Philosophy of AI
Vol. 2, 2026, 186-200
DOI: 10.18716/ojs/phai/2026.3037
©2026 by the author(s)

IN SEARCH OF A BIOLOGICAL CRUX FOR AI CONSCIOUSNESS

Bradford Saad[*]

University of Oxford, Faculty of Philosophy, Oxford, United Kingdom

[*]*brad.saad@philosophy.ox.ac.uk*

Abstract: Whether AI systems could be conscious is often thought to turn on whether consciousness is closely linked to biology. Roughly, the thought is: if consciousness is closely linked to biology, then AI consciousness is impossible; and if consciousness is not closely linked to biology, then AI consciousness is possible—or, at any rate, it's more likely to be possible. A clearer specification of the kind of link between consciousness and biology that is crucial for the possibility of AI consciousness would facilitate inquiry into the prospects for AI consciousness. After motivating desiderata that such a specification should meet, I argue that proposed views about the relationship between consciousness and biology tend not to capture such a crucial link. In addition, I offer a thesis that does capture the link: *the biological requirement* according to which being conscious at least nomically requires having biological states.

Key words: AI consciousness; consciousness and biology; theories of consciousness; substrate independence; functionalism

1. Introduction

Could an AI system be conscious? That is, can we in principle build an intelligent, non-biological machine that has experiences?

Tackling this long-standing philosophical question has recently become a practically urgent task. While few researchers take existing AI systems to be conscious, many take such systems to be a realistic near-term possibility. In a 2025 survey about computer-based systems with the capacity for consciousness, most experts assigned at least a 20% probability to such systems arriving by 2030 and at least a 50% probability to such systems arriving by 2050.[1] Already, philosophers are weighing in on the evaluation of near-term AI systems for consciousness as well as policies concerning the treatment of systems that are deemed serious candidates.[2] The question is being taken up beyond the academy at the frontier of AI development: one leading AI company—Anthropic—is now launching a research program

1. See Caviola and Saad (2025: 16).
2. See, for example, Birch (2024: Part V), Butlin et al. (2023: §1.2), Dung (2025), Long et al. (2024), and Sebo and Long (2025).

on model welfare, following the company's support for a 2024 report on the need to evaluate near-term AI systems for moral patiency partly based on their potential for consciousness.[3]

Whether AI systems could be conscious is often thought to depend crucially on whether consciousness is closely linked to biology.[4] This thought manifests under various guises. For example, philosophy of mind textbooks standardly echo history. These textbooks typically start by reviewing problems with dualism and behaviorism. They then explore a dialectic between functionalist views that characteristically permit consciousness to be realized in non-biological machines and identity theories on which mental states such as consciousness are identical with neurobiological states.[5]

In other contexts, the thought is conveyed with a hardware-software analogy. Thus, Pinker (1997: 3) writes "How can I be so confident that consciousness is secreted by the brain tissue in my skull, rather than lurking in the software that my brain is running — software that … [a] computer could run just as well?" Similarly, in consciousness research a division between biological views and rivals is often marked as key. For example, Block highlights "the battle between two different conceptions of the mind, the biological and the computational" and defends a biological view on which silicon systems cannot be conscious against rival representationalist and functionalist views on which they can (Block, 2007: 169, 249, 279; 2023: 451-457).

Expressions of the thought can also be found in work on AI consciousness. For instance, McClelland (2025: 1) writes, "[i]n the current debate, the main division is between biological views that are sceptical of artificial consciousness and functionalist views that are sympathetic to it." Prinz (2003: 128) says "I'm not sure whether [aspects of my biological makeup] are relevant [to consciousness] or not. With other human beings, this uncertainty is moot because they share my biology. With computers, the uncertainty forces me into a position of agnosticism." And Dung (2024: 2) notes, "The divide between biological and non-biological views seems to be the most important determinant for researchers' views on the prospects of artificial consciousness".

As these examples illustrate, the thought under consideration is not about specific biological features. Rather it concerns a kind of relationship between consciousness and biology while leaving open exactly which biological features participate in that relationship. Put more precisely, then, the thought is: if consciousness is closely linked to some or other biological features, then AI consciousness is impossible; and if consciousness is not closely linked to some such features, then AI consciousness is possible—or, at any rate, it's more likely to be possible. Given this thought, a natural strategy for making progress on whether AI systems could be conscious is thus to examine evidence for and against a close link between consciousness and biology. Although I think this strategy is worth pursuing, it faces a methodological obstacle. This paper will introduce the obstacle, explore its contours along with suggestions for overcoming it, and offer a way forward.[6]

The methodological obstacle is that (i) there are importantly different candidates for kinds of close links between consciousness and biology (e.g. identity, causation, correlation), (ii) it's not clear what kind of close link is crucial to the possibility of AI consciousness, and

3. See Long et al. (2024) and Anthropic (2025).

4. See, for example, Block (2007: 249, passim); compare Block (2023: 452-454), Godfrey-Smith (2023, 2024a, 2024b), Hill (1991: 224-227), and Seth (2021: Ch. 13); compare Chalmers (2023: 18), Karnofsky (2021: 40, fn30), and McLaughlin (2019: 366).

5. See Bayne (2022: Chs. 3, 9), Braddon-Mitchell and Jackson (2007: Ch. 6), Heil (2013: Chs. 3, 4), Kind (2020: Chs. 3-4), Kim (2010: Chs. 3, 4, 10), Mandik (2023: Chs. 6, 8), and Ravenscroft (2005: Ch. 3-4).

6. Aside from noting that the widely-held thought under consideration motivates this strategy, I will assume rather than argue that the strategy is worth pursuing. Those who are skeptical of this strategy may nonetheless find my discussion of the obstacle this strategy faces useful as material for motivating alternative strategies. (I am grateful to an anonymous reviewer for helpful comments here.)

(iii) before we can evaluate how evidence for a close link bears on the possibility of AI consciousness, we first need to be clear about what link is at issue.

This obstacle could be overcome by supplying a crucial thesis that clearly specifies the relevant phenomenal-biological link. However, supplying a crucial thesis turns out to be less straightforward than one might have thought. To find a crucial thesis, we cannot simply take a stock biological view off the shelf. For, as we will see, many of these views—including prominent views that focus on a phenomenal-biological relationship rather than specific biological features—do not supply a clear phenomenal-biological link that is crucial to the possibility of AI consciousness. I hasten to add that there are often good pedagogical, historical, or dialectical reasons for focusing on such views. Thus, my claim that these views do not specify a crucial link should not be read as a blanket criticism of the literature. At the same time, improving our epistemic position regarding the possibility of AI consciousness is now a practically and morally pressing topic.[7] Given the stakes, I think that inquiry into the topic should be beholden to the epistemological joints, even if this requires departure from well-worn paths that have been shaped by other considerations. Further, given the vexed character of the topic, we should be eager to pick low-hanging methodological fruits. For these reasons, it is worth seeking a crucial thesis.

In what follows, I will undertake three tasks. Although I will develop their respective contributions as a package, one could in principle accept some of these but not others. First, I will offer a set of desiderata that should be met by a crucial thesis. Second, I will develop the noted obstacle. More specifically, I will survey an unusually comprehensive range of views about the relationship between consciousness and biology and explain why many of these views fall well short of qualifying as a crucial thesis. Third, I will offer a thesis that captures such a link and which is poised to facilitate the strategy sketched above for making progress on whether AI systems could be conscious. I call this thesis the 'biological requirement'. Roughly, the biological requirement says that, at least in our world, having experiences requires having biological states.[8]

To be clear, the aim here is to defend the biological requirement's standing as a crucial thesis, not its truth. One upshot will be that the biological requirement is well-suited to play a guiding role in investigations of the relationship between consciousness and biology and the possibility of AI consciousness. By asking whether a given consideration concerning the relationship between consciousness and biology bears on the biological requirement, we can filter evidence for relevance to the possibility of AI consciousness. And by asking how different kinds of relevant evidence bear on the biological requirement, we can translate them into a common epistemological currency.

2. What We Seek in a Crucial Thesis

Not all candidate links between consciousness and biology bear significantly on the possibility of AI consciousness. For example, discovering that the locus of human consciousness is the brain rather than the heart has no immediate bearing on whether an AI system could be conscious. For the purpose of investigating how evidence concerning the relationship between consciousness and biology bears on the possibility of AI consciousness, it would be helpful to identify a *crucial thesis*. Roughly, a crucial thesis should be a claim about

7. See, for example, Bradley and Saad (2025).

8. In passing, Chalmers (2023: 10) addresses carbon-based biology as a candidate requirement for consciousness that large language models would fail to meet, while Sebo and Long (2025: §3.1) discuss a biological carbon-based substrate requirement alongside a biological function requirement.

the relationship between consciousness and biology such that evidence concerning that relationship bears on whether AI consciousness is possible via—and only via—that thesis.

More precisely, a crucial thesis should be a clear claim concerning the relationship between consciousness and biology such that:

(1) insofar as evidence confirming a close link between consciousness and biology disconfirms the possibility of AI consciousness, that evidence tends to do so via this thesis,

(2) insofar as evidence disconfirming a close link between consciousness and biology confirms the possibility of AI consciousness, that evidence tends to do so via this thesis, and

(3) the thesis tends to be insensitive to other sorts of evidence.

The stronger a thesis is with respect to the tendencies specified by (1)-(3), the more crucial it is. An ideal crucial thesis would be maximal along each of these dimensions. Although it makes sense to aspire to this ideal, it would be unsurprising if there are tradeoffs between these factors that put an ideal crucial thesis out of reach. A reasonable hope is that we will find a clear, simple, and highly crucial thesis, which can then serve as an evidential focal point through which the large, disparate, and growing body of evidence concerning consciousness and biology can be regimented and brought to bear on the possibility of AI consciousness.[9]

3. Proposed Biological Views as Candidates for a Crucial Thesis

I'll now explain why various biological views of consciousness in the philosophical and scientific literatures on consciousness are not suited for this purpose.

A natural suggestion for finding a crucial thesis is to look to the science of consciousness. The hope is that the science of consciousness will deliver a crucial thesis by way of a proposed neural correlate or scientific theory of consciousness. An initial worry about this suggestion is that—as noted in §1—specific biological features are not at issue in the thought that the possibility of AI consciousness turns on whether there is a close phenomenal-biological link. Consequently, a thesis that appeals to specific scientific features such as neural correlates would not capture the thought. To see that this worry is misplaced, note that we are seeking a crucial thesis that fills the methodological role described in §2, not a conceptual analysis of the noted thought.

However, there are two deeper difficulties with this suggestion. First, there are too many live candidates for the neural correlates of consciousness. The same goes for live candidates for the true scientific theory of consciousness.[10] Discovering that any one of these theories or correlates should be rejected would leave a large field of candidates for links between consciousness and biology that would preclude AI consciousness. Thus, at least at present, there is no scientific theory or candidate for a neural correlate that yields a crucial thesis. Second, work in the science of consciousness is characteristically concerned with investigating what factors make a difference to consciousness within human subjects. But as with

9. In addition to seeking a thesis that scores well with respect to the forms of evidential (in)sensitivity encoded in (1)-(3), one could also seek a thesis that scores well with respect to corresponding suppositional forms of (in)sensitivity, i.e. a thesis such that suppositions about its plausibility suitably modulate the plausibility of the hypothesis that AI consciousness is possible. Satisfying these desiderata would enable a crucial thesis to apply to mysterian views on which whether AI is consciousness turns on whether there is a close link between consciousness and biology and we cannot acquire (significant) evidence concerning that link—see Prinz (2003) and McClelland (2025). I believe much of what I say about how views fare with respect to the evidential desiderata extends to suppositional desiderata. But, for ease of discussion, I will just focus on the former. (I am grateful to an anonymous reviewer for helpful comments here.)

10. For background on the neural correlates of consciousness, see Morales and Lau (2020) and Chalmers (2010b: Chs. 2-4). For background on scientific theories of consciousness, see Butlin et al. (2023) and Seth and Bayne (2022).

discovering that the brain rather than the heart serves as the locus of human consciousness, discovering that consciousness perfectly correlates with one neural state rather than another in humans would have no immediate bearing on the possibility of AI consciousness. A perfect phenomenal-neural correlation in humans is compatible with AI consciousness for two reasons. One is that such a correlation is compatible with a perfect phenomenal-neural-*functional* correlation in humans, and with the correlated functional states giving rise to consciousness in both humans and AI systems.[11] The second is: a perfect phenomenal-neural correlation in humans is compatible with AI systems being conscious in virtue of conditions (e.g. certain silicon states) that no human can satisfy. None of this is to say that results from the science of consciousness are irrelevant to evaluating the possibility of AI consciousness.[12] Rather, the point is just that proposed neural correlates and scientific theories of consciousness are not promising candidates for yielding a crucial thesis.

Next, let's consider some philosophical theses that link consciousness and biology. Within analytic philosophy of mind, the most discussed biological views of consciousness have been forms of *biological reductionism*, the view that every[13] experience in our world is identical with some biological state (presumably a neural state, at least in humans).[14] Biological reductionism would suffice to rule out AI consciousness.

However, biological reductionism is also inapt to serve as a crucial thesis: if we bracket biological reductionism, there remains ample room for evidence concerning consciousness and biology to bear on AI consciousness. One reason for this is that *non-reductive biological views*—i.e. views on which every experience is physically irreducible but has a biological base—also render AI consciousness impossible.[15] On the operative notion of a base, an experience's base is a physical state that (at least) nomically suffices for the experience and does not have any other states as parts that nomically suffice for the experience, where 'nomically suffices for' means entails given the laws of nature.

Non-reductive biological views come in physicalist and dualist forms. *Biological grounding physicalism* says that every experience is distinct from but grounded in a biological state. Similarly, *biological dualism* holds that experiences and physical states are co-fundamental and that experiences have biological states as mere nomic bases. Admittedly, biological dualists could happily accept the possibility of conscious AI systems in worlds with different laws from ours.[16] But that is not the kind of possibility here at issue: the interesting and urgent question in this context is whether we could in principle build a conscious AI—that is, whether the laws of nature permit the creation of such a conscious system. Biological reductionism and non-reductive biological views are on a par in rejecting this possibility.[17] Hence, one need not be a biological reductionist in order to take a close link between

11. Compare Block (2002), Birch (2022: 805), Chalmers (2010b: 99), and Papineau (2003).

12. For example, none of this precludes us from acquiring correlational evidence which confirms a scientific theory of consciousness, invoking philosophical considerations to decide between neural and functional versions of the theory (cf. Chalmers, 1996: Ch. 6; Pautz, 2010b: 348), and then using the favored version to address the possibility of AI consciousness.

13. In formulating universally quantified views about experience, I will take it as read that these views are to be understood as concerning all nomically possible experiences, i.e. all experiences allowed by the laws that hold in our world.

14. See, for example, Block (2009, 2023: Ch. 13), Brown (2012), Hill (1991: 10-12, passim), McLaughlin (2012), Polger (2011), and Papineau (2021).

15. For closely related views, see Kim (2005: Ch. 6), Pautz (2017: 387-9, note 25), and Searle (1997: xiv).

16. In fact, the same is true of biological grounding physicalism and even biological reductionism. This is because, on the adopted formulations, as is standard for formulations of physicalist theses, these views just concern experiences in worlds with our laws, not all metaphysically possible experiences. See, for example, Jackson (1998: 12) and Lewis (1983: 362).

17. Intriguingly, there is a pattern among theorists who at least sympathize with biological views of making concessions to the possibility of consciousness in non-biological systems. For such theorists who favor biological reduction, see, for example, Block (2002, 2023: 445-446), Hill (1991: 224), Lewis (1980, 1994: 420), Papineau (2002: §7.2). For other such theorists, see Brunet and Halina (2020), Kim (2005: 155, 159, 168-173), Godfrey-Smith (2024a: 1665), Searle (1997: 13, 110-111, 175-176, 191, 203), and Seth (2025).

consciousness and biology to rule out the possibility of AI consciousness.[18] By the same token, one need not be a biological non-reductionist in order to rule out the possibility of AI consciousness.

Let's call the hypothesis that every experience has a biological base *biological supervenience*.[19] Biological supervenience leaves open whether experiences are identical with their biological bases, grounded in them, or mere lawful consequences of them. This view captures the shared commitments of biological reductionism and non-reductionist biological views. It also precludes AI consciousness.

Is biological supervenience apt to serve as the crucial thesis? Well, biological supervenience improves on biological reductionism by encompassing some additional biological routes to eliminating the possibility of AI consciousness. However, it does not go far enough, as there remain other routes that it fails to encompass. One route proceeds via *hybrid biological views* on which experiences have bases that are partly biological and partly non-biological.[20] Hybrid biological views have received little attention. But many views that are themselves neutral on whether there is a close link between consciousness and biology come in hybrid versions. To illustrate, let's consider several classes of such theories: externalist theories, internalist theories, and axiarchic theories.

Externalist theories claim that experiences' bases include both internal and external physical factors. On these views, internal physical duplicates might phenomenally differ owing to external differences. Different externalist views differ on what the relevant factors are. *Tracking theories* hold that perceptual experiences' bases include 'tracking' relations—physical relations of covariation—borne to qualities in the environment.[21] *Naive realists* hold that perceptions of such qualities figure in the bases of veridical perceptual experiences.[22] *Embodiment theories* claim that bodily features (such as sensorimotor dispositions) figure in experiences' bases.[23] Proponents of these theories are often silent on what internal factors figure in the basis of experience. In any event, each of these theories can be combined with the hypothesis that neural states figure in the basis of experience or with the hypothesis that they do not. It is also open to tracking theorists to opt for biological-evolutionary accounts of tracking on which tracking a property is a matter of being in a state that covaries with a property under the conditions in which that state was selected for.[24] Similarly, it is open to naive realists to adopt such an account of perception. And it is open to embodiment theorists to maintain that the bodily factors that figure in the bases of experience are biological. None of these commitments is compulsory. But their availability yields hybrid versions of naive realism, tracking theories, and embodiment theories.

---

18. The situation here echoes Chalmers's (1996: 314) observation that the question of whether consciousness is irreducible is quite separate from the question of which systems are conscious. A corollary Chalmers emphasizes is that non-reductionism does not automatically lead to the impossibility of AI consciousness, since some non-reductive views—including the non-reductive form of functionalism Chalmers defends—allow for conscious AIs. The corollary that I am emphasizing is that rejecting biological reductionism does not automatically lead to the possibility of AI consciousness, since some rival views—including some non-reductive biological views—predict the impossibility of AI consciousness. Both of these points merit emphasis because failing to respect the distinction between questions about consciousness's nature and its basis is a persistent source of confusion in discussions of AI consciousness.

19. Compare Kim (2005: 93).

20. For relevant discussion, see Block (2007: 87, Ch. 4, note 14, 340), Braddon-Mitchell and Jackson (2007: 133-4), Chalmers (2010a: "My own view is that any biological theorist should be a "big-state" theorist… Roughly [big-states involve] at least the kind of internal goings-on that would constitute a supervenience base for an internalist functionalist, plus further specifications, e.g. involving biological realization"), Lee (2016: §IV), Lewis (1980), Pautz (2010b: 358-9), Prinz (2012: Ch. 9), and Nagel (2000).

21. For an overview of different tracking theories, see Dalbey and Saad (2022).

22. For recent work on naive realism, see essays in Beck and Masrour (2025).

23. See, for example, O'Regan and Noë (2001).

24. For a tracking theory along these lines, see Dretske (1995).

Internalist theories claim that experiences' bases are exhausted by internal physical factors. Different internalist views differ on what the relevant factors are. There are *scientific theories of consciousness* on which some internal functional or non-phenomenal psychological features figure in the basis of experience. To illustrate, consider a leading scientific theory of consciousness: *the global workspace theory* holds that the basis of experience includes a representational state's being in a global workspace and being broadcast to consumer systems.[25] The global workspace theory comes in a hybrid version—what's known as the *global neuronal workspace theory*—that goes beyond the global workspace theory by specifying the global workspace in neurobiological terms.[26] More generally, we may be able to extract purely functionalist versions of scientific theories by abstracting[27] away from their neural commitments. Similarly, we may be able to extract hybrid versions of these theories by retaining their biological commitments or by leveraging their functional commitments into designators for neural states that play certain functional roles and which are (according to these versions) required for consciousness. In addition, there are philosophical *functionalist* theories on which internal functional differences underpin phenomenal differences. Some theorists have been attracted to the hypothesis that, while functional structure fixes phenomenal structure, further factors are needed to explain why we have one type of experience with a given structure rather than, say, inverted experiences with that structure.[28] This provides a natural opening for a hybrid theory on which biological states figure in the bases of experiences and perform this additional explanatory work. Finally, internalist *Russellian views* hold that the relevant factors include 'quiddities', i.e. the categorical occupants of fundamental physical roles.[29] Russellian monists hold that experiences are at least partly identical with or grounded in quiddities.[30] Similarly, Russellian dualists hold that the nomic bases of experiences are internal physical states that include quiddities. These Russellian views require an explanation of how a sparse stock of quiddities could combine in ways that help generate any experience at all and in ways that help generate the rich and varied experiences of minds like ours.[31] These explanatory demands exert pressure on Russellians to posit non-quiddistic internal factors as part of the basis of experience. This provides a natural opening for Russellians to opt for a hybrid view that appeals to biological factors. For example, Russellians might appeal to the global neuronal workspace to explain why certain combinations of quiddities generate experience.

Axiarchic views claim that the basis of experience includes facts about value or normative reasons. For example, *interpretationist views* hold that the distribution of experience is constrained by what distribution would strike the best balance of a certain set of parameters such as simplicity and rationality.[32] Interpretationist views are motivated by their potential to explain harmonious correlations between experiences and accompanying states (e.g.

25. See Baars (1988). The global workspace theory is sometimes understood as concerning access consciousness in the first instance rather than phenomenal consciousness.

26. Or at least this is so on some formulations of the theory (e.g. see Seth and Bayne, 2022: 441). However, Stanislas Dehaene, one of the main developers of the theory, seems open to AI consciousness (Dehaene, Lau and Kouider, 2017: 54).

27. See Block (2007: 30-32) and Lewis (1970).

28. Compare Kim (2005: 171-173), Shoemaker (1982), and Lee (2016: 216).

29. Whether quiddities count as physical is a merely verbal issue. For convenience, I stipulate that they do.

30. For an overview, see Alter and Pereboom (2023).

31. See, for example, Chalmers (2017).

32. For discussion of axiarchic views and consciousness, see Cutter (2023), Cutter and Saad (2024: §6, 20), Goff (2018: 117, 120). For discussion of interpretationist views of consciousness, see Cutter and Crummett (2025), Saad (2025); compare Pautz (2010a: §9). It might be thought that interpretationist views (and perhaps axiarchic views more generally) are just a special case of functionalist views. This is at least not straightforwardly the case, as there are meta-law interpretationist views on which the first-order psychophysical principles are selected on the basis of which candidates for such principles would yield a maximally rational interpretation (on some measure) and the selected principles are non-functionalist (Saad, 2025). In any event, standard (non-interpretationist) forms of functionalism about experience enjoy different motivations than interpretationist theories.

correlations between experiences and accompanying states that systematically lend themselves to rationalizing interpretations of subjects). However, it is doubtful that the types of factors to which interpretationist views typically appeal could settle all the fine-grained factors about phenomenology—e.g., much of the detail of a typical visual experience seems irrelevant to the rationality of its subject. More generally, since many details of experience seem irrelevant to value and normative reasons, axiarchic theorists are under pressure to countenance further factors in the basis of experience. Because neural factors exhibit sufficient granularity, there is motivation for axiarchic theorists to opt for a hybrid view.

My point in drawing attention to these views is that there is a wide variety of hybrid views, all of which are at odds with biological supervenience but which nonetheless harbor biological commitments that eliminate the possibility of AI consciousness.[33] That possibility is also eliminated by *biological tethering* variants of these views on which experiences' bases are non-biological states that are nomically tethered to biological states. For instance, while it is variously labeled, the most discussed biological tethering view is perhaps a sort of functionalism. Specifically, it's a sort of functionalism on which experiences have non-biological functional bases and these bases are nomically tethered to biological states (e.g. because only a functioning brain can produce the requisite functional states).[34] Another biological tethering view holds that experiences' bases feature energy configurations that can be produced in biological systems but not in artificial systems.[35] The upshot is that a crucial thesis should ideally capture evidential connections between biology and consciousness that flow not just through biological supervenience but also through hybrid and tethering biological views.

It might be suggested that there is little point in investigating complexities that arise from considering hybrid or tethering views or trying to find a crucial thesis that is sensitive to their availability. In response, it should be acknowledged that hybrid and tethering views are often not considered.[36] However, I think that this is because they are rarely relevant in contexts in which rivals to biological supervenience are discussed, not because they have received a fair hearing and been found wanting. For example, naive realist views are usually discussed in the philosophy of perception where they are pitted against representationalist, adverbialist, and sense data rivals, not against biological views. Similarly, Russellian views are often discussed in the metaphysics of mind where they primarily compete against non-Russellian forms of dualism or physicalism. Admittedly, functionalist views are sometimes compared with physicalist forms of biological reductionism. But this is often in contexts (such as philosophy of mind textbooks) where historical precedent or pedagogical considerations tell against considering hybrids or tetherings of functionalist and biological views. Finally, axiarchic views of experience are relative newcomers and have not yet received much discussion. So, it is not surprising that versions of them with biological commitments have not been discussed. However, the fact that hybrid and tethering views are often not worth considering does not mean they are unworthy of consideration in the present context. As I have briefly indicated, at least some hybrid views are motivated developments of views that are already taken seriously. It is morally important that we avoid premature dismissal of such theories that are

33. Or at least perceptual AI consciousness, as on naive realism and tracking theories.

34. For relevant discussion, see Brunet and Halina (2020), Godfrey-Smith (2016), Kleiner (2024), and Wiese (2024).

35. I am not aware of anyone who endorses this view—but for authors who advocate nearby views, see Block (2023: 446), Pockett (2000: Ch. 6), and Thagard (2022).

36. But see Godfrey-Smith (2016, 2023, 2024b); compare Cao (2022) and Shiller (2024). Related proposals contend that there is no sharp divide between the biological and the functional because the biological is functional at some level of functional organization (Lycan, 1981; Dennett, 2001).

relevant to AI consciousness,[37] as such dismissals could lead to errors in decisions about which AI systems to create and how to treat them.

Finally, let us consider whether the doctrine of *substrate independence* can serve as a crucial thesis. On Chalmers's (2022: 93) formulation, substrate independence claims that "consciousness depends only on the [functional] organization of a system and does not depend on the substrate (e.g. biology or silicon) in which the system is implemented."[38] Note that unlike other theses we have considered, substrate independence asserts the absence of a close link rather than the presence of one.

One point against taking substrate independence as our crucial thesis is that substrate independence could be false even in the absence of a close link between consciousness and biology. For example, substrate independence could be false, despite the absence of a close link between consciousness and biology, if consciousness depends partly on a factor that is neither biological nor purely functional (such as tracking qualities in the environment). So, at least on this formulation, substrate independence is not fit to serve as our crucial thesis.

In response, one might suggest that we could extract the crucial thesis by dropping the claim about functional organization. The claim would then be that consciousness is substrate independent in that it does not depend on the substrate in which it is implemented. This formulation fares no better as a candidate for the crucial thesis. To see this, note that substrate independence is often invoked in arguments for taking seriously the hypothesis that we are living in a simulation. But, conceivably, if we are living in a simulation, it might turn out that consciousness depends on a particular non-biological (say, silicon) substrate in the basement level of reality. In that case, consciousness would be substrate dependent and yet not closely linked to biology. Similarly, it might turn out that (i) consciousness can only be implemented by an immaterial substrate, (ii) whether an immaterial substrate implements consciousness depends only on the functional organization of the system to which it is paired, and (iii) immaterial substrates can be paired to both biological and non-biological systems with suitable functional organizations.[39] In that case, consciousness would be substrate dependent but not because of any close link between consciousness and biology. Even if these simulation and substance dualist hypotheses are implausible, they still reveal that the suggested formulation of substrate independence falls short as a candidate for a crucial thesis.

In response, one might suggest we can arrive at a crucial thesis by further tweaking the formulation of substrate independence to say that consciousness does not depend on having a biological substrate. This thesis is close to the complement of the biological requirement I will propose as a crucial thesis in the next section. (How close would turn on a decision about how to unpack 'depends' and 'substrate'.)

4. A Crucial Thesis: the Biological Requirement

Recall that we seek a crucial thesis about the relationship between consciousness and biology. We want a thesis such that agreeing on its plausibility would settle the extent to which the possibility of AI consciousness is (dis)confirmed by evidence against (for) a close link between consciousness and biology. I am now in a position to offer my suggestion for a crucial thesis.

37. This is admittedly a tall order: the range of theories of consciousness to which we should assign a non-negligible credence is wide and varied. Still, we should try. Commendable initial efforts in this direction can be found in Chalmers (2023) and Sebo and Long (2025). These authors evaluate the prospects for AI consciousness by assigning limited confidence to each of many potential markers for consciousness and using those assignments to generate overall estimates.

38. Chalmers (2022: 93) attributes substrate independence to Bostrom (2003), which seems to be the origin of the thesis. Several candidate formulations of substrate independence can be gleaned from Bostrom (2003).

39. Chalmers (2010b: 23-25, 139, fn36) seems open to this possibility.

The biological requirement says that having an experience requires having a biological state.[40] To coordinate with the target hypothesis concerning whether conscious AI systems are nomically possible, the force of this requirement is that of nomic necessity. Because the biological requirement is not committed to biological states nomically sufficing for experiences, it is strictly weaker than biological supervenience.

There is much to recommend the biological requirement as a crucial thesis. For the biological requirement straightforwardly eliminates the possibility of AI consciousness on biological grounds. It also follows from all the biological views we have so far considered that eliminate the possibility of AI consciousness: it is a consequence of all reductive, non-reductive, hybrid, and tethered biological views, and the view that consciousness depends on a biological substrate. More generally, all and only those views of consciousness that rule out the possibility of AI consciousness on biological grounds entail the biological requirement. With respect to crucialness, the biological requirement thus improves on biological reductionism and biological supervenience by encompassing all views that eliminate the possibility of AI consciousness on purely biological grounds. And the biological requirement does this while avoiding various forms of overreach that would encompass views that eliminate the possibility of AI consciousness on purely non-biological grounds. For instance, the biological requirement is appropriately silent on the view that AI systems cannot be conscious because they cannot have souls or free will. The biological requirement also easily adjusts to accommodate restricted versions of the question of whether conscious AI is possible. For example, to adjust the biological requirement to the question of whether AI consciousness will be possible in the next decade, we can ask whether consciousness requires a biological state, *modulo* any bases of consciousness that will only manifest in AI systems more than ten years from now. Finally, the biological requirement is clear and exceedingly simple, and thus well-suited to serve as an organizing principle for inquiry on this topic.

I do not say that the biological requirement is a perfect crucial thesis. One imperfection of the biological requirement is potential sensitivity to evidence that has no bearing on the possibility of AI consciousness. For example, suppose that an oracle told us particles, corporations, stars, and the universe all have experiences and that this has no bearing on the possibility of AI consciousness. Her testimony would tell against the biological requirement without bearing on the possibility of AI consciousness. This imperfection is less pressing than it might seem. For suppose we grant that there could be evidence that supports attributions of consciousness to particles, corporations, stars, or the universe. Still, we might doubt that realistic batches of such evidence could fail to bear on the possibility of AI consciousness. Perhaps any such evidence would support the possibility of AI consciousness indirectly by suggesting that nature favors liberal rather than stringent conditions for consciousness. Even if so, one might still worry that whereas the biological requirement is highly sensitive to such evidence, the possibility of AI consciousness is only mildly sensitive to such evidence. In that case, the biological requirement would still fail to perfectly reflect the *extent* to which evidence against a close link between consciousness and biology supports the possibility of AI consciousness.

Another imperfection is the biological requirement's potential for unwanted sensitivity in response to evidence that bears on the possibility of AI consciousness. We want the crucial thesis to reflect the thought that a close link between consciousness and biology disconfirms the possibility of AI consciousness while the absence of such a link confirms that possibility.

40. To allow for the biological requirement to hold on substance dualism, we can take immaterial minds to 'have' not only their own states but also the states of their physical interaction partners. For argument that we should assign at least a middling credence to AIs having immaterial souls, conditional on AGI and substance dualism, see Cutter (2025).

Yet the biological requirement is potentially sensitive to evidence *for* a close link between consciousness and biology that *confirms* the possibility of AI consciousness; likewise, the biological requirement is potentially sensitive to evidence *against* a close link between consciousness and biology that *disconfirms* the possibility of AI consciousness.

In other words, the biological requirement is potentially sensitive to evidence whose directional bearing on the possibility of AI consciousness is the opposite of what's posited by the target thought. To illustrate, suppose we discovered that recurrent processing of a certain sort is required for consciousness. And suppose that, among the many kinds of non-biological systems that are candidates for conscious subjects, only brains and AI systems that emulate brains engage in that sort of recurrent processing.[41] This evidence would *support* a close link between consciousness and biology and *brighten* the prospects for AI consciousness. Alternatively, suppose we discovered that consciousness covaries with a certain kind of informational integration that does not require biology and which is necessarily absent from AI systems.[42] This is evidence *against* a close phenomenal-biological link that also *dims* the prospects for AI consciousness. Although the biological requirement would be sensitive to each of these discoveries, an ideal crucial thesis that reflects the target thought would be insensitive to these discoveries.

5. Conclusion

To sum up, whether AI systems could be conscious is often taken to depend crucially on whether consciousness is closely linked to biology. Investigating whether consciousness is closely linked to biology is thus a natural strategy for investigating the prospects for AI consciousness. To facilitate that investigation, it would help to have a crucial thesis that clearly specifies the link at issue. I have argued that such a thesis should ideally meet several desiderata, lest it fail to mirror the evidential bearing of the link in question on the possibility of AI consciousness. It turns out that biological views of consciousness in the literature fall well short of this ideal. I offered an alternative candidate for a crucial thesis that improves on this score: the biological requirement.

While the biological requirement does a good job of satisfying the proposed desiderata, we have also seen that the biological requirement does not satisfy all these desiderata perfectly. Should the biological requirement be modified to remove these imperfections? Or replaced entirely? I do not see any general gains to be had by complicating the biological requirement to remove these imperfections that are worth the complexity costs I anticipate. Nor am I aware of a better candidate for a crucial thesis. The moral I draw is that we should by default adopt the biological requirement as a crucial thesis when investigating how the relationship between consciousness and biology bears on the possibility of AI consciousness.

Acknowledgments

For helpful comments, I thank Andreas Mogensen, Derek Shiller, Guido Löhr, and reviewers. For copy editing assistance, I thank Guido Löhr, Claude Sonnet 3.7, 4, 4.6, and 5, and Claude Opus 4.8 and 5.5.

41. Compare Lamme (2010).

42. Compare: proponents of the integrated information theory of consciousness have argued that the type informational integration underpinning consciousness is present in the brain and that even a functionally equivalent neuron-by-neuron simulation of a conscious brain by a classical computer would not replicate consciousness (Tononi and Koch, 2015: 15, fn15; Findlay et al., 2024: 2, 11). Similarly, Wiese (2024) argues for a causal-informational requirement ('causal-flow') on artificial consciousness and contends that the requirement is not met by classical computers. Interestingly, as these theorists note, their challenges to consciousness in classical computers do not necessarily apply to non-classical computers.

References

Alter, T., & Pereboom, D. (2023). Russellian Monism. In *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/archives/fall2023/entries/russellian-monism/

Anthropic. (2025). *Exploring Model Welfare*. https://www.anthropic.com/research/exploring-model-welfare

Baars, B. (1988). *A Cognitive Theory of Consciousness*. Cambridge University Press.

Bayne, T. (2022). *Philosophy of Mind: An Introduction*. Routledge.

Beck, O., & Masrour, F. (2025). *The Relational View of Perception: New Essays*. Routledge.

Birch, J. (2022). Materialism and the moral status of animals. *Philosophical Quarterly*, *72*(4), 795–815. https://doi.org/10.1093/pq/pqab072

Birch, J. (2024). *The Edge of Sentience: Risk and Precaution in Humans, Other Animals, and AI*. Oxford University Press.

Block, N. (2002). The harder problem of consciousness. *The Journal of Philosophy*, *99*(8), 391–425. https://doi.org/10.2307/3655621

Block, N. (2007). *Consciousness, Function, and Representation: Collected Papers*. MIT Press.

Block, N. (2009). Comparing the major theories of consciousness. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (4th ed., pp. 1111–1123). MIT Press.

Block, N. (2023). *The Border Between Seeing and Thinking*. Oxford University Press.

Bostrom, N. (2003). Are we living in a computer simulation? *The Philosophical Quarterly*, *53*(211), 243–255. https://doi.org/10.1111/1467-9213.00309

Braddon-Mitchell, D., & Jackson, F. (2007). *The Philosophy of Mind and Cognition* (2nd ed.). Blackwell.

Bradley, A., & Saad, B. (2025). AI Alignment Versus AI Ethical Treatment: 10 challenges. *Analytic Philosophy*. https://doi.org/10.1111/phib.12380

Brown, R. (2012). The brain and its states. In S. Edelman, T. Fekete, & N. Zach (Eds.), *Being in Time: Dynamical Models of Phenomenal Experience* (pp. 211–230). John Benjamins.

Brunet, T. D. P., & Halina, M. (2020). Minds, machines, and molecules. *Philosophical Topics*, *48*(1), 221–241. https://doi.org/10.5840/philtopics202048111

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2023). *Consciousness in artificial intelligence: Insights from the science of consciousness*. arXiv Preprint. https://doi.org/10.48550/arXiv.2308.08708

Cao, R. (2022). Multiple realizability and the spirit of functionalism. *Synthese*, *200*(6), 506. https://doi.org/10.1007/s11229-022-03524-1

Caviola, L., & Saad, B. (2025). *Futures with digital minds: Expert forecasts in 2025*. https://arxiv.org/abs/2508.00536

Chalmers, D. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

Chalmers, D. (2010a). *Comments on "Why Consciousness Can't Just be in the Head: A New Argument against Biological Theories."* https://consciousnessonline.wordpress.com/2010/02/17/why-consciousness-cant-just-be-in-the-head-a-new-argument-against-biological-theories/

Chalmers, D. (2010b). *The Character of Consciousness*. Oxford University Press.

Chalmers, D. (2017). Panpsychism and Panprotopsychism. In G. Brüntrup & L. Jaskolla (Eds.), *Panpsychism: Contemporary Perspectives* (pp. 19–47). Oxford University Press USA.

Chalmers, D. (2022). *Reality+: Virtual Worlds and the Problems of Philosophy*. Norton.

Chalmers, D. (2023). *Could a large language model be conscious?* https://philarchive.org/rec/CHACAL-3

Cutter, B. (2023). *From moral realism to axiarchism*. https://philarchive.org/archive/CUTFMR
Cutter, B. (2025). The AI ensoulment hypothesis. *Faith and Philosophy*, *41*(1), 1–26. https://doi.org/10.37977/faithphil.2025.41.1.1
Cutter, B., & Crummett, D. (2025). Psychophysical harmony: A new argument for theism. In L. Buchak & D. W. Zimmerman (Eds.), *Oxford Studies in Philosophy of Religion* (Vol. 11, pp. 33–71). Oxford University Press.
Cutter, B., & Saad, B. (2024). The problem of nomological harmony. *Noûs*, *58*(2), 482–504. https://doi.org/10.1111/nous.12463
Dalbey, B., & Saad, B. (2022). Internal constraints for phenomenal externalists: A structure matching theory. *Synthese*, *200*(5), 348. https://doi.org/10.1007/s11229-022-03829-1
Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it? *Science*, *358*(6362), 486–492. https://doi.org/10.1126/science.aan8871
Dennett, D. C. (2001). Are we explaining consciousness yet? *Cognition*, *79*(1–2), 221–237. https://doi.org/10.1016/S0010-0277(00)00130-X
Dretske, F. (1995). *Naturalizing the Mind*. MIT Press.
Dung, L. (2024). *Consciousness without biology: An argument from anticipating scientific progress*. https://philpapers.org/rec/DUNCWB
Dung, L. (2025). How to deal with risks of AI suffering. *Inquiry: An Interdisciplinary Journal of Philosophy*, *68*(7), 2281–2309. https://doi.org/10.1080/0020174X.2023.2238287
Findlay, G., Marshall, W., Albantakis, L., David, I., Mayner, W. G., Koch, C., & Tononi, G. (2024). *Dissociating artificial intelligence from artificial consciousness*. arXiv Preprint. https://doi.org/10.48550/ARXIV.2412.04571
Godfrey-Smith, P. (2016). Mind, matter, and metabolism. *Journal of Philosophy*, *113*(10), 481–506. https://doi.org/10.5840/jphil20161131034
Godfrey-Smith, P. (2023). *Nervous systems, functionalism, and artificial minds*. https://petergodfreysmith.com/wp-content/uploads/2023/12/NYU-Oct-2023-Animals-AI-Functionalism-paper-Post-C3.pdf
Godfrey-Smith, P. (2024a). Inferring consciousness in phylogenetically distant organisms. *Journal of Cognitive Neuroscience*, *36*(8), 1660–1666. https://doi.org/10.1162/jocn_a_02158
Godfrey-Smith, P. (2024b). Simulation scenarios and philosophy. *Philosophy & Phenomenological Research*, *109*(3), 1036–1041. https://doi.org/10.1111/phpr.13124
Goff, P. (2018). Conscious thought and the cognitive fine-tuning problem. *Philosophical Quarterly*, *68*(270), 98–122. https://doi.org/10.1093/pq/pqx039
Heil, J. (2013). *Philosophy of Mind: A Contemporary Introduction* (3rd ed.). Routledge.
Hill, C. S. (1991). *Sensations: A Defense of Type Materialism*. Cambridge University Press.
Jackson, F. (1998). *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Oxford University Press.
Karnofsky, H. (2021). *The most important century*. https://www.cold-takes.com/assets/files/most-important-century-consolidated.pdf
Kim, J. (2005). *Physicalism, or Something Near Enough*. Princeton University Press.
Kim, J. (2010). *Philosophy of Mind* (3rd ed.). Westview Press.
Kind, A. (2020). *Philosophy of Mind* (1st ed.). Routledge.
Kleiner, J. (2024). *Consciousness qua Mortal Computation*. arXiv Preprint. https://doi.org/10.48550/ARXIV.2403.03925
Lamme, V. A. (2010). How neuroscience will change our view on consciousness. *Cognitive Neuroscience*, *1*(3), 204–220. https://doi.org/10.1080/17588921003731586
Lee, G. (2016). Does experience have phenomenal properties? *Philosophical Topics*, *44*(2), 201–230. https://doi.org/10.5840/philtopics201644223

Lewis, D. (1970). How to define theoretical terms. *The Journal of Philosophy*, *67*(13), 427–446. https://doi.org/10.2307/2023861
Lewis, D. (1980). Mad pain and Martian pain. In N. Block (Ed.), *Readings in Philosophy of Psychology, Volume I* (pp. 216–222). Harvard University Press.
Lewis, D. (1983). New work for a theory of universals. *Australasian Journal of Philosophy*, *61*(4), 343–377. https://doi.org/10.1080/00048408312341131
Lewis, D. (1994). Reduction of mind. In S. Guttenplan (Ed.), *Companion to the Philosophy of Mind*. Blackwell.
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*. arXiv Preprint. https://arxiv.org/abs/2411.00986
Lycan, W. G. (1981). Form, function and feel. *Journal of Philosophy*, *78*, 24–50. https://doi.org/10.2307/2025395
Mandik, P. (2023). *This Is Philosophy of Mind: An Introduction* (2nd ed.). Wiley Blackwell.
McClelland, T. (2025). Agnosticism about artificial consciousness. *Mind & Language*. https://doi.org/10.1111/mila.70010
McLaughlin, B. (2019). Could an android be sentient? In A. Pautz & D. Stoljar (Eds.), *Blockheads!: Essays on Ned Block's Philosophy of Mind and Consciousness*. MIT Press.
McLaughlin, B. P. (2012). On justifying neurobiologicalism for consciousness. In S. Gozzano & C. S. Hill (Eds.), *New Perspectives on Type Identity: The Mental and the Physical* (pp. 207–229). Cambridge University Press.
Morales, J., & Lau, H. (2020). The neural correlates of consciousness. In U. Kriegel (Ed.), *The Oxford Handbook of the Philosophy of Consciousness* (pp. 233–260). Oxford University Press.
Nagel, T. (2000). The psycho-physical nexus. In P. Boghossian & C. Peacocke (Eds.), *New Essays on the A Priori*. Oxford University Press.
O'Regan, J. K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, *24*(5), 939–973. https://doi.org/10.1017/S0140525X01000115
Papineau, D. (2002). *Thinking about Consciousness*. Clarendon Press.
Papineau, D. (2003). Could there be a science of consciousness? *Philosophical Issues*, *13*, 205–220. https://doi.org/10.1111/1533-6077.00012
Papineau, D. (2021). *The Metaphysics of Sensory Experience*. Oxford University Press.
Pautz, A. (2010a). A simple view of consciousness. In R. C. Koons & G. Bealer (Eds.), *The Waning of Materialism* (pp. 25–66). Oxford University Press.
Pautz, A. (2010b). Do theories of consciousness rest on a mistake? *Philosophical Issues*, *20*(1), 333–367. https://doi.org/10.1111/j.1533-6077.2010.00189.x
Pautz, A. (2017). The significance argument for the irreducibility of consciousness. *Philosophical Perspectives*, *31*, 349–407. https://doi.org/10.1111/phpe.12104
Pinker, S. (1997). Can a computer be conscious? *US News & World Report*, *123*(7), 63–65.
Pockett, S. (2000). *The Nature of Consciousness*. Writers Club Press.
Polger, T. (2011). Are sensations still brain processes? *Philosophical Psychology*, *24*(1), 1–21. https://doi.org/10.1080/09515089.2010.533263
Prinz, J. (2003). Level-headed mysterianism and artificial experience. *Journal of Consciousness Studies*, *10*(4–5), 111–132.
Prinz, J. (2012). *The Conscious Brain*. Oxford University Press.
Ravenscroft, I. (2005). *Philosophy of Mind: A Beginner's Guide*. Oxford University Press.
Saad, B. (2025). Lessons from the void: What Boltzmann brains teach. *Analytic Philosophy*, *66*(4), 594–621. https://doi.org/10.1111/phib.12353

Searle, J. (1997). *The Mystery of Consciousness*. The New York Review of Books.
Sebo, J., & Long, R. (2025). Moral consideration for AI systems by 2030. *AI and Ethics*, *5*, 591–606. https://doi.org/10.1007/s43681-023-00379-1
Seth, A. (2021). *Being You: A New Science of Consciousness*. Penguin.
Seth, A. (2025). Conscious artificial intelligence and biological naturalism. *Behavioral and Brain Sciences*, 1–42. https://doi.org/10.1017/S0140525X25000032
Seth, A. K., & Bayne, T. (2022). Theories of consciousness. *Nature Reviews Neuroscience*, *23*(7), 439–452. https://doi.org/10.1038/s41583-022-00587-4
Shiller, D. (2024). Functionalism, integrity, and digital consciousness. *Synthese*, *203*(2), 47. https://doi.org/10.1007/s11229-023-04473-z
Shoemaker, S. (1982). The inverted spectrum. *The Journal of Philosophy*, *79*(7), 357–381. https://doi.org/10.2307/2026213
Thagard, P. (2022). Energy requirements undermine substrate independence and mind-body functionalism. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, *89*(1), 70–88. https://doi.org/10.1017/psa.2021.15
Tononi, G., & Koch, C. (2015). Consciousness: Here, there and everywhere? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1668), 20140167. https://doi.org/10.1098/rstb.2014.0167
Wiese, W. (2024). Artificial consciousness: A perspective from the free energy principle. *Philosophical Studies*, *181*, 1947–1970. https://doi.org/10.1007/s11098-024-02182-y