Behavioral and Brain Sciences (forthcoming)

Conscious artificial intelligence and biological naturalism

Anil K. Seth[1,2]

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

Correspondence: a.k.seth@sussex.ac.uk, www.anilseth.com

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.

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

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 et al., 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 us lead to overestimate its likelihood. I will

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

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 impenetrable 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 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 competences may still exhibit intelligent behaviour. More general aspects of intelligence may include the ability to flexibly 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 in digital computation implemented in established architectures (CPUs, GPUs, TPUs, etc), and potentially involving real or virtual interactions with (e.g., robotic) bodies and environments (Russell & Norwig, 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

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

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

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 Kubrik'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).

[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.

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

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, 2023; 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 them with experiential capacities. 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 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

[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.

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

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 organization (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; Shagir, 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]

[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.

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

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 'halting problem' posed by Hilbert (Turing, 1936). Other examples include 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 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 are computable, and since not everything implements computations, it is worth interrogating this assumption.

---

[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 says (unconvincingly, in my 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. M. Anderson & G. Piccinini, 2024; Chalmers, 1994; Floridi, 2010).

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

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 (Shagir, 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]

---

[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

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

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 unlikely – 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 need 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 need 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

---

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 the 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.

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

argues that because mental functions arose through evolution, the realizers 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 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 et al., 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 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.

---

[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).

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, grows quickly.

One implication of this argument is that biological brains, which are highly energy efficient, cannot be implementing 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) (Ullman, 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 et al., 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]

---

[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).

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

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 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 which 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 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

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

consciousness[19]; philosophical considerations about intentionality (Shagir, 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 et al., 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.

---

[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.

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

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, 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)

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

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 noncomputational) 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 area 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

[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'".

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

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 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 that 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

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

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.

[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).

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

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 contents, 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 et al., 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 explaintheir 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 (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

[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*.

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

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 et al., 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

---

[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; 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; 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; Parr et al., 2022).

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

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 et al., 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 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 et al., 2013), see also (Fields et al., 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 is 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 et al., 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-

[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).

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

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 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 a 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.

---

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

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

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

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

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 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), 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).

SCENARIO	DESCRIPTION	ASSUMES COMPUTATIONAL FUNCTIONALISM	DOES THE SUBSTRATE MATTER?	EXAMPLE
Naïve along for the ride	Consciousness will just emerge as AI gets smarter	Yes (Turing)	No	Large language models (LLMs)
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]
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]
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])
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

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,*

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

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).

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 et al., 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 its 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) 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 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

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

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

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 anyway to be true, 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

[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 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.

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

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, 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 (Gomez-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 biological processes (McFadden & Al-Khalili, 2018; Mohseni et al., 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

[39] For an alternative account linking the 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.

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

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, develop 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 et al., 2007; Froese, 2017; Kiverstein, 2020; Thompson, 2007; Varela et al., 1993)].

The relevance of these properties to the prospects for conscious AI various 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 et al., 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

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

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

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, in terms of sufficient 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

[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.

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

as to avoid negatively valenced experience. But this would be hard or 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 & Norwig, 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

[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).

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

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.

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]

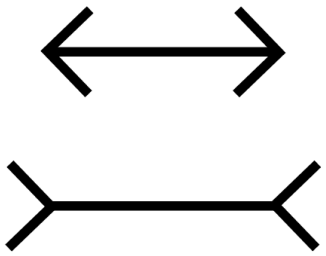

Figure 1. *The Müller-Lyer illusion.*

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 a 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 'West World' 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 West World, and is a line of thinking

[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.

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

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).

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 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 et al., 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

[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.

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

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.

Funding

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).

Conflict of interest

[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

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

References

Aguilera, M., Millidge, B., Tschantz, A., & Buckley, C. L. (2022). How particular is the physics of the free energy principle? *Phys Life Rev*, *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. *Proc Biol Sci*, *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 Comput Biol*, *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.

Anderson, N. M., & Piccinini, G. (2024). *The physical signature of computation*. 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 Neurosci*, *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. *Phys Rev E*, *108*(1-1), 014304. https://doi.org/10.1103/PhysRevE.108.014304

Barrett, L. F., & Simmons, W. K. (2015). Interoceptive predictions in the brain. *Nat Rev Neurosci*, *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 Neurosci*, *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 Cogn Sci*, *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. *Philoosophy 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). 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 P. A. & D. Stoljar (Eds.), *Blockheads! Essays on Ned Block's philosophy of mind and consciousness*. MIT Press.

Blum, L., & Blum, M. (2024). AI Consciousness is Inevitable: A Theoretical Computer Science Perspective. *ArXiv*, arXiv:2403.17101. https://doi.org/ 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. *J Math Psychol*, *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*(11), 243-255.

Brown, R., Lau, H., & LeDoux, J. E. (2019). Understanding the Higher-Order Approach to Consciousness. *Trends Cogn Sci*, *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. *Behav Brain Sci*, *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. (2023). Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. https://doi.org/ https://doi.org/10.48550/arXiv.2308.08708

Cantwell Smith, B. (2002). The Foundations of Computing. In M. Scheutz (Ed.), *Computationalism: New Directions*. Cambridge, MA.

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.

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*. Blackwell Publlishing 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. *Proc Natl Acad Sci U S A*, *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. *Behav Brain Sci*, *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*. 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. (2023). Folk psychological attribuitions of consciousness to large language models. https://doi.org/ https://doi.org/10.31234/osf.io/5cnrv

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. D. (2022). Laying down a forking path: Tensions between enaction and the free energy principle. *Philosophy and the Mind Sciences*, *3*.
Douglas, R., Mahowald, M., & Mead, C. (1995). Neuromorphic analogue VLSI. *Annual Review of Neuroscience*, *18*, 255-281.
Dung, L., & Kersten, L. (2024). Implementing artitifical consciousness. *Mind and Language*. 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 Comput Biol*, *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. *Phys Rev Lett*, *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. *Phys Life Rev*, *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 Netw*, *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). 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/https://doi.org/10.48550/arXiv.2412.04571

Fleming, S. M. (2020). Awareness as inference in a higher-order state space. *Neurosci Conscious*, *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*. https://doi.org/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. *Philos Trans R Soc Lond B Biol Sci*, *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/nrn2787 [pii]
10.1038/nrn2787

Friston, K. J. (2013). Life as we know it. *J R Soc 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?). *Front Psychol*, *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. *Neural Comput*, *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.
Gomez-Marin, A., & Seth, A. K. (2025). A science of consciousness beyond pseudo-science and pseudo-consciousness. *Nat Neurosci*.
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. *Adv Sci (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. *Phys Life Rev*, *11*(1), 39-78. https://doi.org/10.1016/j.plrev.2013.08.002
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. *Nat Commun*, *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*, 1-27. 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. *Front Psychol*, *15*, 1362658
. https://doi.org/doi: 10.3389/fpsyg.2024.1362658
Jarzynski, C. (1997). Nonequilibrium equality for free energy differences. *Phys Rev Lett*, *78*(14), 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.
Jonas, H. (2001). *The phenonomenon 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 e3958. 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 Cent Sci*, *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). https://doi.org/ARTN 169
10.3390/e19040169

Kirkpatrick, K. L. (2022). Biological computation: hearts and flytraps. *J Biol Phys*, *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*, *6*. https://doi.org/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*.
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. *Philos Trans R Soc Lond B Biol Sci*, *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* (Vol. 42). D. Reidel.
McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biology*, *52*, 99-115.
McFadden, J. (2020). Integrating information in the brain's EM field: the cemi field theory of consciousness. *Neurosci Conscious*, *2020*(1), niaa016. https://doi.org/10.1093/nc/niaa016
McFadden, J., & Al-Khalili, J. (2018). The origins of quantum biology. *Proc Math Phys Eng Sci*, *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. *7*, 378. https://doi.org/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). Artificial brains. 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.
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
Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proc Natl Acad Sci U S A*, *120*(13), e2215907120. https://doi.org/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. *Front Artif Intell*, *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. *Neurosci Biobehav Rev*, *170*, 106053. https://doi.org/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. *Behav Brain Sci*, *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. *Nat Neurosci*, *2*(1), 79-87. https://doi.org/10.1038/4580

Rosas, F. E., Geiger, B. C., Luppi, A. I., Seth, A. K., Polani, D., Gastper, 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., & Norwig, 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. *Nat Comput Sci*, *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 of the American Philosophical Society*, *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*. John Wiley and Sons Ltd.

Sengupta, B., Stemmler, M. B., & Friston, K. J. (2013). Information and efficiency in the nervous system--a synthesis. *PLoS Comput Biol*, *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 306-297. 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*

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*.

Seth, A. K., & Bayne, T. (2022). Theories of consciousness. *Nat Rev Neurosci*, *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). https://doi.org/ARTN 20160007

10.1098/rstb.2016.0007

Seth, A. K., & Tsakiris, M. (2018). Being a Beast Machine: The Somatic Basis of Selfhood. *Trends Cogn Sci*, *22*(11), 969-981. https://doi.org/10.1016/j.tics.2018.08.008

Shagir, O. (2010). The rise and fall of computational functionalism. In *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*. 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/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 Cogn Sci*, *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.

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 Ment Health Res*, *3*(1), 12. https://doi.org/10.1038/s44184-024-00056-z

Sterling, P. (2012). Allostasis: a model of predictive regulation. *Physiol Behav*, *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* (Vol. 1478, 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 Cogn Sci*, *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. *Nat Rev Neurosci*, *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*, *49*, 433-460.
Turkle, S. (2021). *The empathy diaries: a memoir*. Penguin.
Ullman, 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 Neurosci*. 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. https://arxiv.org/abs/2410.06633
Wiese, W. (2024). Artificial consciousness: a perspective from the free energy principle. *Philosophical Studies*. https://doi.org/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* (Vol. 2, pp. 185-208). Springer.