In Search of a Biological Crux for AI Consciousness

Bradford Saad (Global Priorities Institute, University of Oxford)

Global Priorities Institute | July 2024

GPI Working Paper No . 18-2024

Please cite this working paper as: Saad, B. In Search of a Biological Crux for AI Consciousness. *Global Priorities Institute Working Paper Series, No. 18-2024.* Available at https://globalprioritiesinstitute.org/in-search-of-a-biological-crux-for-ai-consciousness-bradford-saad

GLOBAL PRIORITIES INSTITUTE

UNIVERSITY OF OXFORD

In Search of a Biological Crux for AI Consciousness

Bradford Saad (Global Priorities Institute)

Abstract: Whether AI systems could be conscious is often thought to turn on whether consciousness is closely linked to biology. The rough thought is that 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 help organize inquiry into the topic. However, I argue, proposed views about the relationship between consciousness and biology tend not to capture a link that is crucial for the possibility of AI consciousness. In addition, I offer a crucial thesis, namely *the biological requirement* according to which being conscious at least nomically requires having biological states.

Key words: AI consciousness; machine consciousness; biological theories of consciousness; functionalism; substrate independence; reductionism; physicalism; dualism; identity theory; supervenience; phenomenal internalism; phenomenal externalism; Russellian monism; tracking theories; naive realism; interpretationism; neural correlates of consciousness

Word count: 4,488 (main text)

1. Introduction

Could an AI system be conscious? That is, can we in principle build an intelligent, non-biological machine that has experiences? Aside from its theoretical interest, this question is also practically urgent. It will soon bear on which AI systems are serious candidates for conscious subjects and moral patients. Already, philosophers are weighing in on the evaluation of AI systems for consciousness and policies concerning the treatment of systems that are deemed serious candidates.[1]

Whether AI systems could be conscious is often thought to turn on whether consciousness is closely linked to biology.[2] In broad strokes, the thought is that 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 since there are no biological barriers to it. Given this thought, a natural strategy for making progress on whether AI systems could be conscious is thus to examine evidence for and against close links between consciousness and biology. One obstacle to pursuing this strategy is that it's not clear what the relevant notion of a close link is. To overcome this obstacle, we might look to clear views in the literature that posit a close link between consciousness and biology. However, an obstacle to pursuing this strategy is that many discussions of the relationship between consciousness and biology are not focused on the phenomenal-biological links that are most crucial to the possibility of AI consciousness. For instance, when discussing close links between biology and consciousness or mental states more generally, philosophy of mind textbooks typically retain the historical focus on theories that identify experiences with brain states and challenges to such theories from functionalism and dualism.[3]

In what follows I'll explain why a wide range of views that posit close links between consciousness and biology do not come close to capturing the sort of candidate link that is most crucial for whether AI consciousness is possible. I'll then offer a thesis that captures a candidate link that is better poised to serve as such an evidential conduit. I call this thesis the 'biological requirement'. Roughly, the biological requirement says that, at least in our world,

---

[1] See, e.g., Birch (2024: Part V), Butlin et al. (2023: §1.2), Dung (2023), Long et al. (2024), and Sebo & Long (2023).

[2] See, e.g., Block (2007: 65, 249, passim; cf. 2023: 452-454), Godfrey-Smith (2023a; 2023b); Hill (1991: 224-227), and Seth (2021: Ch. 13); cf. Chalmers (2023: 18), Karnofsky (2022: 40, fn30), and McLaughlin (2019: 366).

[3] See Bayne (2022: Chs. 3, 9), Braddon-Mitchell & 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).

having experiences requires having biological states.[4] My goal here is not to defend the biological requirement. Rather, I aim to show that the biological requirement is a thesis concerning a close link between consciousness and biology that is especially crucial to whether AI systems could be conscious. One moral 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.

I should acknowledge that I do not take the argument of this paper to be particularly ambitious. I suspect that much of what I say will strike some readers as obvious once said. Nor should my argument be read as a blanket criticism of the existing literature for failing to focus on the biological requirement. There are often good pedagogical, historical, or dialectical reasons for focusing on other theses in the vicinity. Still, given the practical importance of improving our epistemic position with respect to AI consciousness, our 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 rather different 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 showing that the biological requirement is especially crucial to the possibility of AI consciousness and why other theses in the vicinity are not.

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 the relationship between consciousness and biology such that agreeing on its plausibility would settle the extent to evidence concerning a close link between consciousness

[4] 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 & Long (2023: §3.1) discuss a biological carbon-based substrate requirement alongside a biological function requirement

and biology confirms or disconfirms the possibility of AI consciousness. More precisely, a crucial thesis should be a 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 to 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.

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. However, there are two 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.[5] 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. There is also a second, perhaps deeper difficulty. Work in the science of consciousness is characteristically concerned with investigating what factors make a difference to consciousness within human subjects. But as

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

with 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. One reason that a perfect phenomenal-neural correlation in humans would not rule out the possibility of AI consciousness is: a perfect phenomenal-neural correlation in humans is compatible with a perfect phenomenal-neural-*functional* correlation in humans and with the correlated functional states rendering some AI systems conscious even in the absence of neural states.[6] Another reason 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.[7] 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 experience in our world is identical with some biological state (presumably a neural state, at least in humans).[8] 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.[9] 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. Non-reductive biological views come

---

[6] Cf. Block (2002), Birch (2022: 805), Chalmers (2010*a*: 99), and Papineau (2003).

[7] 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 (2010*b*: 348)), and then using the favored version to address the possibility of AI consciousness.

[8] See, e.g, Block (2009; 2023: Ch. 13), Brown (2012), Hill (1991: 10-12, *passim*), McLaughlin (2012), Polger (2011), Papineau (2021). Intriguingly, there seems to be a pattern among those with sympathies for biological reductionism of occasionally making concessions to the possibility of consciousness in non-biological systems—see, e.g., Block (2002; 2019: 376), Hill (1991: 224), Lewis (1980; 1994: 420), Papineau (2002: §7.2); cf. Kim (2005: 155, 159, 168-173) and Godfrey-Smith (2024: 1665).

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.

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

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.[10] But that is not the kind of possibility here at issue: the interesting and practically urgent question about the possibility of AI consciousness 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 with respect to this candidate nomic possibility. Upon attending to the fact that this is a relevant candidate possibility, it should be clear that each of these biological views would eliminate the possibility of AI consciousness and hence that one need not be a biological reductionist in order to reject the possibility of AI consciousness because one thinks there is a close link between consciousness and biology.[11]

Let's call the hypothesis that every experience has a biological base *biological supervenience*.[12] 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

[10] 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, e.g., Jackson (1998:12) and Lewis (1983: 362).

[11] The situation here echoes Chalmers's (1996) 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.

[12] Cf. Kim (2005: 93).

non-biological.[13] 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* claim that experiences' bases include both internal and external physical factors. On these views, internal physical duplicates might 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.[14] *Naive realists* hold that perceptions of such qualities figure in the bases of veridical perceptual experiences.[15] *Embodiment theories* claim that bodily features (such as sensory-motor dispositions) figure in experiences' bases.[16] 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 naive realists and tracking theorists to opt for biological-evolutionary accounts of perception and tracking on which perceiving or 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.[17] Likewise, 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 are compulsory. But their availability yields hybrid versions of naive realism, tracking theories, and embodiment theories.
- *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.[18] 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.[19] More generally, we may be able to extract purely functionalist versions of scientific theories by abstracting away from their neural commitments (through 'Ramseyfication')[20] as well as hybrid versions by retaining their biological commitments or by leveraging their functional commitments into designators for neural states that play certain functional roles and which are

[13] For relevant discussion, see Block (2007: 340), Braddon-Mitchell & Jackson (2007: 133-4), Chalmers (2010*b*: "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"), Lewis (1980), Lee (2016: §IV), Pautz (2010*b*: 358-9), Prinz (2012: Ch. 9), and Nagel (2000).

[14] For an overview of different tracking theories, see Dalbey & Saad (2022).

[15] For recent work on naive realism, see essays in Beck & Masrour (forthcoming).

[16] See, e.g., O'Regan & Noë, A. (2001)

[17] For a tracking theory along these lines, see Dretske (1995).

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

[19] Or at least this is so on some formulations of the theory (e.g. see Seth & Bayne (2022: 441)). However, Stanislas Dehaene, one of the main developers of the theory, seems open to AI consciousness (Dehaene, 2017: 79).

[20] See Block (2007: 30-32) and Lewis (1970).

(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.[21] 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.[22] Russellian monists hold that experiences are at least partly identical with or grounded in quiddities.[23] 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.[24] 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.[25] Interpretationist views are motivated by their potential to explain harmonious correlations between experiences and accompanying states (e.g. correlations between experiences and accompanying states that systematically lend 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

[21] Cf. Kim (2005: 171-173) and Shoemaker (1982); cf. Lee (2017: 216).
[22] Whether quiddities count as physical is a merely verbal issue. For convenience, I stipulate that they do.
[23] For an overview, see Alter & Pereboom (2023).
[24] See,e.g., Chalmers (2017).
[25] For discussion of axiarchic views and consciousness, see Cutter (2023), Cutter & Saad (2023: §6, 20), Goff (2018: 117, 120). For discussion of interpretationist views of consciousness, see Cutter & Crummett (forthcoming), Saad (forthcoming); cf. Pautz (2010*a*: §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—see Saad (forthcoming). In any event, standard (non-interpretationist) forms of functionalism about experience enjoy different motivations than interpretationist theories.

biological commitments that eliminate the possibility of AI consciousness.[26] 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 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). 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.[27] 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. While functionalist views are sometimes compared with physicalist forms of biological reductionism, this is often in contexts (such as philosophy of mind textbooks) where historical precedent and/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 that impose the biological requirement 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

[26] Or at least perceptual AI consciousness, as on naive realism and tracking theories.
[27] But see Godfrey-Smith (2016; 2023*a*; 2023*b*); cf. 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).

dismissal of such theories that are relevant to AI consciousness,[28] as such dismissals could lead to errors in decisions about what 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. As Chalmers (2022: 93) formulates this thesis, it 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."[29] 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 does not specify the (absence of) a crucial phenomenal-biological link.

---

[28] 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 & Long (2023). 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.

[29] Chalmers (2022: 93) attributes substrate independence to Bostrom (2003), which seems to be the origin of the thesis. However, Bostrom's remarks suggest at least four different formulations of substrate independence.

> [1] [M]ental states can supervene on any of a broad class of physical substrates.
> [2] Provided a system implements the right sort of computational structures and processes, it can be associated with conscious experiences.
> [3] Neurotransmitters, nerve growth factors and other chemicals affect subjective experience only via their direct or indirect influence on computational activities.
> [4] It is not an essential property of consciousness that it is implemented on carbon-based biological neural networks inside a cranium: silicon-based processors in a computer could in principle do the trick as well. (Bostrom, 2003: 244-5; numerals added for ease of reference)

Are any of these claims apt to serve as the crucial thesis? No. Briefly, it is unclear how to interpret [1]: it cannot plausibly be read in terms of a standard notion of supervenience (on which variation in a supervenient class requires variation in a base class); but an alternative notion of supervenience is not provided. In any case, [1] concerns mental states and physical substrates in general, not close links between consciousness in particular and biology. Although [2] concerns consciousness in particular, strictly speaking it is silent on whether the requisite computational processes could be implemented in a non-biological system. Granting that they can, [2] is then in effect a very specific (computationalist) hypothesis about how consciousness could fail to be closely linked by biology and is hence inapt to qualify as a crucial thesis since it fails to capture many other ways that link could fail—the same goes for [3]. In addition, [3] concerns how biological factors can affect *which* experiences an individual has, not just how those factors can influence *whether* an individual is conscious—but only the latter is at issue when we ask whether an AI system could be conscious. Finally, although [4] concerns the (absence of) a close link between consciousness and biology, it also characterizes that link in terms of details that are not crucial.

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 fairs 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.[30] 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 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. *The biological requirement* says that having an experience requires having a biological state.[31] 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

[30] Chalmers (2010*a*: 23-25, 139, fn36) seems open to this possibility.

[31] 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 (forthcoming).

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 purely 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 found only in kinds of AI systems that will become available more than ten years from now. Finally, the biological requirement is clear and exceedingly simple and so 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. Even granting that there could be evidence that supports attributions of consciousness to particles, corporations, stars, or the universe, one 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—meaning 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 capture 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. Yet the biological requirement is potentially sensitive to evidence *for* a close link between consciousness and biology that *confirms* the possibility of AI consciousness and to evidence *against* a close link between consciousness and biology that *disconfirms* the possibility of AI consciousness. In short, 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 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.[32] This evidence could *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.[33] 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 captures the target thought would be insensitive to these discoveries.

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 an alternative that is a better candidate for a crucial thesis. A better suggestion, I think, is to adopt the biological requirement as a default crucial thesis in discussions of how close links between consciousness and biology bear on the possibility of AI consciousness and then modify the biological requirement as appropriate in context. The

[32] Cf. Lamme (2010).
[33] Cf. Tononi & Koch (2015).

moral I draw is that the biological requirement is well-suited to serve as a crucial thesis in guiding investigations into the relationship between consciousness and biology and the possibility of AI consciousness.

References

Alter, T. and D. Pereboom, "Russellian Monism", (2023) *The Stanford Encyclopedia of Philosophy* E.N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/fall2023/entries/russellian-monism/>.

Bayne, T. (2022). *Philosophy of mind : an introduction*. Routledge.

Beck, Ori & Masrour, Farid (eds.) (forthcoming). *The Relational View of Perception: New Essays*. Routledge.

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

————(2007). Consciousness, Function, and Representation: Collected Papers. Bradford.

————(2009). Comparing the major theories of consciousness. In Michael Gazzaniga (ed.), *The Cognitive Neurosciences IV*. pp. 1111-1123.

————(2019) Functional Role, Superficialism, and Commander Data: Reply to Brian McLaughlin. In A. Pautz & D. Stoljar (eds) *Blockheads!: Essays on Ned Block's Philosophy of Mind and Consciousness*, 335.

————(2023). *The border between seeing and thinking*. OUP.

Birch, J. (2022). Materialism and the moral status of animals. *The Philosophical Quarterly*, 72(4)795-815.

————(2024) *The edge of sentience: risk and precaution in humans, other animals, and AI.* OUP.

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

Braddon-Mitchell, D., & Jackson, F. (2007). *The philosophy of mind and cognition* (2nd edition). Blackwell.

Butlin P. et al. (2023) Consciousness in artificial intelligence: insights from the science of consciousness. arXiv

Brown, R. (2012). The brain and its states. In *Being in time: Dynamical models of phenomenal experience,* 88:211-38.

Campbell, J. (2010) "Demonstrative Reference, the Relational View of Experience and the Proximality Principle." In R. Jeshion (ed.) *New Essays on Singular Thought*, 193–212. OUP.

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

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

————(2010*a*). *The Character of Consciousness*. OUP.

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

————(2017). The combination problem for panpsychism. In *Panpsychism: contemporary perspectives*.

————(2022). *Reality+: Virtual worlds and the problems of philosophy*. Penguin UK.

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

Cutter, B. (2023). From moral realism to axiarchism. URL:https://philarchive.org/archive/CUTFMR

————(forthcoming). The AI Ensoulment Hypothesis. *Faith and Philosophy.*

Cutter, B. & D. Crummett (2022). Psychophysical Harmony: A New Argument for Theism. *Oxford Studies in Philosophy of Religion*.

Cutter, B., & Saad, B. (2024). The problem of nomological harmony. *Noûs*, 58(2)482-504.

Dehaene, S. What Is Consciousness, and Could Machines Have It?. Power and Limits of Artificial Intelligence, 75.

Dennett, D. (2001). Are we explaining consciousness yet? *Cognition* 79(1–2): 221–237.

Dung, L. (2023). How to deal with risks of AI suffering. *Inquiry*, 1-29.

Godfrey-Smith, P. (2016). Mind, matter, and metabolism. *The Journal of Philosophy*, *113*(10), 481-506.

————(2023*a*) "Nervous Systems, Functionalism, and Artificial Minds" URL:https://petergodfreysmith.com/wp-content/uploads/2023/12/NYU-Oct-2023-Animals-AI-Functionalism-paper-Post-C3.pdf

————(2023*b*) Simulation Scenarios and Philosophy. URL:https://petergodfreysmith.com/wp-content/uploads/2023/09/Simulations-etc-PGS-2023-N-Dst.pdf

————(2024). Inferring consciousness in phylogenetically distant organisms. *Journal of Cognitive Neuroscience*, 1-7.

Goff, P. (2018). Conscious thought and the cognitive fine-tuning problem. *The Philosophical Quarterly*, 68(270)98-122.

Heil, J. (2013). *Philosophy of mind : a contemporary introduction* (3rd ed.). Routledge.

Hill, C.S. (1991). *Sensations: A defense of type materialism*. Cambridge.

Jackson, F. (1998). *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. OUP.

Karnofsky, H. (2022) URL:https://www.cold-takes.com/assets/files/most-important-century-consolidated.pdf

Kim, J. (2005) *Physicalism or Something Near Enough*. Princeton University Press.

Kind, A. (2020). *Philosophy of mind* (1st ed.). Routledge.

Lamme, V. A. (2010). How neuroscience will change our view on consciousness. *Cognitive neuroscience*, 1(3)204-220.

Lee, G. (2016). Does experience have phenomenal properties?. *Philosophical Topics*, 44(2), 201-230.

Lewis, D. (1970). How to define theoretical terms. *The Journal of Philosophy*, 67(13), 427-446.
————(1980). Mad pain and Martian pain. In *The Language and Thought Series* (pp. 216-222). Harvard University Press.
————(1983). New work for a theory of universals. *Australasian Journal of Philosophy* 61(December):343-377.
————(1994). "Reduction of Mind." In S. Guttenplan (ed.), Companion to the Philosophy of Mind. Oxford: Blackwell.
Lycan, W. (1981) Form, function and feel. *Journal of Philosophy* 78: 24–50.
Mandik, P. (2023). *This is philosophy of mind : an introduction* (Second edition.). Wiley Blackwell.
McLaughlin, B.P. (2012). On justifying neurobiologicalism for consciousness. *New Perspectives on Type Identity: The Mental and the Physical*.
McLaughlin, B. P. (2019). Could an Android Be Sentient?. A. Pautz & D. Stoljar (eds) *Blockheads!: Essays on Ned Block's Philosophy of Mind and Consciousness*, 335.
Morales, J. & Lau, H. (2020). The Neural Correlates of Consciousness. In U. Kriegel (ed.), *The Oxford Handbook of the Philosophy of Consciousness*. OUP.
Nagel (2000) "The Psycho-physical Nexus", in P. Boghossian and C. Peacocke (eds.). *New Essays on the A Priori*. OUP.
O'Regan, J.K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and brain sciences*, 24(5)939-973.
Papineau, D. (2002). *Thinking about consciousness*. Clarendon.
Papineau, D. (2003). Could there be a science of consciousness?. *Philosophical Issues*, 13, 205-220.
Papineau, D. (2021). *The metaphysics of sensory experience*. OUP..
Pautz, A. (2010*a*). A Simple View of Consciousness. In R. Koons & G. Bealer (eds.), *The Waning of Materialism*. OUP 25--66.
Pautz, A. (2010*b*). Do theories of consciousness rest on a mistake?. *Philosophical Issues*, 20, 333-367.
Pautz, A. (2017). The significance argument for the irreducibility of consciousness. *Philosophical Perspectives*, 31, 349-407.
Polger, T. W. (2011). Are sensations still brain processes?. *Philosophical Psychology*, 24(1)1-21.
Prinz, J. (2012). *The Conscious Brain*. OUP.
Saad, B. (forthcoming). Lessons from the Void: What Boltzmann Brains Teach. *Analytic Philosophy*.
Searle, J., (1997), *The Mystery of Consciousness*, New York: The New York Review of Books.
Seth, A. (2021). *Being you: A new science of consciousness*. Penguin.
Seth, A. K., & Bayne, T. (2022). Theories of consciousness. *Nature Reviews Neuroscience,* 23(7), 439-452.
Sebo, J., & Long, R. (2023). Moral consideration for AI systems by 2030. *AI and Ethics*, 1-16.
Shoemaker, S. (1982). The inverted spectrum. *The Journal of Philosophy*, *79*(7), 357-381.
Shiller, D. (2024). Functionalism, integrity, and digital consciousness. *Synthese* 203(2):1-20.
Tononi, G., & Koch, C. (2015). Consciousness: here, there and everywhere?. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1668), 20140167.