This draft is subject to change. Feedback is welcome!

Consciousness without biology: An argument from anticipating scientific progress

Leonard Dung (leonard.dung@rub.de)

Abstract

I develop the *anticipatory argument* for the view that it is nomologically possible that some non-biological creatures are phenomenally conscious, including conventional, silicon-based AI systems. This argument rests on the general idea that we should make our beliefs conform to the outcomes of an ideal scientific process and that such an ideal scientific process would attribute consciousness to some possible AI systems. This kind of ideal scientific process is an ideal application of the *iterative natural kind (INK) strategy,* according to which one should investigate consciousness by treating it as a natural kind which iteratively explains observable patterns and correlations between potentially consciousness-relevant features. The relevant AI systems are *psychological duplicates*. These are hypothetical non-biological creatures which share the coarse-grained functional organization of humans. I argue that an ideal application of the INK strategy would attribute consciousness to psychological duplicates because this gives rise to a simpler and more unifying explanatory account of biological and non-biological cognition. If my argument is sound, then creatures made from the same material as conventional AI systems can be conscious, thus removing one of the main uncertainties for assessing AI consciousness and suggesting that AI consciousness may be a serious near-term concern.

1. Introduction

Due to progress in AI, the question when artificial systems might be phenomenally conscious[1] has received increasing attention and scrutiny. Researchers examine what theories of consciousness imply for the distribution of artificial consciousness (Butlin et al. 2023), what empirical tests for artificial consciousness might be (Dung 2023; Elamrani and Yampolskiy 2019; Perez and Long 2023), and what the ethical significance of questions of artificial consciousness is (Birch 2024; Ladak 2023; Saad and Bradley 2022).

A central point of contention is whether consciousness requires biological processes (Saad 2024, see also footnote 2 in that paper). On *biological substrate views*, creatures can only be conscious if they possess a biological, carbon-based substrate (for this distinction, see Sebo

[1] A creature is phenomenally conscious if and only if there is something it is like to be it, it has subjective experience, and it has a first-person point of view. These three conditions are equivalent. For brevity, I will omit the qualifier "phenomenal" henceforth.

and Long 2023). That is, implementing the right functions is not sufficient for consciousness; instead, conscious creatures need to be made out of the right kind of material. An example is a view which posits that consciousness is type-identical to a certain brain state (Place 1956; Smart 1959), and type-individuates brain states partly in terms of the biological material they are composed from. According to the *biological function view*, realizing certain fine-grained biological functions is necessary for consciousness (Cao 2022; Godfrey-Smith 2016, 2020; Seth 2021). These functions could be having a metabolism, system-wide synchronization or oscillation properties, or other functions whose implementation depends on specifics of the physical features of neurons or brain biochemistry. Some other views tie consciousness not explicitly to biological processes, but nevertheless posit constraints on the substrates or functions of consciousness which conventional silicon-based computing systems cannot fulfill, while biological organisms can (Piccinini 2021; Shiller 2024; Wiese 2024; Wiese and Friston 2021). I will call all these positions "biological views of consciousness". I will call creatures "non-biological" when they do not satisfy the necessary conditions on consciousness these views posit because they are not made from biological (or other suitable) material.

The divide between biological and non-biological views seems to be the most important determinant for researchers' views on the prospects of artificial consciousness. On biological views, artificial consciousness is probably far in the future, if at all possible. Moreover, if conscious artificial systems are developed, they will not conform to currently dominant AI paradigms. Neural organoids (Birch 2023) and molecular computers (Brunet and Halina 2020) are better candidates. By contrast, Butlin et al. (2023) base their report on computationalism about consciousness which entails that biology is not necessary for consciousness. Accordingly, their analysis concludes that, while no AI systems of their time of writing are conscious, "there are no obvious barriers to building conscious AI systems" (ibid.). Many authors even argue that, on computationalism, near-term AI consciousness is a serious possibility (Chalmers 2023; Goldstein and Kirk-Giannini n.d.; Long et al. 2024; Sebo and Long 2023).

In this paper, I will develop a new argument for the view that AI consciousness is nomologically possible and against biological views. The argument rests on the general idea that we should make our beliefs conform to the outcomes of an ideal scientific process and that such an ideal scientific process would attribute consciousness to some possible (non-biological) AI systems. *Psychological duplicates* are these kinds of AI systems, and an ideal application of the *iterative natural kind (INK) strategy* is that kind of ideal scientific process.

So, in section 2, I explain the INK strategy as an approach to the scientific investigation of consciousness. In section 3, I elaborate on the notion of psychological duplicates and argue

that they are possible. This paves the way for the core argument of this paper (in section 4): the anticipatory argument from consciousness science. I argue that we can foresee that an ideal scientific process, based on the iterative natural kind strategy, converges on the view that non-biological AI consciousness is possible. Section 5 replies to the objection that my argument may be question-begging. Section 6 generalizes the argument of section 4: Approaches to consciousness science beyond the iterative natural kind strategy also support the view that biological views of consciousness are false.

2. The iterative natural kind strategy in consciousness science

The iterative natural kind (INK) strategy consists in treating consciousness as a natural kind which iteratively explains observable patterns and correlations between potentially consciousness-relevant features (Bayne et al. 2024; Bayne and Shea 2020; Birch 2022; Boyle forthcoming; Mckilliam forthcoming; Shea 2012). Grouping entities in terms of natural kinds means classifying them according to underlying, not socially constructed, similarities and differences in their natures (Bird and Tobin 2023). According to this strategy, researchers should search for clusters of effects and capacities which may be related to consciousness and iteratively apply inference to the best explanation to confirm the underlying causal processes and mechanisms.

To find such clusters, researchers need to agree on some measures of consciousness, without being able to already presuppose a comprehensive theory of consciousness (since such measures are necessary to support such a theory in the first place). Such measures have to be established based on pre-theoretical principles (Michel 2023). Such principles include (principle 1) "[t]he better one sees something, the more likely one is to be conscious of it" (ibid., p. 837) and (principle 2) "people can usually tell whether they are conscious of something or not" (ibid., p. 838). The former explains why perceptual discrimination is a pre-theoretical measure of consciousness with some degree of epistemic standing, the latter does the same for metacognitive verbal report (e.g. "I saw the stimulus"). These can then be used to evaluate and improve ("calibrate") further measures of consciousness, i.e., features which – as measured by the preceding criteria – are robustly caused by conscious experience. For example, Birch (2022) proposes trace conditioning, reversal learning, and cross-modal learning as measures of consciousness in animals, since they seem to be facilitated by consciousness in humans, when the latter is measured by verbal report.

Importantly, Michel (2023) explains how pre-theoretical principles can be used to calibrate, including possibly to revise, our initial measures of consciousness, including verbal

report. For instance, based on principle 1 (and other background information) we can identify cases where subjects are very likely conscious of a stimulus. If a subject verbally reports that it is not conscious of this stimulus, and particularly if we have additional reason to distrust a subject's report (for example, damage to brain areas involved in metacognitive judgement, a neural response which is similar to the neural response to conscious stimuli, or the presence of a wide range of other potentially consciousness-relevant abilities), then this may point to cases where verbal report is not a valid measure of consciousness (Mckilliam forthcoming).

Once we have a large variety of measures of consciousness, they can be used to calibrate each other. Moreover, we can examine how they covary. If two measures are both measures of consciousness, their results should agree. So, if many of them cluster, this is evidence that there is an underlying natural kind – consciousness – which explains such clustering. This process can also weaken our confidence that certain features are measures of consciousness: for example, if they don't cluster with other measures we are confident in or if our growing theoretical understanding of the consciousness kind suggests that these are the wrong types of features to measure consciousness.[2] This process of mutual correction is iterative, it applies to all our measures and models of consciousness, and it is driven by considerations "of theoretical unification, simplicity, explanatory power, and predictive success" (Bayne et al. 2024).

The INK strategy has been prominently advocated for and defended as a methodology for consciousness science. It is plausible that this strategy describes how consciousness should be researched, since it seems – on a high level of abstraction – to describe the standard approach for scientific investigation, at least in domains which deal with natural kinds (Chang 2004). While it is not guaranteed that consciousness turns out to be a natural kind,[3] the INK strategy has the resources to test whether it is one.

There are disagreements between proponents of the INK strategy, particularly with respect to the measures or principles which form our initial starting point to get the process of calibration of measures going (see Mckilliam forthcoming, section 4 for a brief overview). However, the basic guidelines of the strategy are agreed-upon: 1. Use pre-theoretical principles to establish an initial set of measures of consciousness with at least minimal epistemic standing. 2. Use these measures to calibrate each other and to establish and calibrate further "derived"

[2] This is related to *model calibration* which calibrates a measure by developing a better model of how the measure works and what it would output in certain situations if it was accurate (Michel 2023, p. 834).

[3] The features which are pre-theoretically linked to consciousness may, for example, form a heterogenous array, which is not bound together by a few underlying causal mechanisms. However, this arguably seems relatively unlikely, since consciousness science has already discovered many robust correlations between consciousness and other (behavioral, psychological, as well as neural) features. Eliminativism about consciousness (Frankish 2016; Irvine and Sprevak 2020) – the view that consciousness does not exist – would also imply that there is no natural kind of consciousness.

measures of consciousness. 3. Use these measures to search for clusters of consciousness-relevant features. 4. Iteratively apply inference to the best explanation to confirm the underlying natural kind which is then identified with consciousness as well as to further calibrate proposed measures.

When applied to non-human consciousness, it has been suggested to perform the INK strategy hierarchically (Bayne et al. 2024): When validating measures of consciousness, we start with neurotypical, adult humans as a population where consciousness is agreed-upon (step 1). In step 2, we see whether different measures also correlate in a new population (for example, human infants or non-human mammals), including new measures that are only applicable to that population. Then, we increase our credences in measures, including the ones from step 1, where correlations are preserved and decrease it in others. Subsequently, we again apply the measures to new populations, working our way downwards from the human case where there is consensus about the presence of consciousness until we reach hotly contested cases (e.g. invertebrates or AI).

Before we can turn to my core argument (section 4), section 3 introduces the notion of psychological duplicates that this argument relies on.

3. Psychological duplication

3.1 What are psychological duplicates?

Henceforth, when not specified otherwise, "possible", and synonymous notions, mean "nomologically possible": compatible with the laws of nature.[4] When not specified otherwise, "AI system" refers to conventional, silicon-based and non-biological systems, not AI systems made from unusual material (e.g. morphological computers). In this section, I want to make the following assumption plausible.

Psychological Duplication: Psychological duplicates are possible.

Let me stipulate that psychological duplicates are non-biological creatures which are made from the same material as conventional AI systems and share all coarse-grained functional as well as behavioral properties of normal conscious humans.

So, psychological duplicates are behaviorally equivalent to normal conscious humans: they exhibit the same behavior in all possible situations. Coarse-grained functional equivalence is harder to spell out. Functional properties are individuated by their causal role, i.e. by input-

[4] For reasons why nomological possibility is most relevant in the context of debates on the possibility of AI consciousness, see Saad (2024).

output mappings. The functional properties of the human brain can be characterized on various levels of abstractions. On a very fine-grained level, it is possible to talk about the causal roles played by individual neurons, their sub-neural components, and brain biochemistry. On such a fine-grained level, it may not be nomologically possible (Cao 2022) that conventional computer hardware duplicates the functional properties of brains.

An example of a coarse-grained functional level is the algorithmic level of description (Marr 1982). Here, the brain can be understood in terms of the high-level tasks, characterized as mathematical functions, it solves and the algorithms, operating over representations, it implements to solve these tasks.[5] So, psychological duplicates can be said to be type-identical to humans on the algorithmic level.

While the algorithmic level is one example for a coarse-grained functional description of the brain, there may be other levels of functional abstraction which are coarse-grained enough that duplication of the functioning of the human brain may be possible on that level, and which are fine-grained enough that this duplication – in an interesting sense – duplicates human "psychology". Overall, psychological duplicates are AI systems that mirror the coarse-grained functional organization, i.e. abstractly described causal processes, of humans, while the exact fineness of grain on which this duplication operates is up for debate.

Coarse-grained functional equivalence implies behavioral equivalence: if a system implements the same functions as the human mind, then it produces the same outputs, given the same inputs. Moreover, the same behavior can be produced by a wide variety of mechanisms. For these reasons, it is relatively uncontroversial that AI systems which are behaviorally equivalent to humans are (nomologically) possible.[6]

3.2 Why should we think that psychological duplicates are possible?

Let me give three reasons in favor of Psychological Duplication. Minimally, it is important to notice that Psychological Duplication is a weaker assumption than the assumption that biological views of consciousness are false. Many biological views are best interpreted as implying that psychological duplicates are not conscious, not as claiming that they are impossible. This is true, for instance, of views which identify consciousness with some brain process (e.g. Smart 1959) or posit constraints on implementing consciousness that silicon-based systems cannot meet (e.g. Shiller 2024). For some other researchers, it is not obvious whether

[5] Marr (1982) calls the former the "computational level".

[6] Behavioral duplicates might not be possible, if one individuates behavior very finely, e.g. counting very precise patterns of physiological responses as types of behavior. However, the notion of behavior relevant for my argument is coarse-grained.

their view entails that psychological duplicates are impossible, that they are not conscious, or both (e.g. Godfrey-Smith 2020).

So, the debate between biological and non-biological views has mainly been about whether coarse-grained functional duplicates of humans are conscious or whether something is missing: a biological substrate, fine-grained biological functions, or some other property which cannot be possessed by systems made from silicon-based material. For this reason, proponents of biological views need additional arguments to motivate the view that psychological duplicates are not possible. Specifically, proponents of biological views should endorse Psychological Duplication if they think consciousness' dependence on particular physical substrates is something distinctive of consciousness, not shared by other mental states.

Second, on the algorithmic level, it is plausible that psychological duplicates are possible. It is common in cognitive science to describe humans as storing representations and processing them using algorithms which are characterized independently of neural properties and sufficiently abstractly that they can also be implemented by AI systems.

Third, Psychological Duplication is also supported by views which hold that other types of mental states can be possessed by AI – like beliefs (Goldstein and Kirk-Giannini forthcoming) or concepts (Butlin 2021) – or that computational models implement key components of cognitive processes such as episodic memory (Boyle and Blomkvist forthcoming) or attention (Lindsay 2020). For, if these views are true, human and AI cognition must have coarse-grained functional correspondences sufficient to share (key components of) central cognitive states.

My three reasons do not entail that Psychological Duplication can be established conclusively. For example, Cao (2022) appears skeptical of Psychological Duplication. She grants that cognitive scientists and philosophers often assume that there is a functional level (like the algorithmic level) which is, in terms of abstractness, "below behavior and above biology" (ibid., p. 506). Yet she goes on (ibid.):

> But we have no theoretical guarantee that any such efficient, intermediate level of description exists; perhaps there are no real patterns at the sub-personal level that are well above the level of biological description.

If there are no patterns relevant to explaining human psychology which are more coarse-grained than biological levels of description, then psychological duplicates are not possible: there are no relevant coarse-grained functional structures to duplicate. While I cannot rule out this view, it seems less likely to me than Psychological Duplication. In nature, patterns typically exist at a variety of levels of description, and the success of cognitive science supports the view that there are relevant patterns at a coarse-grained functional level.

3.3 Alternatives to assuming Psychological Duplication

In what follows, I will assume Psychological Duplication. However, the main argument of this paper is also of interest for researchers who reject Psychological Duplication. First, they can read my conclusion as conditional: If Psychological Duplication is true, then AI consciousness is possible. This insight informs how biological views of consciousness are best developed in the future, namely that they need to focus on explaining how they reject Psychological Duplication. Second, one may try to develop the argument of this paper with weaker assumptions. Candidates are:

Behavioral Duplication: AI systems which are *behaviorally* (but not necessarily coarse-grained functionally) equivalent to humans are possible.

Psychological Similarity: AI systems which are coarse-grained *functionally* similar (but not equivalent) to humans are possible.

One of these weaker assumptions may be sufficient to support my subsequent argument, although this is not obvious.

Third, one could aim to establish a more modest conclusion. Consider:

Non-biological Duplication: Creatures without biology (which do not need to be AI systems) which are coarse-grained functionally equivalent to humans are possible.

If, except for assuming Psychological Duplication, my subsequent argument is sound, then Non-biological Duplication – a weaker assumption than Psychological Duplication – can be used to show that consciousness without biology is possible, without entailing that AI consciousness is possible. In other words, the resulting argument would still establish that consciousness is possible without *biology* specifically (e.g., in non-biological extraterrestrials), although it would not follow that systems made from the same material as conventional AI systems can be conscious.

Now, let us turn to the main argument.

4. The anticipatory argument from consciousness science

4.1 The argument and premise 1

Here is the argument:

P1. If the outcome of an ideal application of the INK strategy is that some possible AI systems are conscious, then some possible AI systems are conscious.

P2. The outcome of an ideal application of the INK strategy is that some possible AI systems are conscious.

C. Some (nomologically) possible (non-biological) AI systems are conscious.

P1 expresses the following: If an ideal use of the INK strategy converges on the result that p, then it is the case that p – at least with respect to the distribution of AI consciousness. What is an ideal application of the INK strategy? We have already outlined the INK strategy – roughly, treating consciousness as a natural kind to iteratively explain patterns and correlations of potentially consciousness-relevant features. An *ideal* application is a fictional application of the INK strategy where researchers have unlimited resources (including time), are perfect reasoners, and are immune to performance errors. So, we imagine researchers to do all relevant experiments, data analyses, and so on, while flawlessly constructing interpretations of these experiments and theories which skillfully explain these results. Then, they iterate by constructing improved measures, experiments, theories and so forth, until no further information or scientific insight can be gained that way. Given this idealization, the following conditional holds: If the INK strategy *can* tell us which creatures are conscious, then the *ideal* INK strategy gives us the correct result (such that P1 holds).

Of course, in principle, someone could object that the INK strategy itself is flawed. However, as we have seen, the INK strategy is simply a high-level description of the process science generally employs to investigate putative natural kinds. It is unclear how else science could find determinate answers to questions about consciousness and comparative cognition (Boyle forthcoming) than by asking whether members of the same natural kind are present across situations and creatures.[7]

If we take a scientific perspective on the question of AI consciousness, then we should adopt whatever view is the outcome of an ideal scientific investigation of this question. It does seem that we should take a scientific perspective to the question whether AI consciousness is possible. A reason is that it is unclear how metaphysics, or some other non-scientific domain of investigation, could settle this question. In general, the question what possible physical realizers of certain states are – i.e. whether consciousness can be realized in non-biological processes – appears like a scientific question. It lacks the degree of generality or fundamentality often associated with metaphysical issues.[8]

[7] Also, my argument aims to be compatible with approaches to consciousness science beyond the INK strategy (see section 6).

[8] I only assume here that the question which physical material consciousness can (nomologically) be realized in is one where we should defer to science. I am not assuming that all questions about consciousness can or should be settled by science.

If science can give us an answer to the question whether AI consciousness is possible and the appropriate scientific methodology is captured by the INK strategy, then P1 is true. The only alternative is to hold that nothing can give us an answer to the question whether consciousness requires biology. In this case, P1 is false and the anticipatory argument fails. So, I am assuming here that it can be found out in principle whether AI consciousness is possible.

However, even if questions about AI consciousness are metaphysically indeterminate or forever epistemically outside our reach, it might – even under this assumption – still be relevant what the outcome of an ideal application of the INK strategy consists in. It could be that this outcome would be similarly indeterminate. However, if it can be shown that an ideal application of the INK strategy delivers a determinate outcome, that must (under this assumption) be because non-epistemic considerations guide the process in a determinate direction. For instance, considerations of theoretical unification, simplicity, explanatory power, and predictive success (Bayne et al. 2024) might favor certain views about the distribution of consciousness, even if – by assumption – these values cannot be construed as truth-tracking in this case. Whether such values favor attributions of AI consciousness would still be pragmatically relevant, even if there is no fact of the matter about whether certain AI systems are conscious.

4.2 Premise 2

So, P1 says that the outcome of the ideal application of the INK strategy matches what is true about the distribution of consciousness. P2 expresses that the outcome of the ideal application of the INK strategy would be that AI consciousness is possible. If both premises are true, then biological views are false.

Why believe P2? Again, we can focus on psychological duplicates as test cases of AI consciousness. Since we stipulated that psychological duplicates are silicon-based coarse-grained functional and behavioral equivalents to humans, we already have knowledge about the features of psychological duplicates that the INK strategy will discover. If consciousness is a natural kind, then an ideal application of the INK strategy will find rich clusters of cognitive capacities and patterns of cognitive and behavioral effects (Taylor et al. 2022) shared by humans, other animals (insofar as they are conscious), and psychological duplicates. These clusters will also manifest across phylogeny, ontogeny, and situations. Moreover, there will be shared computational mechanisms which explain these clusters.

At the same time, some biological properties which are part of this cluster of capacities, effects, and mechanisms in human and non-human animals are absent in psychological duplicates. This includes, for instance, the presence of a neocortex or a biological structure with

a similar functional role (Stacho et al. 2020), or of certain global oscillations in the brain (Godfrey-Smith 2020). I hold that, in this case, an ideal application of the INK strategy would identify consciousness with the mechanisms underlying the shared, coarse-grained functional properties, rather than the fine-grained biological properties. The reason is that the theoretical virtues driving the INK strategy militate in favor of the former option.

First, positing consciousness as a coarse-grained functional kind which explains clusters of phenomena in biological and non-biological creatures provides *theoretical unification*. It posits the same kind of process – consciousness – to explain phenomena related to biological and non-biological creatures. By being unified, the resulting account is *simple*. It provides a single kind of explanation – consciousness – for clusters of cognitive effects and capacities in all biological and non-biological creatures. For this reason, the non-biological view also has a lot of *explanatory power*. Assuming non-biological consciousness licenses explanations of behavior across biological and non-biological creatures. This is why the non-biological view also gives consciousness *predictive power*: By positing consciousness, it can predict the appearance of parts of the cluster of effects and capacities in biological and non-biological creatures alike.

Biological views are inferior with respect to these virtues. They cannot use consciousness to give a unified and simple explanation for shared clusters of capacities and effects in biological and non-biological species. They have two options. First, they can posit two mechanisms: a biological one which explains the clusters of capacities and effects, including the biological ones, we find in biological creatures, plus a non-biological one which explains the same phenomena, excluding the biological ones, in non-biological creatures. The biological mechanism would then be identified with consciousness. However, positing two distinct mechanisms is obviously less unified, less simple, less explanatorily powerful, and less predictive than the non-biological explanation. There would be no explanation for why all coarse-grained functional and behavioral properties are shared between biological and non-biological creatures. Explaining this requires positing a shared mechanism, which can then be identified with consciousness.

Second, biological views could grant that there is a shared, coarse-grained functional mechanism which is causally responsible for these shared clusters in biological and non-biological creatures. However, they could argue that consciousness is distinct from this shared mechanism, and that consciousness depends on the specific properties that we only find in biological creatures.

However, this move is inconsistent with the INK strategy. According to the INK strategy, consciousness should be identified with whatever natural kind underlies and explains the clusters of phenomena that our different putative measures of consciousness target. When making this objection, proponents of biological views concede that all properties that humans and psychological duplicates share are explained via the same coarse-grained functional mechanism. These shared properties include all behavior as well as consciousness' effects on all cognitive capacities, including learning, reasoning, and attention.[9] So, most of the phenomena targeted by putative measures of consciousness are part of this cluster. Given this, the INK strategy is committed to identifying the natural kind which underlies this cluster with consciousness. Saying that consciousness is distinct from this natural kind contradicts the INK strategy.

For this reason, this response also makes it questionable how empirical research can produce determinate answers to the question which creatures are conscious at all, since it eliminates the option of saying that creatures are conscious if and only if they share the natural kind which underlies the cluster of consciousness-relevant features in humans. It is not obvious what an alternative, empirically accessible interpretation of the question which creatures are conscious could be.

Here are some further considerations which favor identifying consciousness with the coarse-grained functional kind, rather than the biological one. First, a theoretically elegant move for proponents of non-biological views is to hold that biology-based measures of consciousness (e.g. based on the presence of certain brain structures) are simply *not applicable* to non-biological creatures, rather than supporting the view that these creatures lack consciousness. Then, the absence of the biological properties these measures target does not need to be treated as speaking against attributions of consciousness to non-biological creatures.

This is plausible since the INK strategy generally requires us to realize that certain measures of consciousness are applicable in some populations, but not in others (Bayne et al. 2024). For example, verbal report cannot be used on (most) non-human animals. In addition, one can make a symmetry argument: If we – assuming that psychological duplicates are conscious – would devise measures of consciousness based on the physical realizer of psychological duplicates and then apply them to humans, the measures would give the result that humans are not conscious (since they are made out of other material). However, it seems more reasonable to say that the measure is simply inapplicable to humans, because it

[9] Potentially, one could individuate these capacities so finely that they are not shared by psychological duplicates, but – even then – the key explananda of consciousness science are ordinarily not individuated at that fineness of grain.

presupposes that its target is made from silicon. Yet, if so, by the same principle, it is plausible to say that biology-based measures are simply inapplicable to psychological duplicates, since these measures presuppose that their targets are biological creatures.

Second, it is not only the case that the cluster explained by a coarse-grained functional kind explains most of the phenomena targeted by putative measures of consciousness. In addition, these are phenomena – like verbal report and visual discrimination – which are particularly central to our pre-theoretical conception of consciousness. Pre-theoretically, we characterize consciousness as the kind which explains central phenomena like verbal report and visual discrimination, not necessarily certain biological traits. Since this pre-theoretical conception, encapsulated in pre-theoretical principles, is the initial starting point for consciousness science, it should be – to some extent – reflected in the natural kind consciousness science eventually converges on.

The third argument is the most speculative one I will make: By definition, psychological duplicates talk and interact with us – over their whole life span – in the same ways as our fellow humans. In addition, there is no reason why psychological duplicates could not have a body which mirrors human outward appearance, including facial expressions and outside material which feels like human skin. In that situation, it is plausible that most laypeople would have high and unshakeable confidence that psychological duplicates are conscious, if they encountered them.[10] If so, scientific accounts which identify consciousness with a coarse-grained functional kind conform better to the deeply held convictions of the society which embeds this science.

This consideration may be relevant if and because science ought to be (to some extent) receptive to societal demands (e.g. Douglas 2009). A science which depicts psychological duplicates as not conscious might contradict what most laypeople would take to be an obvious assumption, for example when thinking about how to interact with non-biological creatures. Therefore, scientific results could arguably not fruitfully inform societal decision-making since they would not be acceptable to the concerns and deep convictions of most laypeople.

To recap, I have argued that the outcome of an ideal application of the INK strategy is that it is nomologically possible for AI systems (made out of conventional materials) to be conscious. The reason is that this view possesses theoretical virtues, in addition to fit with our pre-theoretical conception of consciousness and possibly also with deeply held convictions of wider society, which make it preferrable to biological views, given the INK strategy. Moreover, since we should take a scientific perspective to the question whether consciousness requires

[10] See Shevlin (n.d.) for a related view.

biology and the INK strategy is the best way to implement a scientific investigation of this question, the outcome of an ideal application of the INK strategy corresponds to the facts of the matter. Thus, non-biological AI consciousness is possible.

Importantly, the possibility of consciousness in psychological duplicates entails the general conclusion that consciousness can be realized by the materials constituting conventional AI systems. A view which states that consciousness can be realized by such materials in psychological duplicates, but not in other AI systems, would be objectionably ad hoc. The latter view is also inconsistent with the view – that I argued for – that consciousness should be identified with a coarse-grained functional kind.

The previous argument rests on two important assumptions worth highlighting again: that psychological duplicates are (nomologically) possible and that questions of AI consciousness should be settled by science, rather than by metaphysics. If one remains skeptical of these views, despite the reasons I provided in their favor, then the conclusion of this argument should be taken as a conditional: If psychological duplicates are possible and if science has priority to metaphysics in questions of AI consciousness, then (non-biological) AI consciousness is possible.

5. Objections from begging the question

Let me discuss one type of objection to my argument. A natural first inclination is to think that the anticipatory argument is question-begging: After all, it seems like the belief that the outcome of an ideal scientific process is p (e.g. *that psychological duplicates are conscious*) is only plausible if one already believes p. However, this is not true. Being strongly convinced of p is a *sufficient* reason to believe that an ideal epistemic process of some kind would converge on p. However, it is not a necessary reason. As my argument shows, consideration of the theoretical virtues implicit in science can also provide a good reason for believing that the outcome of an ideal scientific process is p, independently of already assuming p.

Let me also note that the reasoning supporting my argument does not generalize to other scientific debates, since most (or all) of them turn on uncertain scientific evidence and theoretical virtues which push into competing directions. By contrast, the debate on necessary conditions for AI consciousness depends on judgements about hypothetical cases (like cases of psychological duplication) in which all the relevant empirical considerations can be stipulated and, I have argued, the consideration of theoretical virtues strongly favors one particular view.

To see more concretely why the anticipatory argument is not begging any questions against biological views, consider three assumptions, and types of assumptions, I rely on which are prima facie candidates for being question-begging:

A. The assumption that psychological duplicates are possible.
B. Assumptions about what features are measures of consciousness.
C. Assumptions about pre-theoretical principles in consciousness science, according to which certain behaviors (such as verbal report) constitute evidence of consciousness.

A is not question-begging, since proponents of biological views are not committed to the view that psychological duplicates are impossible. Instead, many biological views are best interpreted as implying that psychological duplicates are not conscious, not as claiming that they are impossible (see section 3.2).

B is not question-begging since my argument allows that a wide variety of biological features are measures of consciousness. In fact, I assumed that the cluster of consciousness-relevant features we find in humans comprises many biological, e.g. neuroscientific, features. However, I argued that – even given this assumption – it is overall more plausible to identify consciousness with a coarse-grained functional kind, rather than a biological one.

C is not question-begging since all consciousness researchers which take a third-person, empirically driven (observational or experimental) approach must make the pre-theoretical assumption that certain kinds of behaviors (for example, verbal report, visual discriminations or intentional behavior) are (defeasible) evidence of consciousness. Otherwise, we have no initial measures of consciousness such that a science of consciousness cannot get off the ground. This assumption is compatible with biological views, since it specifies that pre-theoretical measures are defeasible. So, assuming that (e.g.) verbal report is a pre-theoretical measure of consciousness is compatible with the view that some creatures which provide verbal reports are not conscious. This is because pre-theoretical principles only have a minimal degree of epistemic standing (Michel 2023) which means that they can be outweighed – for instance, if a creature lacks certain neural properties and we have good reasons to believe that these neural properties are relevant to consciousness.

In total, none of my assumptions are question-begging against biological views.

6. Generalizing beyond the INK strategy

I will conclude with a brief argument suggesting that my argument does not depend on the INK strategy specifically, since other accounts of how to study non-human consciousness either independently support the view that psychological duplicates are conscious or need to rely on

the INK strategy (or something similar) to make assessments of consciousness in psychological duplicates. [11] Hence, while this matter requires further exploration, the following considerations should at least raise one's credence in the view that methodologies beyond the INK strategy can support my anticipatory argument.

It seems to me that there are two main alternatives to the INK strategy in the literature:

1. Propose self-standing behavioral measures of consciousness, without an iterative search for natural kinds.

For example, Tye (2017) argues that types of behaviors which are caused by conscious experience in humans provide (defeasible) evidence of consciousness, when they occur in other animals species. Since psychological duplicates are behaviorally equivalent to humans, it is clear that we would attribute consciousness to some possible non-biological AI systems, if we think behavioral measures are decisive. If we think behavioral measures are only heuristically useful and must eventually be superseded by more reliable types of measures, then we need an independent account (like the INK strategy) which tells us what these further measures consist in and how they should be calibrated.

2. Make assessments of consciousness based on comprehensive theories of consciousness (Carruthers 2020; Doerig et al. 2021; Seth and Bayne 2022).

My argument may generalize to such *theory-based strategies* in consciousness science. It has often been noted that theories of consciousness are underspecified, having the consequence that they can be applied in multiple and conflicting ways to non-human systems (Butlin et al. 2023; de Weerd 2024; Dung 2022; Michel 2019; Shevlin 2021). Moreover, they arguably *should* initially be under-specified in this way. The choice whether a theory of consciousness should be specified further such that it allows for non-biological consciousness or not should be made in an evidence-based manner, not by stipulation (see Shevlin 2021). We need independent evidence to tell us whether a theory of consciousness should be interpreted as attributing consciousness to AI systems.

For example, the perceptual reality monitoring theory (PRMT) (Lau 2022; Michel forthcoming) claims that consciousness involves monitoring the reliability of one's own sensory signals. According to the theory, a PRM mechanism in the brain produces pointer representations which encode information about how reliably neuronal signals represent the

[11] That being said, my argument does not generalize to methodologies in consciousness science which are based on introspective evidence such as the attempted axiomatic justification of the integrated-information theory (Bayne 2018; Tononi and Koch 2015). I only consider approaches which focus on third-person empirical evidence, first-person data are beyond my scope. Panpsychism (Goff 2017) is also outside the scope of my argument, since taking panpsychism as a basis corresponds to a metaphysically driven investigation of non-human consciousness, as opposed to the science-based methodology my argument rests on.

world as it is right now. Then, a neuronal representation of a feature is taken to be conscious in virtue of being flagged as reliable by a (meta-representational) pointer produced by a PRM mechanism.

As it happens, proponents of PRMT are sympathetic to an interpretation according to which the theory is compatible with non-biological consciousness (Butlin et al. 2023, section 2.3; Michel and Lau forthcoming). However, the empirical evidence for the theory is mostly based on brain imaging studies in humans. This evidence cannot distinguish between the hypothesis that consciousness requires biology and that it does not: after all, all humans are biological creatures. So, to examine whether PRMT should be interpreted such that it supports AI consciousness or not, one would need to draw on a wider range of evidence from a wider range of creatures, including non-biological creatures.

This suggests that a theory-based strategy – if it wants to be informed by relevant evidence – may need to draw on the same kinds of factors when assessing which artificial systems to attribute consciousness to (if any) as the INK strategy: a) a range of consciousness-relevant measures that have been calibrated in biological creatures (especially humans), including behavioral criteria, b) clusters between these measures, c) considerations of theoretical virtues, d) fit with our pre-theoretical conception of conciousness, and e) (potentially) fit with the deep convictions of larger society. If so, since the sources of considerations for assessing biological views would be the same as on the INK strategy, a theory-based approach to consciousness science would reach the same verdict: A coarse-grained functional interpretation of PRMT (or any other theory) is supported by theoretical virtues, our pre-theoretical conception of consciousness, and fit with the beliefs of the surrounding society. Thus, there are some reasons to think that an ideal theory-based investigation also reaches the verdict that AI systems can be conscious.

References

Bayne, T. (2018). On the axiomatic foundations of the integrated information theory of consciousness. *Neuroscience of Consciousness*, *2018*(1). https://doi.org/10.1093/nc/niy007

Bayne, T., Seth, A. K., Massimini, M., Shepherd, J., Cleeremans, A., Fleming, S. M., et al. (2024). Tests for consciousness in humans and beyond. *Trends in Cognitive Sciences*, *0*(0). https://doi.org/10.1016/j.tics.2024.01.010

Bayne, T., & Shea, N. (2020). Consciousness, Concepts and Natural Kinds. *Philosophical Topics*, *48*(1), 65–83. https://doi.org/10.5840/philtopics20204814

Birch, J. (2022). The search for invertebrate consciousness. *Noûs*, *56*(1), 133–153. https://doi.org/10.1111/nous.12351

Birch, J. (2023). When is a brain organoid a sentience candidate? *Molecular Psychology: Brain, Behavior, and Society*, *2*, 22. https://doi.org/10.12688/molpsychol.17524.1

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

Bird, A., & Tobin, E. (2023). Natural Kinds. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Spring 2023.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2023/entries/natural-kinds/. Accessed 13 September 2023

Boyle, A. (forthcoming). Disagreement & classification in comparative cognitive science. *Noûs*, *n/a*(n/a). https://doi.org/10.1111/nous.12480

Boyle, A., & Blomkvist, A. (forthcoming). Elements of Episodic Memory: Insights From Artificial Agents. *Philosophical Transactions of the Royal Society B*.

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. (2021). Sharing Our Concepts with Machines. *Erkenntnis*. https://doi.org/10.1007/s10670-021-00491-w

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., et al. (2023, August 22). Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. arXiv. 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

Carruthers, P. (2020). *Human and Animal Minds: The Consciousness Questions Laid to Rest*. Oxford, New York: Oxford University Press.

Chalmers, D. J. (2023). Could a Large Language Model Be Conscious? *Boston Review*, *1*. https://philarchive.org/rec/CHACAL-3. Accessed 17 November 2024

Chang, H. (2004). *Inventing Temperature: Measurement and Scientific Progress*. New York: Oxford University Press. https://doi.org/10.1093/0195171276.001.0001

de Weerd, C. R. (2024). A credence-based theory-heavy approach to non-human consciousness. *Synthese*, *203*(5), 171. https://doi.org/10.1007/s11229-024-04539-6

Doerig, A., Schurger, A., & Herzog, M. H. (2021). Hard criteria for empirical theories of consciousness. *Cognitive Neuroscience*, *12*(2), 41–62. https://doi.org/10.1080/17588928.2020.1772214

Douglas, H. (2009). *Science, Policy, and the Value-Free Ideal*. Pittsburgh: University of Pittsburgh Press.

Dung, L. (2022). Assessing tests of animal consciousness. *Consciousness and Cognition*, *105*, 103410. https://doi.org/10.1016/j.concog.2022.103410

Dung, L. (2023). Tests of Animal Consciousness are Tests of Machine Consciousness. *Erkenntnis*. https://doi.org/10.1007/s10670-023-00753-9

Elamrani, A., & Yampolskiy, R. V. (2019). Reviewing Tests for Machine Consciousness. *Journal of Consciousness Studies*, *26*(5–6), 35–64.

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

Godfrey-Smith, P. (2016). Mind, Matter, and Metabolism. *Journal of Philosophy*, *113*(10), 481–506. https://doi.org/10.5840/jphil20161131034

Godfrey-Smith, P. (2020). *Metazoa: Animal minds and the birth of consciousness*. William Collins.

Goff, P. (2017). *Consciousness and Fundamental Reality* (Vol. 1). Oxford University Press. https://doi.org/10.1093/oso/9780190677015.001.0001

Goldstein, S., & Kirk-Giannini, C. D. (n.d.). *A Case for Ai Consciousness: Language Agents and Global Workspace Theory*.

Goldstein, S., & Kirk-Giannini, C. D. (forthcoming). AI Wellbeing. *Asian Journal of Philosophy*. https://philarchive.org/rec/GOLAWE-4. Accessed 29 October 2024

Irvine, E., & Sprevak, M. (2020). Eliminativism About Consciousness. In U. Kriegel (Ed.), *The Oxford Handbook of the Philosophy of Consciousness* (pp. 347–370). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198749677.013.16

Ladak, A. (2023). What would qualify an artificial intelligence for moral standing? *AI and Ethics*, *4*. https://doi.org/10.1007/s43681-023-00260-1

Lau, H. (2022). *In Consciousness we Trust: The Cognitive Neuroscience of Subjective Experience*. Oxford, New York: Oxford University Press.

Lindsay, G. W. (2020). Attention in Psychology, Neuroscience, and Machine Learning. *Frontiers in Computational Neuroscience*, *14*. https://doi.org/10.3389/fncom.2020.00029

Long, R., Sebo, J., Butlin, P., Finlinson, K., Fish, K., Harding, J., et al. (2024, November 4). Taking AI Welfare Seriously. arXiv. https://doi.org/10.48550/arXiv.2411.00986

Marr, D. (1982). *Vision. A computational investigation into the human representation and processing of visual information*. Cambridge, MA: MIT Press. https://mitpress.mit.edu/9780262514620/vision/. Accessed 23 October 2024

Mckilliam, A. (forthcoming). Natural kind reasoning in consciousness science: An alternative to theory testing. *Noûs*, *n/a*(n/a). https://doi.org/10.1111/nous.12526

Michel, M. (2019). Fish and microchips: on fish pain and multiple realization. *Philosophical Studies*, *176*(9), 2411–2428. https://doi.org/10.1007/s11098-018-1133-4

Michel, M. (2023). Calibration in Consciousness Science. *Erkenntnis*, *88*(2), 829–850. https://doi.org/10.1007/s10670-021-00383-z

Michel, M. (forthcoming). The Perceptual Reality Monitoring Theory. In M. Herzog, A. Schurger, & A. Doerig (Eds.), *Scientific Theories of Consciousness: The Grand Tour*. Cambridge University Press.

Michel, M., & Lau, H. (forthcoming). Higher-Order Theories Do Just Fine. *Cognitive Neuroscience*.

Perez, E., & Long, R. (2023, November 14). Towards Evaluating AI Systems for Moral Status Using Self-Reports. arXiv. https://doi.org/10.48550/arXiv.2311.08576

Piccinini, G. (2021). The Myth of Mind Uploading. In I. Hipólito, R. W. Clowes, & K. Gärtner (Eds.), *The Mind-Technology Problem : Investigating Minds, Selves and 21st Century Artefacts* (pp. 125–144). Springer Verlag.

Place, U. T. (1956). Is Consciousness a Brain Process? *British Journal of Psychology*, *47*(1), 44–50. https://doi.org/10.1111/j.2044-8295.1956.tb00560.x

Saad, B. (2024, July 12). In search of a biological crux for AI consciousness. *Global Priorities Institute*. https://globalprioritiesinstitute.org/in-search-of-a-biological-crux-for-ai-consciousness-bradford-saad/. Accessed 24 October 2024

Saad, B., & Bradley, A. (2022). Digital suffering: why it's a problem and how to prevent it. *Inquiry: An Interdisciplinary Journal of Philosophy*, *0*(0), 1–36. https://doi.org/10.1080/0020174X.2022.2144442

Sebo, J., & Long, R. (2023). Moral consideration for AI systems by 2030. *AI and Ethics*. https://doi.org/10.1007/s43681-023-00379-1

Seth, A. K. (2021). *Being You: A New Science of Consciousness*. Penguin Random House.

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

Shea, N. (2012). Methodological Encounters with the Phenomenal Kind. *Philosophy and Phenomenological Research*, *84*(2), 307–344. https://doi.org/10.1111/j.1933-1592.2010.00483.x

Shevlin, H. (2021). Non-human consciousness and the specificity problem: A modest theoretical proposal. *Mind & Language*, *36*(2), 297–314. https://doi.org/10.1111/mila.12338

Shevlin, H. (n.d.). *Consciousness, Machines, and Moral Status*.

Shiller, D. (2024). Functionalism, integrity, and digital consciousness. *Synthese*, *203*(2), 47. https://doi.org/10.1007/s11229-023-04473-z

Smart, J. J. C. (1959). Sensations and Brain Processes. *The Philosophical Review*, *68*(2), 141–156. https://doi.org/10.2307/2182164

Stacho, M., Herold, C., Rook, N., Wagner, H., Axer, M., Amunts, K., & Güntürkün, O. (2020). A cortex-like canonical circuit in the avian forebrain. *Science*, *369*(6511), eabc5534.

https://doi.org/10.1126/science.abc5534

Taylor, A. H., Bastos, A. P. M., Brown, R. L., & Allen, C. (2022). The signature-testing approach to mapping biological and artificial intelligences. *Trends in Cognitive Sciences*, *26*(9), 738–750. https://doi.org/10.1016/j.tics.2022.06.002

Tononi, G., & Koch, C. (2015). Consciousness: here, there and everywhere? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1668). https://doi.org/10.1098/rstb.2014.0167

Tye, M. (2017). *Tense Bees and Shell-Shocked Crabs: Are Animals Conscious?* Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190278014.001.0001

Wiese, W. (2024). Artificial consciousness: a perspective from the free energy principle. *Philosophical Studies*, *181*(8), 1947–1970. https://doi.org/10.1007/s11098-024-02182-y

Wiese, W., & Friston, K. J. (2021). The neural correlates of consciousness under the free energy principle: From computational correlates to computational explanation. *Philosophy and the Mind Sciences*, *2*. https://doi.org/10.33735/phimisci.2021.81