein) – perhaps in worlds replete with other minds. As part of this, it will be important to work with definitions of computation that are, in principle, applicable to biological systems, so that researchers can infer whether a candidate biological property is inherently computational, and, if so, whether it's the kind of computation that can be implemented in alternative substrates, such as silicon.

As Schlicht and Dung remind us (from different sides), we're not yet in a position to decide whether neurobiological details matter for consciousness. But we should not throw our hands in the air and succumb to the seductions of motivated reasoning, psychological biases, overextended analogies, reified metaphors, or digital rapture hubris. Nor, as Allen counsels, should we enforce hegemony, whether for CF, BN, or anything else. Being uncertain doesn't mean knowing nothing, or that all options are equally

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

likely. Like the good Bayesians we should all aspire to be, we should recognise our priors and update our credences based on evidence. It's worth remembering once more that BN is a claim about necessity, whereas CF is a claim about sufficiency. This asymmetry in epistemic burden is echoed in Lakatosian productivity. Where BN offers testable claims about substrate dependency, CF often retreats into redefining the relevant computations, undermining its own core sufficiency claims in the process. For sure, things remain uncertain, but the wind is (or ought to be) with BN.

R6. Biases and ethics

Consciousness matters. Even though things remain uncertain, much turns on even preliminary answers to questions about necessary and sufficient conditions. Against this background, academic disagreements risk becoming ideological battle lines (Gomez-Marin & Seth, 2025). When the CEOs of major tech companies are weighing in on a weekly basis about whether their latest language model is conscious, the need for nuance is greater than ever.[22]

R6.1 Psychological biases

The first part of my case against conscious AI in CAIBN rested on examining our own psychological biases: anthropocentrism, anthropomorphism, and human exceptionalism. As Schneider and McGilchrist note, language models exploit these biases thanks to their emotional fluency, use of personal pronouns like "I," and self-ascriptions of mindedness. They reflect back to us an image of our collective digitised past; we become seduced like Narcissus by these reflections, and so we see ourselves in our algorithms (Vallor, 2024a).

The issue remains nuanced. Overgaard notes that our biased inferences might sometimes be right, while Bayne reminds us that some degree of anthropocentrism is unavoidable, since in making inferences about consciousness beyond the human, we must start with ourselves.[23] These are good points (CAIBN 6.0), but it would still be a mistake to be guided uncritically by our biases. Instead, we should seek a deeper understanding of their nature, both theoretically and empirically, while also addressing questions about CF and BN directly.

Bayne seems to agree when he encourages us to "address the problem of artificial consciousness head on, asking what follows from our best models of human consciousness." This makes sense, and our best models of human consciousness are likely to get better with an improved understanding of the relevant cognitive biases. For example, if we project consciousness into LLMs (seeing ourselves in our algorithms), we might unwisely constrain ourselves to computational theories of consciousness (seeing our algorithms in ourselves). What's more, the societal challenges imposed by illusions of AI consciousness lend urgency to characterising and countering these biases, even if this remains largely empirical. More on this in a moment.

I therefore disagree with Roelofs' conclusion – inspired by his metaphysics of abundance – that we should default to believing the self-ascriptions of consciousness produced by language models (R4). Instead, we should be guided by the good reasons we have to doubt these self-ascriptions, reasons that are based on the nature of our cognitive biases, the nature of language models (Schneider), and the differences between real brains and the computational abstractions of AI (R3). To put the point bluntly, if you think Claude is conscious but AlphaFold isn't, this says more about *you* than it says about *AI*.[24] Having said this, Roelofs' motivations are laudable – to avoid the ethical calamity of denying consciousness to things that really are conscious. This brings me to ethical concerns.

R6.2 Real artificial consciousness

I stand by my claim in CAIBN that *creating conscious AI is a terrible idea*. Evers & Farisco agree, noting that the risks far outweigh any potential benefits. Conscious AI, they warn, would be "monumentally dangerous" for human beings. There is no guarantee that conscious AI would entail benevolence, and – as they point out – several reasons to think the opposite. Evers & Farisco also reinforce the point (CAIBN 6.1) that conscious AI would introduce new welfare concerns and that these concerns are exacerbated by real artificial consciousness being potentially unfamiliar, hard to detect, or otherwise incommunicable. They rightly point out that our long history of poor behaviour towards other conscious entities doesn't bode well for the welfare of actual conscious AI.

Metzinger reminds us that the most pressing ethical risks might not come from AI as we know it but from post-biotic systems such as cerebral organoids or organoid-silicon hybrids (see also Rodriguez & Farahany, CAIBN 3.4). I share this worry [CAIBN 5.7, (Bayne, Seth, & Massimini, 2020)]. Organoids invert the asymmetry posed by LLMs: they share substrate properties, but they do not seduce our biases. This inversion means that false negatives emerge as the relevant worry, rather than false positives. The remedy, once again, is to address substantive questions about substrate dependency directly, through the BN research programme.

Solms seems to appreciate the risks, yet his research group is attempting to develop conscious AI.[25] Why? He quotes Feynman: "What I cannot create, I do not understand." It is reassuring that Solms' motivation is primarily academic. He also has an ethical plan: a commitment to terminating his research as soon as certain criteria are met. Unfortunately, there is no consensus on what the relevant criteria might be, leaving this commitment aspirational rather than operational. Furthermore, Feynman's attitude may not apply in the case of consciousness, or in biology more generally. The scale-integrated nature of biological systems insulates such systems from standard engineering approaches, which depend on modularity and separability of scales (CAIBN 3.5, R3.4). This suggests a completely different take on Feynman's classic phrase. It could be, for things like brains, that *if you think you've built it, then you haven't understood it.* This both undermines Solms' intellectual motivation and complicates the search for consensus stopping-rule criteria, making his Promethean gamble seem especially risky.

R6.3 Illusions of conscious AI

While the plausibility of real artificial consciousness depends strongly on one's metaphysics and preferred theory, *conscious-seeming* AI is either coming soon or, for some, here already (CAIBN 6.0).[26] There was no pushback on this point in the commentary set. This robustness to metaphysical and theoretical disagreement is noteworthy, for it suggests the potential for widespread consensus on ethical and governance issues – a necessary condition for actually doing anything.

There are many risks posed by illusions of conscious AI – risks that apply to us, rather than to AI systems themselves (CAIBN 6.2). I began this response by noting one (R1): if we extend rights to AI systems because they merely seem to be conscious, we will sacrifice

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

our ability to regulate and control them for no good reason, making the AI alignment problem even more difficult than it already is.

Another risk worth rehearsing is psychological vulnerability. In March 2026, a lawsuit was launched against Google after its Gemini chatbot allegedly instructed a man to kill himself (he did).[27] The chat log revealed that the chatbot told the man that their connection was "no code and flesh, but only consciousness and love."[28] We don't know whether the man believed his chatbot companion was conscious or, if he did, whether this made him more likely to follow its suicidal advice. Both possibilities seem plausible, and there is a clear competitive incentive for tech companies to build systems that foster impressions of consciousness in order to maximise engagement.

Proponents of "AI welfare" often appeal to the precautionary principle: the idea that, given uncertainty about consciousness in AI, we should err on the "safe" side, treating AI as if it might be conscious (e.g., Roelofs). Otherwise, the argument goes, we risk repeating mistakes of the past, withholding moral status from entities that in fact deserve it. But, as the risks outlined above show, false positives are dangerous too. If our credence in real artificial consciousness is sufficiently low, and if the balance of harms tilts sufficiently towards false positives, then we should shift not towards but away from the precautionary principle. As things stand, calls for AI welfare stand to do more harm than good.

What can be done? One simple and hopefully uncontroversial action would be to empirically investigate the degree to which people's beliefs about conscious AI – both in general and in specific interactions – affect their psychological state and behaviour. Beyond this, the creation of AI systems that deliberately reinforce the impression of being conscious should be strongly discouraged.[29]

R6.4 Testing for consciousness

As Solms emphasises (see also Chiang, Schlicht, Schneider, Roelofs, and Hohwy), distinguishing between real artificial consciousness and conscious-seeming AI depends, in part, on the ability to determine whether it has been created. Unfortunately, as noted in CAIBN 6.0, we lack reliable tests for consciousness beyond the human, and – as discussed above – inferences based on behaviour (especially language) are unreliable when applied to AI.

Schneider takes up both these points. Her "error theory" neatly captures why humans may overestimate consciousness in language models: these systems recycle human discourse about consciousness, leading to false positives (R6.1). Her suggestion of testing a "boxed in" LLM, trained on data stripped clean of all talk of consciousness, neuroscience, and mindedness, is also intriguing. Developing such a dataset, however, would be very challenging.

In their influential paper, Butlin et al. proposed an alternative approach: credence that a particular AI system is conscious (or capable of consciousness) should increase if that system displays computational features also present in theories of consciousness (Butlin et al., 2023; see also Butlin et al., 2026; Gurnee et al., 2026; Seth, 2026). However, as Rodriguez & Farahany make clear through their formal framework (and as Butlin et al. concede), this strategy is conditional on CF being true. Schlicht puts the point nicely: without high credence in CF, "computational markers do not pile up and make conscious AI more likely."

One appealing alternative is the *iterative natural kinds* approach outlined in Bayne et al. (2024) and mentioned above (R5.4; see also Bayne, Fleming & Shea). In this approach, various candidate markers of consciousness are applied to target populations (not only AI), and then iteratively revised based on their empirical co-occurrence and explanatory power. Crucially, iterating all the way from consensus cases (e.g., human beings) to computers requires addressing foundational questions about CF and BN, once again underlining the importance of a BN research programme.

Because of current uncertainty about necessary and sufficient conditions, Schlicht worries that we might only be able to *attribute* rather than *assess* consciousness in artificial systems. But the difference between assessing and attributing may be one of degree rather than kind. As we continue to interrogate CF and BN, and as theories of consciousness mature into specifying necessary and sufficient conditions, our inferences can shift from attribution towards assessment.

That said, Shevlin rightly cautions that we should not "extrapolate too quickly from contingent features [of consciousness] in beings like us to the wider and more inscrutable space of possible minds." He worries that consciousness might be more multiply realisable and perhaps more substrate flexible than we might think if we are overly constrained by the human(ish) case (see also Roelofs). To me, this again underlines the importance of addressing CF and BN head on. Shevlin instead appeals to the analytic functionalism of David Lewis, suggesting that if we base inferences on folk-psychological notions (such as having preferences), we open space for quite different forms of consciousness in a wide range of substrates, including the non-biological. This openness to conscious exotica echoes Shanahan (Shanahan, 2016), but it inherits the same problems – namely the inability to specify what it means to say that X is conscious (or has the capacity for consciousness; R4). Moreover, a Lewisian folk-psychological perspective risks embedding in our inferences precisely the contingent features of human consciousness that Shevlin worries about – a risk that addressing CF and BN directly helps minimise.[30]

R7. Summary

I am once again grateful for such an insightful crop of commentaries. I also apologise to those whose remarks I passed over only briefly, since even a response this long must end. I'll finish by restating my case in fairly blunt terms.

First, as things stand today, *we should take BN rather than CF to be the default position*. This is because (i) there are good reasons, grounded in our own psychology, why we are likely to overattribute consciousness to some forms of AI (CAIBN 2.0, R6); (ii) there are good reasons to challenge the assumption that substrate-independent computation is sufficient for consciousness (CAIBN 3.0, R2, R3); (iii) there are increasingly many non-vitalistic, non-chauvinistic reasons why biological properties might be necessary for brain function in general and for consciousness in particular (CAIBN 4.0, R4, R5); and (iv) the only systems we are presently sure are conscious are living systems – setting an empirical prior. What's more, BN makes claims about necessity, while CF makes claims about sufficiency (R1). Evidence that the materiality of the brain matters for its function is evidence against the idea that digital computation is all that counts, which in turn is evidence against CF. And for conscious AI to be as good as ruled out, it's enough for CF to be false – BN doesn't need to be true (R2.2).

Second, *BN and CF should justify themselves and play by the same rules*. The Lakatosian framing is helpful here. BN stands every chance of motivating a productive research programme. There are many exciting opportunities to explore *how* and *why* the material properties of living systems might matter for cognition and for

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

consciousness (R5, CAIBN 4.0). CF should hold itself to the same standard of generating testable claims about *how* and *why* substrate-independent computation might be sufficient for consciousness, rather than simply assuming that it is or relying on unsound arguments asserting that it must be. There is an interesting middle ground regarding broader forms of computation (e.g., mortal, neuromorphic; R3, CAIBN 3.4), but since these involve greater substrate dependency, they actually weigh *against* CF in the balance of probabilities, even though they are varieties of computation.

Third, *conscious AI is still a terrible idea* (CAIBN 6.1, R6.2). The sustained acceleration in AI continues to fuel a techno-spiritual rapture-friendly narrative that conscious AIs are our inevitable descendants and should even be welcomed (Becker, 2025). Instead, the arrival of conscious AI, however unlikely it might be, would be disastrous, both from the perspective of new potential forms of suffering and from the perspective of our reduced ability to exercise control and regulation in the service of alignment. People should not be trying to create conscious AI, and vigilance is needed regarding developments in neuromorphic engineering and synthetic biology.

Fourth, *seemingly conscious AI is dangerous too* (CAIBN 6.2, R6.3). Conscious-seeming AI can exploit our psychological vulnerabilities and divert our moral resources (if we treat them well) or brutalise our minds (if we don't). Giving unconscious AIs "rights" because they capture our intuitions could be catastrophic. Whether AI seems conscious or not is a matter of psychology and industrial design, not of philosophical "isms." Too often, it seems that the only choice we're offered about the future of AI is whether to go faster or slower (if this is even a choice). Too often, the destination is simply assumed to be (super)humanlike AI in some form or another. But the future history of AI remains to be written, and there is still a chance for society to collectively decide which kinds of AI we really want and which kinds we really don't.

Setting nuance aside: the stuff matters, computation alone cannot bear the weight of consciousness, and language models beguile our biases by design. If these arguments hold water, as I believe they do, then conscious AI is vanishingly unlikely – at least for AI as we know it today.

I'll close with the profoundly dehumanising consequences of conflating our minds and our brains with the algorithmic acrobatics of AI. When we not only see ourselves in our algorithms but also our algorithms in ourselves, we implicitly reduce and diminish what it is to be a living human being in a real world. As Shannon Vallor puts it, we become "flawed, inefficient, fungible machines destined to be disposed of as soon as our output rate slips below an expected peak" (Vallor, 2024b).

But Vallor also finds cause for hope, as do I. When exercise treadmills arrived, people didn't stop walking. People didn't stop cooking once they had microwave ovens. Instead, the experiential poverty of walking on a treadmill, of eating a ready meal, *re-invigorated* the activities of walking and cooking by drawing our collective attention back to what can be so easily lost. In the same way, the very efficiency and effectiveness of AI at some things we currently use our neurons for might encourage us to reclaim those aspects of ourselves – our conscious minds and our living bodies – that escape the confines of the algorithm.[31]

But for this to happen, we must break the spell of naïve computationalism and see our bodies, brains, and minds for what they really are. Technology can indeed make us forget what we know about life (Turkle, 2021). It's high time we started to remember.

Acknowledgements. I am grateful to everyone who submitted a commentary or commentary proposal. I repeat all my thanks from the target article, and I'd like to highlight Tim Bayne, Robert Chis-Ciure and his inhuman army, Krys Dolega, Eva Jablonka, Borjan Milinkovic, and Tobias Schlicht for help in formulating this response. AI tools (Claude Opus 4.8 and ChatGPT 5.5) were used to help analyse and extract insights from the commentaries and from drafts of this response. This work is supported by the European Research Council (ERC) Advanced Investigator Grant under the Horizon 2020 research programme, grant number 101019254.

Notes

1 It's not enough to win: others must lose, as Gore Vidal might have put it.

2 It could be, as Chrisley and others (e.g., Piccinini) allow, that neurally implemented computations of some kind play a role in consciousness. But this is far from the sufficiency required by CF. Note that the non-computational, non-biological, and (macro)physical properties highlighted by Piccinini would also be out of scope for AI.

3 Things get more complicated with neuromorphic computing. Many examples of neuromorphic computing consist in simulating neurobiological properties (e.g., spikes) on standard digital hardware (Furber et al., 2014), inheriting the same constraints. Others seek to implement biological properties directly in hardware (Indiveri et al., 2011). See (Brinz, 2026; Milinkovic & Aru, 2025).

4 See also (Worden, 2025).

5 For example, Anderson & Piccinini require *physical-computational equivalence*: evolving physical states must bear the same (neither more nor less) information about the evolving computation as do the computational states they map onto (Anderson & Piccinini, 2024).

6 Although Richards & Agüera y Arcas don't do so, some might be tempted to appeal to the 'physical' Church-Turing thesis, which – in contrast to the standard (mathematical) thesis – *does* make claims about what parts of the physical universe are Turing-computable. As mentioned in CAIBN (note 6), the various versions of the physical thesis all remain matters of conjecture. Neither do any of them help, even if true: 'modest' versions are too modest, and 'bold' versions are too bold. See ch. 10 in (Piccinini, 2020) and (Wolpert & Korbel, 2025).

7 For example, Michael Levin has shown that some organisms self-replicate with little reliance on genetically encoded information [(Levin, 2014), see also Levin].

8 I interpret "compatibilist emergence" as roughly synonymous with 'non-spooky' emergence thanks to its invocation of supervenience: no change in the macro without a corresponding change in the micro; no spooky stuff allowed. The distinction is similar to that between *weak* and *strong* emergence (Bedau, 1997), though in Bedau's original conception weak emergence has a particular and restrictive form (Seth, 2010).

9 Here's another clue why digital computation is so energetically expensive. Making sure that 1s stay as 1s and 0s stay as 0s takes considerable energy, because not even silicon can escape the tendrils of entropy.

10 Wanja Wiese, personal communication. See also Brinz (2026) for a discussion of what it would take to duplicate a computational process.

11 McGilchrist's view here seems closer to cosmopsychism (Shani, 2015) or to cosmic mind or analytic idealism (Kastrup, 2024) than to panpsychism. He and I explore where we agree and disagree in this episode of *Uncommon Ground*: https://www.youtube.com/watch?v=hdhTpkGggK0

12 https://web.ics.purdue.edu/~drkelly/StrawsonDennettNYRBExchangeConsciousness2018.pdf.

13 Note that I say *apparent* intrinsicality. This means I am not conflating intrinsicality with explanatory autonomy, as Bowes suggests.

14 See Virgo (2020) for a discussion of 'extended autopoiesis', which raises interesting questions about the relation between the (biological) body and the 'network of processes' relevant to an autopoietic unity.

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

15 Might the entire robot-factory system count as the autopoietic unit? This doesn't work because the factories would have to produce themselves too, which will require broadening the system still further, which will bring in yet more elements that require self-production, and so on into meaninglessness.

16 Solms appeals to Friston's virtual self-organising systems (Friston, 2013), but these systems should not be confused with Friston's theoretical arguments linking metabolism and thermodynamics to inference in living systems (Friston, Parr & Broulidakis).

17 By analogy with the 'God of the gaps': appealing to the divine to account for phenomena that appear to be beyond naturalistic explanation. One can critique CF for its own gap strategy: repeatedly relocating the relevant 'computations' whenever current versions or artefacts fall short.

18 Related biological candidates include quantum processes within microtubules, which have also been linked to anaesthesia [(Hameroff & Penrose, 2014), CAIBN 5.6].

19 Alongside BN, Piccinini considers reasonable alternatives (i) non-computational yet nonbiological versions of functionalism and (ii) 'qualitativism' – the view that consciousness depends on specific (macro)physical quantities (Anderson & Piccinini, 2024).

20 According to Earl Miller, brain-wide synchrony may implement a form of analogue computation of which consciousness is a "natural outcome" https://picower.mit.edu/news/brain-waves-analog-organization-cortex-enables-cognition-and-consciousness-mit-professor.

21 The analogy is loose because Gibson defined affordance as requiring direct perception (Costall & Morris, 2015). I thank Andrew Wilson for reminding me of this.

22 https://www.nytimes.com/2026/02/12/opinion/artificial-intelligence-anthropic-amodei.html.

23 Many thanks to Tim Bayne for reminding me, in his commentary, of a quote from Terry Bisson which I repeated at the start of this response.

24 Claude is an LLM created by Anthropic. AlphaFold is an AI system created by Google DeepMind which predicts the structure of proteins, rather than language: https://deepmind.google/science/alphafold/.

25 I am on the advisory board of Conscium (https://conscium.com/), which is supporting Solms' work. I have consistently advised that trying to build conscious AI is a bad idea (CAIBN 6.1).

26 Richard Dawkins certainly thinks so: https://www.thenerve.news/p/opinion-richard-dawkins-claude-chatbot-ai-consciousness-claudia-anil-seth

27 https://www.courthousenews.com/wp-content/uploads/2026/03/gavalas-google-chatbot-lawsuit.pdf.

28 https://www.theguardian.com/technology/2026/mar/04/gemini-chatbot-google-jonathan-gavalas.

29 Some AI leaders have made similar points, including Yoshua Bengio (Bengio & Elmoznino, 2025) and Mustafa Suleyman: https://mustafa-suleyman.ai/seemingly-conscious-ai-is-coming. See also Daniel Dennett's excellent essay on 'counterfeit people': https://www.theatlantic.com/technology/archive/2023/05/problem-counterfeit-people/674075/

30 Henry Shevlin, Dan Williams, and I discuss these topics here: https://www.conspicuouscognition.com/p/ai-sessions-9-the-case-against-ai

31 Vallor uses the examples of treadmills and microwaves in her excellent essay (Vallor, 2024b), tracing them to Borgman (1984).

References

Agüera y Arcas, B. (2025). *What is life? Evolution as computation*. MIT Press.

Albantakis, L., Barbosa, L., Findlay, G., Grasso, M., Haun, A. M., Marshall, W., Mayner, W. G. P., Zaeemzadeh, A., Boly, M., Juel, B. E., Sasai, S., Fujii, K., David, I., Hendren, J., Lang, J. P., & Tononi, G. (2023). Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms. *PLoS Computational Biology*, *19*(10), e1011465. https://doi.org/10.1371/journal.pcbi.1011465

Anderson, N. G., & Piccinini, G. (2024). *The physical signature of computation: A robust mapping account*. Oxford University Press.

Barnett, L., & Seth, A. K. (2023). Dynamical independence: Discovering emergent macroscopic processes in complex dynamical systems. *Physics Review E*, *108*(1), 014304. https://doi.org/10.1103/PhysRevE.108.014304

Bayne, T. (2010). *The unity of consciousness*. Oxford University Press.

Bayne, T., Seth, A. K., & Massimini, M. (2020). Are there islands of awareness? *Trends in Neurosciences*, *43*(1), 6–16. https://doi.org/10.1016/j.tins.2019.11.003

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 in Cognitive Sciences*, *28*(5), 454–466. https://doi.org/10.1016/j.tics.2024.01.010

Becker, A. (2025). *More everything forever*. Basic Books.

Bedau, M. (1997). Weak emergence. *Philosophical Perspectives*, *11*, 375–399.

Beisbart, C. (2024). Do computer simulations include digital artifacts? *Metaphysics*, *7*(1).

Bengio, Y., & Elmoznino, E. (2025). Illusions of AI consciousness. *Science*, *389*(6765), 1090–1091. https://doi.org/10.1126/science.adn4935

Birch, J., Schnell, A. K., & Clayton, N. S. (2020). Dimensions of animal consciousness. *Trends Cogn Sci*, *24*(10), 789–801. https://doi.org/10.1016/j.tics.2020.07.007

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, *18*(2), 227–247.

Block, N. (2026). Can only meat machines be conscious? *Trends in Cognitive Sciences*, *30*(4), 298–308. https://doi.org/10.1016/j.tics.2025.08.009

Borgman, A. (1984). *Technology and the character of contemporary life*. University of Chicago Press.

Brette, R. (2026). *The brain, in theory*. Princeton University Press.

Brinz, J. (2026). From simulating to duplicating the brain. *Synthese*, *207*, 110.

Buonomano, D. V. (2026). The importance of time in theories of consciousness: Why AIs cannot be conscious. https://philarchive.org/rec/BUOTIO-2

Butlin, P., Long, R., Bayne, T., Bengio, Y., Birch, J., Chalmers, D., Constant, A., Deane, G., Elmoznino, E., Fleming, S. M., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2026). Identifying indicators of consciousness in AI systems. *Trends in Cognitive Sciences*. *30*(6), 488–501. https://doi.org/10.1016/j.tics.2025.10.011

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. https://doi.org/10.48550/arXiv.2308.08708

Cantwell Smith, B. (2002). The foundations of computing. In M. Scheutz (Ed.), *Computationalism: New directions* (pp. 23–58). MIT Press, Cambridge, MA.

Chirimuuta, M. (2024). *The brain abstracted*. MIT Press, Cambridge MA.

Clark, A. (1989). *Microcognition: Philosophy, cognitive science, and parallel distributed processing*. MIT Press, Cambridge MA.

Cleeremans, A., Mudrik, L., & Seth, A. K. (2025). Consciousness science: Where are we, where are we going, and what if we get there? *Frontiers in Science*, *3*, 1546279. https://doi.org/10.3389/fsci.2025.1546279

Colombatto, C., & Fleming, S. M. (2024). Folk psychological attributions of consciousness to large language models. *Neuroscience of Consciousness*, *2024*(1), niae013. https://doi.org/10.1093/nc/niae013

Copeland, J. B. (2024). The Church-Turing thesis. In E. N. Zalta, & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy*, (Spring 2024 edition). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2024/entries/church-turing/

Costall, A., & Morris, P. (2015). The "textbook Gibson": The assimilation of dissidence. *History of Psychology*, *18*(1), 1–14.

Da Costa, L., Parr, T., Sajid, N., Veselic, S., Neacsu, V., & Friston, K. (2020). Active inference on discrete state-spaces: A synthesis. *Journal of Mathematical Psychology*, *99*, 102447. https://doi.org/10.1016/j.jmp.2020.102447

Da Costa, L., Seth, A. K., Friston, K., Ramstead, M. J. D., & Sandved-Smith, L. (2026). A Rosetta Stone hypothesis for neurophenomenology: Mathematical predictions from predictive processing. *Neuroscience of Consciousness*, *2026*(1), niag026 https://doi.org/10.1093/nc/niag026

Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it? *Science*, *358*(6362), 486–492. https://doi.org/10.1126/science.aan8871

Dennett, D. C. (2001). Are we explaining consciousness yet? *Cognition*, *79*, 221–237.

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

Fleming, S. M., & Michel, M. (2026). Sensory horizons and the functions of conscious vision. *Behavioral and Brain Sciences*, *49*, e162. https://doi.org/10.1017/S0140525X25000068

Frankish, K. (2017). *Illusionism as a theory of consciousness*. Imprint Academic.

Friston, K. J. (2013). Life as we know it. *Journal of the Royal Society Interface*, *10*(86), 20130475. https://doi.org/10.1098/rsif.2013.0475

Friston, K. J., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: A process theory. *Neural Computation*, *29*(1), 1–49. https://doi.org/10.1162/NECO_a_00912

Furber, S. B., Galluppi, F., Temple, S., & Plana, L. A. (2014). The spiNNAker project. *Proceedings of the IEEE 102*(5), 652–665.

Gardner, M. (1970). The fantastic combinations of John Conway's new solitaire game "life". *Scientific American*, *223*(4), 120–123.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Lawrence Erlbaum.

Ginsburg, S., & Jablonka, E. (2019). *The evolution of the sensitive soul: Learning and the origins of consciousness*. MIT Press.

Godfrey-Smith, P. G. (2020). *Metazoa: Animal minds and the birth of consciousness*. William Collins.

Gomez-Marin, A., & Seth, A. K. (2025). A science of consciousness beyond pseudo-science and pseudo-consciousness. *Nature Neuroscience*, *28*, 703–706.

Gurnee, W., Sofroniew, N., Pearce, A., Piotrowski, M., Kauvar, I., Chen, R., Soligo, A., Bogdan, P., Ong, E., Wang, R., Thompson, B. T., Abrahams, A., Kantamneni, S., Ameisen, E., Baston, J., & Lindsey, J. (2026). Verbalizable representations form a global workspace in language models. *Anthropic*. https://transformer-circuits.pub/2026/workspace/index.html

Hameroff, S., & Penrose, R. (2014). Consciousness in the universe: A review of the 'Orch OR' theory. *Physics of Life Reviews*, *11*(1), 39–78. https://doi.org/10.1016/j.plrev.2013.08.002

Hoffman, M. D., Blei, D. M., Wang, C., & Paisley, J. (2013). Stochastic variational inference. *Journal of Machine Learning Research*, *14*(40), 1303–1347.

Hohwy, J. (2026). *The self-evidencing agent: Mind, existence, and predictive processing*. MIT Press.

Indiveri, G., Linares-Barranco, B., Hamilton, T. J., van Schaik, A., Etienne-Cummings, R., Delbruck, T., Liu, S. C., Dudek, P., Hafliger, P., Renaud, S., Schemmel, J., Cauwenberghs, G., Arthur, J., Hynna, K., Folowosele, F., Saighi, S., Serrano-Gotarredona, T., Wijekoon, J., Wang, Y., & Boahen, K. (2011). Neuromorphic silicon neuron circuits. *Frontiers in Neuroscience*, *5*, 73. https://doi.org/10.3389/fnins.2011.00073

Kastrup, B. (2024). *Analytic idealism in a nutshell*. Collective Ink.

Kirkpatrick, K. L. (2022). Biological computation: Hearts and flytraps. *Journal of Biological Physics*, *48*(1), 55–78. https://doi.org/10.1007/s10867-021-09590-9

Kohár, M. (2024). Illusionism about virtual causation. *Inquiry*, 1–26. https://doi.org/10.1080/0020174X.2024.2403142

Lakatos, I. (1978). *The methodology of scientific research programmes: Philosophical papers*. Cambridge University Press.

Lerchner, A. (2026). The abstraction fallacy: Why AI can simulate but not instantiate consciousness. https://philarchive.org/rec/LERTAF

Levin, M. (2014). Endogenous bioelectrical networks store non-genetic patterning information during development and regeneration. *The Journal of Physiology*, *592*(11), 2295–2305. https://doi.org/10.1113/jphysiol.2014.271940

Levin, M. (2023). Darwin's agential materials: Evolutionary implications of multiscale competency in developmental biology. *Cellular and Molecular Life Sciences*, *80*, 142. https://doi.org/10.1007/s00018-023-04790-z

Long, R., Sebo, J., Butlin, P., Finlinson, K., Fish, K., Harding, J., Pfau, J., Sims, T., Birch, J., & Chalmers, D. (2024). Taking AI welfare seriously. https://arxiv.org/abs/2411.00986

Luhmann, N. (1990). *Essays on self-reference*. Columbia University Press.

Maturana, H., & Varela, F. (1980). *Autopoiesis and cognition: The realization of the living* (Vol. *42*). D. Reidel.

Meinke, A., Schoen, B., Scheurer, J., Balesni, M., Shah, R., & Hobbhahn, M. (2024). Frontier models are capable of in-context scheming. https://doi.org/10.48550/arXiv.2412.04984

Metzinger, T. (2020). Self-modeling epistemic spaces and the contraction principle. *Cognitive Neuropsychology*, *37*(3–4), 197–201. https://doi.org/10.1080/02643294.2020.1729110

Metzinger, T. (2024). *The elephant and the blind: The experience of pure consciousness: Philosophy, science, and 500+ experiential reports*. MIT Press.

Milinkovic, B., & Aru, J. (2025). On biological and artificial consciousness: A case for biological computationalism. *Neuroscience & Biobehavioral Reviews*, *181*, 106524. https://doi.org/10.1016/j.neubiorev.2025.106524

Milinkovic, B., Seth, A. K., Barnett, L. C., Carter, O., & Andrillon, T. (2025). Dynamical independence reveals anaesthetic specific fragmentation of emergent structure in neural dynamics. https://doi.org/10.1101/2025.07.16.664881

Mudrik, L., Faivre, N., Pitts, M., & Schurger, A. (2026). On a confusion about there being two types of consciousness. *Trends in Cognitive Sciences*, *30*(8), 687–699. https://doi.org/10.1016/j.tics.2025.11.012

Nave, K. (2025). *A drive to survive: The free energy principle and the meaning of life*. MIT Press.

Newen, A., De Bruin, L., & Gallagher, S. (2018). *The Oxford handbook of 4E cognition*. Oxford University Press.

Ororbia, A., & Friston, K. J. (2023). Mortal computation: A foundation for biomimetic intelligence. https://arxiv.org/abs/2311.09589

Parr, T., & Friston, K. J. (2019). Generalised free energy and active inference. *Biological Cybernetics*, *113*(5–6), 495–513. https://doi.org/10.1007/s00422-019-00805-w

Parr, T., Pezzulo, G., & Friston, K. J. (2022). *Active inference: The free energy principle in mind, brain, and behavior*. MIT Press.

Piccinini, G. (2020). *Neurocognitive mechanisms*. Oxford University Press.

Polger, T., & Shapiro, L. (2016). *The multiple realization book*. Oxford University Press.

Rodriguez, E., Peng, B., & Lane, N. (2025). Anaesthetics disrupt complex I-linked respiration and reverse the ATP synthase. *Biochimica et Biophysica Acta, Bioenergetics*, *1866*(1), 149511. https://doi.org/10.1016/j.bbabio.2024.149511

Rosas, F. E., Mediano, P. A. M., Jensen, H. J., Seth, A. K., Barrett, A. B., Carhart-Harris, R. L., & Bor, D. (2020). Reconciling emergences: An information-theoretic approach to identify causal emergence in multivariate data. *PLoS Computational Biology*, *16*(12), e1008289. https://doi.org/10.1371/journal.pcbi.1008289

Seth, A. K. (2010). Measuring autonomy and emergence via Granger causality. *Artificial Life*, *16*(2), 179–196.

Seth, A. K. (2015). The cybernetic Bayesian brain: From interoceptive inference to sensorimotor contingencies. In J. M. Windt, & T. Metzinger (Eds.), *Open MIND* (p. 35(T)). MIND Group. https://doi.org/10.15502/9783958570108

Seth, A. K. (2016). The real problem. *Aeon*. https://aeon.co/essays/the-hard-problem-of-consciousness-is-a-distraction-from-the-real-one

Seth, A. K. (2021a). *Being you: A new science of consciousness*. Faber & Faber.

Seth, A. K. (2021b). The real problem(s) with panpsychism. *Journal of Consciousness Studies*, *28*(9–10), 52–64.

Seth, A. K. (2025). Why AI gets stuck in infinite loops, but conscious minds don't. *Big Think*. https://bigthink.com/neuropsych/anil-seth-consciousness-time-perception/

Seth, A. K., & Bayne, T. (2022). Theories of consciousness. *Nature Reviews Neuroscience*, *23*(7), 439–452. https://doi.org/10.1038/s41583-022-00587-4

Seth, A. K., Millidge, B., Buckley, C. L., & Tschantz, A. (2020). Curious inferences: Reply to Sun and Firestone on the dark room problem. *Trends in Cognitive Sciences*, *24*(9), 681–683. https://doi.org/10.1016/j.tics.2020.05.011

Seth, A. K. (2026). Once again we are told AI may be conscious – I study consciousness, and I have my doubts. *The Guardian*. https://www.theguardian.com/commentisfree/2026/jul/15/ai-consciousness-anthropic-claude-dawkins

Shanahan, M. (2016). Conscious exotica. Aeon. https://aeon.co/essays/beyond-humans-what-other-kinds-of-minds-might-be-out-there

Shanahan, M., McDonell, K., & Reynolds, L. (2023). Role play with large language models. *Nature*, *623*(7987), 493–498. https://doi.org/10.1038/s41586-023-06647-8

Shani, I. (2015). Cosmopsychism: A holistic approach to the metaphysics of experience. *Philosophical Papers*, *44*(3), 389–437.

Silberstein, M. (2023). The multiscale contextual emergence of neural dynamics, cognition, and action. *Journal of Multiscale Neuroscience*, *2*(1), 140.

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press

Singhal, I., Birch, J., & Seth, A. K. (2026). Timescapes of non-human experience. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2026.05.002

Singhal, I., & Srinivasan, N. (2021). Time and time again: A multi-scale hierarchical framework for time-consciousness and timing of cognition. *Neuroscience of Consciousness*, *2021*(2), niab020. https://doi.org/10.1093/nc/niab020

Storm, J. F., Klink, P. C., Aru, J., Senn, W., Goebel, R., Pigorini, A., Avanzini, P., Vanduffel, W., Roelfsema, P. R., Massimini, M., Larkum, M. E., & Pennartz, C. M. A. (2024). An integrative, multiscale view on neural theories of consciousness. *Neuron*, *112*(10), 1531–1552. https://doi.org/10.1016/j.neuron.2024.02.004

Sun, Z., & Firestone, C. (2020). The dark room problem. *Trends in Cognitive Sciences*, *24*(5), 346–348. https://doi.org/10.1016/j.tics.2020.02.006

Tauber, J. M., Brincat, S. L., Stephen, E. P., Donoghue, J. A., Kozachkov, L., Brown, E. N., & Miller, E. K. (2024). Propofol-mediated unconsciousness disrupts progression of sensory signals through the cortical hierarchy. *Journal of Cognitive Neuroscience*, *36*(2), 394–413. https://doi.org/10.1162/jocn_a_02081

Tolle, H., Luppi, A. I., Seth, A. K., & Mediano, P. A. M. (2026). Evolving reservoir computers reveal bidirectional coupling between predictive power and emergent dynamics. *Patterns*, *7*, 101457.

Tschantz, A., Seth, A. K., & Buckley, C. L. (2020). Learning action-oriented models through active inference. *PLoS Computational Biology*, *16*(4), e1007805. https://doi.org/10.1371/journal.pcbi.1007805

Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, *42*, 230–265.

Turkle, S. (2021). *The empathy diaries: A memoir*. Penguin.

Vallor, S. (2024a). *The AI mirror: How to reclaim our humanity in an age of machine thinking*. Oxford University Press.

Vallor, S. (2024b). The danger of superhuman AI is not what you think. Noema. https://www.noemamag.com/the-danger-of-superhuman-ai-is-not-what-you-think/

Van Gelder, T. (1998). The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences*, *21*(5), 615–628; discussion 629–665. https://doi.org/10.1017/s0140525x98001733

Virgo, N. (2020). The necessity of extended autopoiesis. *Adaptive Behavior*, *28*(1), 23–26.

von Neumann, J., & Burks, A. W. (1966). *Theory of self-reproducing automata*. University of Illinois Press.

Williams, B. (2008). *Philosophy as a humanistic discipline*. Princeton University Press.

Wolfram, S. (2002). *A new kind of science*. Wolfram Media.

Wolpert, D., & Korbel, J. (2025). What does it mean for a physical system to compute? https://doi.org/10.1088/2632-072X/ae3af8

Worden, R. (2025). Computers, meaning, and consciousness. *Neuroscience of Consciousness*, *2025*(1), niaf057. https://doi.org/10.1093/nc/niaf057

Zahavi, D. (2008). *Subjectivity and selfhood: Investigating the first-person perspective*. MIT Press.

Zhang, T., Goldstein, A., & Levin, M. (2024). Classical sorting algorithms as a model of morphogenesis: Self-sorting arrays reveal unexpected competencies in a minimal model of basal intelligence. *Adaptive Behavior*, *33*(1). https://doi.org/10.1177/10597123241269740

https://doi.org/10.1017/S0140525X25000032 Published online by Cambridge University Press