If consciousness is dynamically relevant, artificial intelligence isn't conscious

Johannes Kleiner[1,2,3] and Tim Ludwig[4]

[1]Munich Center for Mathematical Philosophy, LMU Munich
[2]Munich Graduate School of Systemic Neurosciences, LMU Munich
[3]Association for Mathematical Consciousness Science
[4]Institute for Theoretical Physics, Utrecht University
Princetonplein 5, 3584 CC Utrecht, The Netherlands

Abstract. We demonstrate that if consciousness is relevant for the temporal evolution of a system's states—that is, if it is *dynamically relevant*—then AI systems cannot be conscious. That is because AI systems run on CPUs, GPUs, TPUs or other processors which have been designed and verified to adhere to computational dynamics that systematically preclude or suppress deviations. The design and verification preclude or suppress, in particular, potential consciousness-related dynamical effects, so that if consciousness is dynamically relevant, AI systems cannot be conscious.

The question of whether artificial intelligence (AI) systems are conscious has emerged as one of critical scientific, philosophical, and societal concern. While empirical support to differentiate theories of consciousness is still nascent and while current measures of consciousness (the simplest example of which is interpretation of verbal reports) cannot justifiably be applied to AI systems, our best hope for reliable answers is to link AI's potential for consciousness with fundamental properties of conscious experience that have empirical import or philosophical credibility.

Significant progress in this regard has already been achieved. In [12], David Chalmers assesses evidence for or against AI consciousness based on an extensive array of features that a system or organism might possess or lack, such as self-report, conversational ability, general intelligence, embodiment, world or self-models, recurrent processing, or the presence of a global workspace. In [48], Wanja Wiese proposes a criterion for distinguishing between conscious and non-conscious AI, anchored in the desiderata of the neuroscientific Free Energy Principle.[1]

In this paper, we propose a result of similar nature, which however does not rely on system features and how they relate to consciousness, but on a general property of theories of consciousness: *dynamical relevance*. Here, *dynamical* refers to the temporal evolution of a system's states. Consciousness is *relevant* to a system's time evolution if the time evolution with consciousness, as described by a theory of consciousness, differs from the time evolution without consciousness, as described by a physical theory. Crucially, consciousness can be dynamically relevant in both physicalist and non-physicalist ontologies. An example of the former is a theory which posits that consciousness serves a specific functional role that is absent in systems without consciousness; an example of the latter is a theory that violates the closure of the physical.

What sets AI systems apart in the context of consciousness is not the specific computational architecture that is employed; architectures that closely resemble the mammalian brain's computational structure can arguably also be used, after all [18]. Instead, the distinctive aspect is the hardware on which an AI architecture operates, namely CPUs, GPUs, TPUs, or other processors. This hardware is designed and verified to ensure that the physical dynamics evolve precisely as described by a computational theory during what is known as *functional* and *post-silicon verification*. These verification processes ensure that the physical design of the chip (the layout of integrated circuits in terms of semiconductors), as well as the actual

[1]These are examples of research whose aim is to evaluate *whether* AI systems of the more recent form are or can be conscious. Other interactions between AI research and consciousness science include the use of AI inspired tools and concepts to model consciousness, for example [5, 23, 25], and studies of how models of consciousness might help to build better AI, for example [6, 24, 37]. The question of whether machines in general can be conscious has guided much of the debate in philosophy of mind over the decades, notable contributions include [4, 8, 11, 14, 15, 16, 20, 21, 40, 43, 44, 46, 47].

physical product (the processing unit after production), yield dynamics exactly as specified by the computational theory. Any dynamical effects that violate the specification of this theory are excluded or dynamically suppressed by error correction.

The *intuition* behind our result is summarised below. The objective of the paper is to delineate all concepts involved in this intuition carefully, so as to present a theorem that underwrites the intuition both in scope and precision.

(A1) Verification of processing units ensures that any dynamical effects that change the computational dynamics of a processing unit are precluded or suppressed.
(A2) If consciousness is dynamically relevant, and AI systems are conscious, then there are dynamical effects that change the computational dynamics of an AI system.
(A3) AI systems run on processing units.

(C) If consciousness is dynamically relevant, AI systems cannot be conscious.

The conclusion (C) follows because qua (A3) and (A1), verification ensures that any dynamical effects that change the computational dynamics of an AI system are precluded or suppressed. (A2) states that if consciousness is dynamically relevant, and AI systems are conscious, then there are dynamical effects that change the computational dynamics of an AI system. Therefore, if consciousness is dynamically relevant, then AI systems cannot be conscious. The crucial work of the formalisation we introduce below is to make sure this reasoning is also sound if consciousness' dynamical effects apply on a "level below" the computational level.

In a nutshell, this paper shows that if consciousness makes a difference to how a system evolves in time—as it should if consciousness is to have any evolutionary advantage, for example—then any system design which systematically precludes or suppresses diverging dynamical effects systematically precludes or suppresses the system from being conscious.

1. Preliminaries

The central notion which underlies our result is that of the time evolution of a system's states. Given a scientific theory T and a system S within the scope of the theory, we denote by $k_T(S, s)$ the *dynamical evolution* (also called 'trajectory') of S with initial state s. This dynamical evolution describes how the state s evolves in time according to T. An example is the evolution of a brain state according to a neuroscientific theory. We will mostly abbreviate $k_T(S, s)$ by k_T if it is clear from context that we're talking about one system and one initial state.

The class of theories which is relevant in the present context are physical theories, on the one hand, and theories of consciousness, on the other hand. We use the symbol Υ to denote physical theories that have been discovered by the natural sciences, in so far as they are relevant for AI or consciousness. Examples are theories of neuroscience, biology, chemistry, computer science and physics.

Different theories describe systems at different levels [31], and in some cases, the states of a system posited by one theory T (the "lower" level) can (in principle) be mapped to sates of another theory T' (the "higher" level). If this is the case, we write $T < T'$. Because dynamical evolutions are sequences of states, if $T < T'$, we can map any dynamical evolution k_T of T to a (not necessarily dynamical) evolution of T', which we denote as $k_T|_{T'}$. We assume that there is a fundamental physical theory $T_F \in \Upsilon$ that can be mapped to states of any other physical theory in Υ, so that $T_F < T$ for all $T \in \Upsilon$. It is likely that the states of quantum theory can, in principle, be mapped to states of all physical theories in Υ, which is why for all practical purposes we can think of T_F as quantum theory.

We also assume that there is a fact to the matter of what the real (that is: actual) dynamics of any system are, even if that fact may not be knowable. We denote the description of the real dynamics in terms of the states of any physical theory $T \in \Upsilon$ (any "level" of description, so to speak) by $k^*|_T$. If $T < T'$, the description of the real dynamics in terms of the states of both theories are compatible, that is $k^*|_T|_{T'} = k^*|_{T'}$.

2. Theories of Consciousness

The second class of theories that are relevant in this context are theories of consciousness (tocs), which are sometimes also called models of consciousness. Tocs express a relation between a physical description of a system, on the one hand, and a description of its conscious experiences, on the other hand. The latter could be a description of its phenomenal character (cf. e.g. [26, 29]), or simply an expression of whether a system S has conscious experiences at all. Together, the physical description and the description of conscious experiences applied by a toc M constitute a state s of the toc, and the dynamical evolution $k_M(S, s)$ of this state expresses how the physical and consciousness relate according to the toc M. Independently of what the description is that a toc applies on the side of consciousness, there is a fact to the matter of whether a system is conscious or not in

$k_M(S,s)$, that is: whether the system S has conscious experiences at least at one point of time in the dynamical evolution $k_M(S,s)$.

Because tocs contain a physical description of a system at some level, for every toc M, there is at least one physical theory $T_P \in \Upsilon$ such that the physical part of any state s of M, and therefore also any dynamical evolution k_M, can be expressed in T_P. We denote this by $s|_{T_P}$ and $k_M|_{T_P}$, respectively. So, $k_M|_{T_P}$ is what M says about the evolution of physical states on T_P's level of description. We call any such T_P an *underlying* physical theory of M.

Making use of this important link between tocs and physical descriptions, we can say that a system S *is conscious* in a physical evolution k_{T_P} iff there is a dynamical evolution k_M of M such that (a) we have $k_M|_{T_P} = k_{T_P}$ and (b) the system is conscious in k_M.

Whether a toc has anything original to say about the dynamical evolution of its physical states, or simply presumes the dynamical evolution of an underlying physical theory, is precisely the question of dynamical relevance, defined as follows. Let M denote a toc and $T_P \in \Upsilon$ an underlying physical theory of M.

Definition 1. *Consciousness is dynamically relevant* according to M with respect to T_P iff

$$S \text{ is conscious in } k_M \;\Rightarrow\; k_M|_{T_P} \neq k_{T_P}\ .$$

Here, the right-hand-side is short-hand for $k_M(S,s)|_{T_P} \neq k_{T_P}(S, s|_{T_P})$, where $s|_{T_P}$ denotes the restriction of the state s of M to T_P. The left-hand side is a shorthand for 'S is conscious in $k_M(S,s)$', meaning that there is at least one point of time in $k_M(S,s)$ so that S has a conscious experience at that time according to M. The definition expresses the intuition that if S is conscious according to a toc M, then the dynamical evolution as specified by M differs from the dynamical evolution as specified by the underlying physical theory alone.

We have already referenced the 'real' dynamics of a system and introduced the symbol $k^*|_{T_P}$ to denote what the real dynamics of a system would look like in terms of the states of T_P. There is also a fact to the matter of whether a system in a trajectory k^* is conscious and how conscious experiences relate to the physical. That is, there is a 'true' or 'real' theory of consciousness, which we denote by M^*. As in the physical case, M^* may be unknown and or unknowable. We will denote its dynamical evolutions by k_{M^*}. Because these describe what really happens, we have $k_{M^*}|_{T_P} = k^*|_{T_P}$ for all T_P. Using M^*, we can define dynamical relevance simpliciter:

Definition 2. *Consciousness is dynamically relevant* (CDR) only if it is dynamically relevant according to the 'true' toc M^* with respect to *some* physical theory $T_P \in \Upsilon$.

3. Verification

What is unique about AI systems in the present context is not the particular architecture that is employed; AI can also be built on architecture derived from the brain; cf. e.g [18]. What is unique is rather that the architecture runs on CPUs, GPUs, TPUs or other processors that have been designed and *verified* in the lab.

There are two major verification steps in processor development, called functional and post-silicon verification. *Functional verification* [34, 49] is applied once the design of a processor in terms of integrated circuits has been laid out, but before the manufacturing phase begins. It applies simulation tools, formal verification tools and hardware emulation tools to ensure that the design of the chip meets the intended specifications as described by a computational theory T_{comp}. *Post-silicon verification* [35, 36] is applied after the silicon waver has been fabricated. It applies in-circuit testing, functional testers, failure analysis tools and reliability testing, among other things, to ensure that the physical product works as T_{comp} would have it.

Functional verification is a theoretical endeavour. It applies simulation and emulation tools based on a theoretical account on how the substrate, on which a processor is to be built, behaves. Because this substrate is a semi-conductor, this theoretical account is based on quantum theory T_F. Put in terms of dynamics, functional verification aims to ensure that whatever happens in the quantum realm implements or is compatible with the dynamics as described by T_{comp}, formally:

$$k_{T_F}|_{T_{\text{comp}}} = k_{T_{\text{comp}}} \tag{3.1}$$

for all dynamical evolutions of a processor S.

Post-silicon verification, on the other hand, is applied to a chip once it has been built. It ensures that the dynamics of the actual physical product comply with T_{comp}. Making use of the k^* notation to denote the actual dynamical evolution of a system, post-silicon verification enforces that

$$k^*|_{T_{\text{comp}}} = k_{T_{\text{comp}}} \tag{3.2}$$

for all dynamical evolutions of a processor S.

Being an AI system means running on CPUs, GPUs, TPUs or other processors that have been designed and verified. That's what makes the system "artificial". And because processor dynamics compose (the output of one is the input of the next), verification holds for AI systems as well: there is an underlying computational theory T_{comp}

that accounts for what "happens" on the processors while the system is running, and the computational dynamics satisfy (3.1) and (3.2).

4. AI Consciousness

With all this in place, we can formulate the question that is being asked precisely. The term 'artificial intelligence' is used very broadly, comprising many different computational architectures and applications. What one means when one asks whether an AI system is conscious is whether the computational architecture that is applied by this system, with the specific quirks of its implementation and training, potentially in a specific task, has conscious experiences. The architecture and these specifics determine the computational dynamics the system is capable of. Thus, the question is whether the system has a computational evolution $k_{T_{\text{comp}}}$ such that it is conscious in this computational evolution according to a theory of consciousness M; cf. Section 2 for a definition of what this means in terms of dynamics k_M of M.[2] In summary:

Definition 3. An *AI system S is conscious* according to a theory of consciousness M only if there is at least one dynamical evolution $k_{T_{\text{comp}}}$ in which the system is conscious according to M.

This is a very weak condition, which however has one important consequence: that the question of AI consciousness is determined by facts on the computational level and above; it is independent of what happens on a sub-computational level. That is, if we have a a trajectory k_{T_P} on a sub-computational level ($T_P < T_{\text{comp}}$) with $k_{T_P}|_{T_{\text{comp}}} = k_{T_{\text{comp}}}$ then S is conscious in $k_{T_{\text{comp}}}$ only if it is conscious in k_{T_P}.

5. Main Result

Our main result is the following theorem.

Theorem 4. If consciousness is dynamically relevant, then AI systems aren't conscious.

Before giving the proof, we first illustrate the result for the simpler case where consciousness is dynamically relevant with respect to the computational level T_{comp} itself. The power of the theorem is to extend this result to all other cases. Subsequent to this illustration, we prove a lemma needed for the main theorem, and then proceed to prove the theorem itself.

So let us consider the case where T_P in Definition 2 is T_{comp}. Let S be an AI system. Because of post-silicon verification (3.2), all of the dynamical evolutions of S satisfy

$$k^*|_{T_{\text{comp}}} = k_{T_{\text{comp}}} \,. \tag{5.1}$$

Application of Definition 2 for the case $T_P = T_{\text{comp}}$ implies, via Definition 1, that if S is conscious in a k_{M^*}, then $k_{M^*}|_{T_{\text{comp}}} \neq k_{T_{\text{comp}}}$. The converse of this statement is that if $k_{M^*}|_{T_{\text{comp}}} = k_{T_{\text{comp}}}$, then S is not conscious in k_{M^*}. Because $k_{M^*}|_{T_{\text{comp}}} = k^*|_{T_{\text{comp}}}$, the identity (5.1) establishes the prerequisite of this condition for all dynamical evolutions of S. Therefore, it follows that S is not conscious in any k_{M^*}. Thus, Definition 3 implies that S is not conscious, as claimed.

The remainder of this section is devoted to the proof of the theorem in the general case. To this end, we first state and prove the following lemma.

Lemma 5. *Dynamical relevance passes downward*, in the sense that if $T_P < T'_P$ and consciousness is dynamically relevant according to M with respect to T'_P, then it is also dynamically relevant according to M with respect to T_P.

Proof of the Lemma. Consciousness is dynamically relevant according to M with respect to T'_P, iff

$$S \text{ is conscious in } k_M \;\Rightarrow\; k_M|_{T'_P} \neq k_{T'_P} \,.$$

Because $T_P < T'_P$, there is a function which maps states—and therefore also dynamical evolutions—from T_P onto T'_P. Therefore, we have

$$k_M|_{T'_P} \neq k_{T'_P} \;\Rightarrow\; k_M|_{T_P} \neq k_{T_P} \,.$$

Together with the above, this gives

$$S \text{ is conscious in } k_M \;\Rightarrow\; k_M|_{T_P} \neq k_{T_P} \,,$$

which is the case iff consciousness is dynamically relevant according to M with respect to T_P. □

We now proceed to the proof of the theorem.

Proof of the Theorem. We first consider the case where T_P in Definition 2 is T_F.

Let S be an AI system. Because of functional and post-silicon verification, we have

$$k_{T_F}|_{T_{\text{comp}}} = k_{T_{\text{comp}}} = k^*|_{T_{\text{comp}}} \tag{5.2}$$

for all dynamical evolutions of S. Because consciousness is (by assumption) dynamically relevant and we have assumed $T_P = T_F$, Definition 1 applies to give

$$S \text{ is conscious in } k_{M^*} \;\Rightarrow\; k_{M^*}|_{T_F} \neq k_{T_F} \tag{5.3}$$

for all dynamical trajectories k_{M^*} of M^*.

[2]The point here is to restrict downwards, not upwards. Any question "above" the computational level can be posed in terms of computational dynamics.

Let us now assume that S is conscious in some trajectory k_{M^*} of M^*. According to the last implication, we thus have

$$k_{M^*}|_{T_F} \neq k_{T_F} \ .$$

Because $T_F < T_{\text{comp}}$, we can map both of these trajectories to T_{comp}. For $k_{M^*}|_{T_F}$, this gives

$$\begin{aligned} k_{M^*}|_{T_F}|_{T_{\text{comp}}} &= k^*|_{T_F}|_{T_{\text{comp}}} \\ &= k^*|_{T_{\text{comp}}} = k_{M^*}|_{T_{\text{comp}}} \ , \end{aligned}$$

where we have made use of identities established in Sections 1 and 2. Equation (5.2) furthermore establishes that

$$k_{M^*}|_{T_{\text{comp}}} = k^*|_{T_{\text{comp}}} = k_{T_{\text{comp}}} \ .$$

The two facts that (a) $k_{M^*}|_{T_{\text{comp}}} = k_{T_{\text{comp}}}$ and (b) that S is conscious in k_{M^*} establish that S is conscious in $k_{T_{\text{comp}}}$.

Equation (5.2) also establishes that

$$k_{T_F}|_{T_{\text{comp}}} = k_{T_{\text{comp}}} \ .$$

Because of this equation and $T_F < T_{\text{comp}}$, the implication of Definition 3 explained in the last paragraph of Section 4 applies and establishes that S is conscious in k_{T_F}.

Unwrapping what 'S is conscious in k_{T_F}' means by definition, we find that there must by a dynamical evolution $\tilde{k}_{M^*}$ of M^* such that

$$\begin{aligned} &\text{(a) } \tilde{k}_{M^*}|_{T_F} = k_{T_F} \quad \text{and} \\ &\text{(b) } S \text{ is conscious in } \tilde{k}_{M^*} \ . \end{aligned}$$

Together, these two conditions violate (5.3). Thus we have arrived at a contradiction.

The assumptions that went into the derivation of this contradiction were that consciousness is dynamically relevant with respect to the T_F level, that S is an AI system, and that S is conscious in a trajectory k_{M^*} of M. The first assumption is stated as a condition in the theorem. Thus it follows that the latter two cannot be both the case.

Because k_{M^*} was arbitrary, it follows that an AI system S cannot be conscious in any trajectory k_{M^*} of M^*. Consequently, applying Definition 3, it cannot be conscious at all. This establishes the claim that if consciousness is dynamically relevant with respect to T_F, then AI systems aren't conscious.

It remains to consider all other cases of T_P in Definition 2. Therefore, let us assume that consciousness is dynamically relevant with respect to some $T_P \neq T_F$. Because $T_F < T_P$ for all $T_P \in \Upsilon$, and because dynamical relevance passes downward (Lemma 5), it follows that consciousness is also dynamically relevant with respect to T_F. Hence the previous case applies and the result follows in full generality. □

6. Is Consciousness Dynamically Relevant?

There are at least three routes to answer this question. Dynamical relevance is an epistemic assumption which is partially related to an ontological assumption known as 'causal closure of the physical' or 'completeness of the physical' [42]: if the physical is not causally closed in virtue of consciousness, then consciousness is dynamically relevant. Therefore, a first route to determine whether consciousness is dynamically relevant is via *philosophy of mind*.

There have been extensive arguments for and against the causal closure of the physical in the literature, cf. [42] for a short summary. To the best of our knowledge, there is to date no conclusive argument *against* the causal closure. On the other hand, no argument *for* causal closure can establish dynamical *ir*relevance because if consciousness is physical, causal closure holds, yet consciousness can still be dynamically relevant.

The second route is to study necessary conditions of some of the *practices* we engage in as researchers or as a society; those are conditions which are presupposed by these practices.

A great example is the empirical investigation of consciousness itself, as pursued broadly now under the roof of the Association for the Scientific Study of Consciousness, among other organisations. Any empirical investigation of consciousness presupposes behavioural measures (such as, but not limited to, reports) that can be used to infer the state of consciousness of a subject in certain contexts. These means of inference are often referred to as *measures of consciousness* [22]. Dynamical relevance is a premise for any measure of consciousness to work as intended. That is because if consciousness does not make a difference to the time evolution of any physical states, its presence or absence cannot be inferred from the physical states that account for body movement (pressing of a button, say) or sound waves (verbal report). Empirical distinguishability of theories of consciousness hinges on dynamical relevance.[3]

Another good example of the second route is the investigation of consciousness from an AI engineering perspective. If consciousness makes a difference to how a system performs, it is dynamically relevant. Therefore, any AI engineering perspective

[3] [27] proves this point by analysing what data is and how it is used in experiments in consciousness science. Data is determined by physical states such as charge position or magnetic orientation; dynamical relevance is required for two theories to cause different such states and hence different data. For details, cf.[27], noting that dynamical relevance is referred to as an 'empirical version of the closure of the physical'.

which asks how the implementation of consciousness can make a difference to a system's evolution presupposes that consciousness is dynamically relevant.

The third route, finally is via existing *theories of consciousness*, where it is helpful to distinguish between metaphysical theories and scientific theories. Metaphysical theories primarily target the question of what consciousness *is*, whereas scientific theories primarily model what consciousness *does*.[4]

Some metaphysical theories, such as type identity theory [41], Russelian-type panpsychism [19], or Chalmers-style dualism [10], presuppose consciousness to not be dynamically relevant. Others, such as interactive dualism [17] or dual aspect monism [1], render consciousness dynamically relevant. Yet others leave the question open, for example most versions of functionalism [30]: it remains unclear whether the function that consciousness is identified with has a dynamical relevance over and above the physical theories that are thought to implement it.

The situation isn't much better in the case of scientific theories, unfortunately. The only unambiguous example we know of is Integrated Information Theory (IIT) [38], which despite its non-physicalist ontology and emphasis of the primacy of conscious experiences proposes a mathematical model in which consciousness is not dynamically relevant.[5] Models such as Global Neuronal Workspace Theory [32] or Higher Order Thought Theory [9] do not imply either case, as far as we can see.

In summary, it seems to us that the only conclusive route *to date* seems to be route number two, which largely speaks in favour of dynamical relevance.

7. Objections

In this section, we discuss a few immediate responses to our result.

7.1. **Verification is imperfect.** Verification is an industrial process that may not be perfect: despite functional and post-silicon verification, the actual dynamics of a processor may not adhere to the computational theory targeted by verification in all cases. Verification may leave a bit of wiggle-room for the dynamics to diverge from the computational theory. Could this wiggle-room suffice for consciousness to unfold its dynamical effects?

Any answer to this question depends on how exactly consciousness is dynamically relevant and which imperfections arise in day-to-day verification. It is natural to expect that consciousness' dynamical relevance is systematic in nature: dynamical effects should systematically occur if a system is conscious and make a systematic difference to how the system evolves in time. The imperfections in day-to-day verification, on the other hand, are likely to be mostly random in nature, meaning that the deviation in dynamical evolution they fail to suppress are random too, both in time (when such a deviation can occur) and in the extend to which they can make a difference. If this is true, it is unlikely that the wiggle-room left open due to imperfections suffices for consciousness to unfold its dynamical effects.

7.2. **Emergence.** Our result is compatible with emergence. If consciousness is weakly emergent from a physical substrate [39], consciousness is not dynamically relevant with respect to this substrate, so that our result does not apply. If consciousness is strongly emergent, it is dynamically relevant: the "fundamental higher-level causal powers" [39, Sect. 4] which exist in this case make a difference to the time evolution of the substrate states, a difference that is excluded or suppressed by any design which is verified to comply to non-emergent substrate dynamics.

7.3. **Probabilistic processing.** Verification as applied in industry targets deterministic computational theories. Would our result also hold in case of verified probabilistic processing?

The mathematical framework we apply is compatible with probabilistic processing: we do not make an assumption as to whether the notions of state and dynamical evolution are deterministic or not; a state may well be a probability distribution and its dynamical evolution a stochastic process. Verification, in this case, implies that a system conforms to the stochastic process as described by a stochastic computational theory. This leaves room for consciousness to have a dynamical effect, but only if this effect conforms to the probability distributions as described by the stochastic computational theory. That is, consciousness may determine how the probability distributions of the stochastic computational theory are sampled, but it cannot change them. As in the case of imperfect verification, we remain sceptical as to whether this limited freedom is compatible with the systematic

[4]Thanks to Kobi Kremnitzer for pointing us to this terminology.

[5]The algorithm that IIT presupposes takes the form of a mapping (function) from a physical descriptions of systems to a space of states of conscious experiences [28], so that the time evolution of the physical state is precisely as described by the underlying physical theory.

nature of consciousness' dynamical effects that are to be expected.

7.4. **Quantum computing.** Does our result also hold true in the case of quantum computing? Quantum computing is a young industry and it is not yet clear which type of verification, if any, will need to be deployed. It is likely, however, that any type of verification will need to presuppose a notion of *measurement*, which is an inherently vague concept in quantum theory [2] that is partially external to the account of quantum dynamics by the Schrödinger equation. If consciousness were related to measurement (for example via consciousness-induced dynamical collapse as proposed in [13]), then verification might leave enough room for consciousness to have a systematic and meaningful effect. If, on the other hand, consciousness is not related to measurement in quantum theory, it is likely that verification of quantum computers to adhere to quantum dynamics will preclude any potential dynamical effects of consciousness; just as in the classical case.

8. Conclusion

This paper addresses the question of whether AI systems are conscious. Its objective is to introduce a new formal tool, in the form of a theorem, that provides an answer to this question which is independent of the specific computational architecture that an AI system utilises, and which does not rely on any specific cognitive feature that an AI system might possess or lack that may be related to conscious experience.

Our result is based on what we take to be the only property that distinguishes AI systems from other cognitive systems, a property that might well embody the actual meaning of the word 'artificial' in 'artificial intelligence': that the system runs on a substrate that has been designed and verified, rather than naturally evolved.

Ultimately, we believe that any statement about whether a system is conscious needs to be based on a theory of consciousness that is supported by theoretical, philosophical, and most importantly empirical evidence. The Science of Consciousness[6] searches for such theories. The crucial premise in our result—dynamical relevance—is a *property* which theories ascribe to consciousness, so that our theorem can be regarded as establishing a fact about AI's capability for consciousness for a whole class of theories of consciousness: all those that posit consciousness to be dynamically relevant. Results of this form are important as long as evidence in favour of any single theory of consciousness, as well as evidence to distinguish among them, is still in its early stages, and while the space of possible theories remains only partially explored.

Our result has a few interesting, slightly funny, and potentially relevant implications for AI engineering and AI interpretability. The most notable of these is that our result shows that *if an AI system states that it is conscious, then this cannot be because it is conscious.* That is to say, the cause of any such statement cannot be that the AI system is conscious. This follows because if such a cause existed, consciousness would have to be dynamically relevant, in which case our theorem implies that the system isn't conscious. Another implication is that if consciousness has functions that could improve a system's information processing, then, to make use of those functions, theories of consciousness should be taken into account when designing the substrate on which an AI system will run.

The question of whether AI systems are conscious is of major societal concern. It has important ethical [7, 33], legal [3, 45], and technological consequences, and will likely play a major role in shaping governance of AI and how individuals interact with this technology. Our result aims to deliver a rigorous and justified answer to this question that does not rely on contingent assumptions, such as the truth of a particular theory of consciousness, or the validity of a particular test of consciousness when applied to AI systems.

Acknowledgements. We would like to thank the participants of the *Modelling Consciousness Workshops 2022 and 2023* of the *Association for Mathematical Consciousness Science* for valuable discussions on the topic of AI consciousness and feedback about this result, specifically Alexandra Proca, Cameron Beebe, Sophie Taylor, Peter Thestrup Waade, Joscha Bach, Mathias Gutmann, George Deane, Jordan O'Byrne, Ian Durham and Sean Tull. This research was supported by grant number FQXi-RFP-CPW-2018 from the Foundational Questions Institute and Fetzer Franklin Fund, a donor advised fund of the Silicon Valley Community Foundation. We would like to thank the Dutch Research Council (NWO) for (partly) financing TL's work on project number 182.069 of the research programme Fluid Spintronics, and the Mathematical Institute of the University of Oxford for hosting JK while working on this project.

[6]Also called *Scientific Study of Consciousness* to emphasise the importance of contributions from humanities, most notably philosophy.

References

[1] H. Atmanspacher and C. A. Fuchs. *The Pauli-Jung conjecture and its impact today*. Imprint Academic, 2014.

[2] J. Bell. Against 'measurement'. *Physics world*, 3(8):33, 1990.

[3] C. Benzmüller and B. Lomfeld. Reasonable machines: A research manifesto. In *KI 2020: Advances in Artificial Intelligence: 43rd German Conference on AI, Bamberg, Germany, September 21–25, 2020, Proceedings 43*, pages 251–258. Springer, 2020.

[4] N. Block. Troubles with functionalism. In *The Language and Thought Series*, pages 268–306. Harvard University Press, 1980.

[5] L. Blum and M. Blum. A theory of consciousness from a theoretical computer science perspective: Insights from the conscious turing machine. *PNAS*, 2022.

[6] L. Blum and M. Blum. A theoretical computer science perspective on consciousness and artificial general intelligence. *arXiv preprint arXiv:2303.17075*, 2023.

[7] N. Bostrom and E. Yudkowsky. The ethics of artificial intelligence. In *Artificial intelligence safety and security*, pages 57–69. Chapman and Hall/CRC, 2018.

[8] Z. Bronfman, S. Ginsburg, and E. Jablonka. When will robots be sentient? *Journal of Artificial Intelligence and Consciousness*, 8(02):183–203, 2021.

[9] R. Brown, H. Lau, and J. E. LeDoux. Understanding the higher-order approach to consciousness. *Trends in cognitive sciences*, 23(9):754–768, 2019.

[10] D. J. Chalmers. *The conscious mind: In search of a fundamental theory*. Oxford Paperbacks, 1997.

[11] D. J. Chalmers. The singularity: A philosophical analysis. *Journal of Consciousness Studies*, 2010.

[12] D. J. Chalmers. Could a large language model be conscious? *arXiv preprint arXiv:2303.07103*, 2023.

[13] D. J. Chalmers and K. J. McQueen. Consciousness and the collapse of the wave function. In S. Gao, editor, *Consciousness and Quantum Mechanics*. Oxford University Press, forthcoming.

[14] W. J. Clancey. The strange, familiar, and forgotten: An anatomy of consciousness. *Artificial Intelligence*, 60(2):313–356, 1993.

[15] A. Clark. *Being there: Putting brain, body, and world together again*. MIT press, 1998.

[16] D. C. Dennett. *Consciousness Explained*. Little, Brown and Co., 1991.

[17] J. C. Eccles and K. Popper. *The self and its brain: an argument for interactionism*. Routledge, 1977.

[18] K. J. Friston, M. J. Ramstead, A. B. Kiefer, A. Tschantz, C. L. Buckley, M. Albarracin, R. J. Pitliya, C. Heins, B. Klein, B. Millidge, et al. Designing ecosystems of intelligence from first principles. *arXiv preprint arXiv:2212.01354*, 2022.

[19] P. Goff. *Consciousness and fundamental reality*. Oxford University Press, 2017.

[20] J. Haugeland. *Artificial intelligence: The very idea*. MIT press, 1989.

[21] O. Holland. *Machine consciousness*. Imprint Academic, 2003.

[22] E. Irvine. Measures of consciousness. *Philosophy Compass*, 8(3):285–297, 2013.

[23] X. Ji, E. Elmoznino, G. Deane, A. Constant, G. Dumas, G. Lajoie, J. Simon, and Y. Bengio. Sources of richness and ineffability for phenomenally conscious states. *arXiv preprint arXiv:2302.06403*, 2023.

[24] A. Juliani, K. Arulkumaran, S. Sasai, and R. Kanai. On the link between conscious function and general intelligence in humans and machines. *Transactions on Machine Learning Research*, 2022.

[25] A. Juliani, R. Kanai, and S. S. Sasai. The perceiver architecture is a functional global workspace. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44, 2022.

[26] J. Kleiner and T. Ludwig. What is a mathematical structure of conscious experience? *arXiv preprint arXiv:2301.11812*, 2023.

[27] J. Kleiner and H. Stephan. The closure of the physical, consciousness and scientific practice. *arXiv preprint arXiv:2110.03518*, 2023.

[28] J. Kleiner and S. Tull. The mathematical structure of Integrated Information Theory. *Frontiers in Applied Mathematics and Statistics*, 6:602973, 2021.

[29] A. Y. Lee. Modeling mental qualities. *Philosophical Review*, 130(2):263–298, 2021.

[30] J. Levin. Functionalism. In E. N. Zalta and U. Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Stanford University, 2023.

[31] C. List. Levels: descriptive, explanatory, and ontological. *Noûs*, 53(4):852–883, 2019.

[32] G. A. Mashour, P. Roelfsema, J.-P. Changeux, and S. Dehaene. Conscious processing and the Global Neuronal Workspace Hypothesis. *Neuron*, 105(5):776–798, 2020.

[33] T. Metzinger. Artificial suffering: An argument for a global moratorium on synthetic phenomenology. *Journal of Artificial Intelligence and Consciousness*, 8(01):43–66, 2021.

[34] P. Mishra and N. D. Dutt. *Functional verification of programmable embedded architectures: a top-down approach*. Springer Science & Business Media, 2005.

[35] P. Mishra, R. Morad, A. Ziv, and S. Ray. Post-silicon validation in the soc era: A tutorial introduction. *IEEE Design & Test*, 34(3):68–92, 2017.

[36] S. Mitra, S. A. Seshia, and N. Nicolici. Post-silicon validation opportunities, challenges and recent advances. In *Proceedings of the 47th Design Automation Conference*, pages 12–17, 2010.

[37] D. C. Mollo and R. Millière. The vector grounding problem. *arXiv preprint arXiv:2304.01481*, 2023.

[38] M. Oizumi, L. Albantakis, and G. Tononi. From the phenomenology to the mechanisms of consciousness: Integrated Information Theory 3.0. *PLoS computational biology*, 10(5):e1003588, 2014.

[39] T. O'Connor. Emergent Properties. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Stanford University, 2021.

[40] R. Penrose. *The emperor's new mind*. Oxford University Press, 1989.

[41] U. T. Place. Is consciousness a brain process? *British Journal of Psychology*, 47(1):44–50, 1956.

[42] D. Robb, J. Heil, and S. Gibb. Mental Causation. In E. N. Zalta and U. Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Stanford University, 2023.

[43] J. R. Searle. Minds, brains, and programs. *Behavioral and brain sciences*, 3(3):417–424, 1980.

[44] S. W. Smoliar. The remembered present: A biological theory of consciousness: Gerald m. edelman. *Artificial Intelligence*, 52(3):295–318, 1991.

[45] R. Susskind. *Online courts and the future of justice*. Oxford University Press, 2019.

[46] M. Tegmark. *Life 3.0: Being human in the age of artificial intelligence*. Knopf, 2017.

[47] A. M. Turing. Computing machinery and intelligence. *Mind*, 1950.

[48] W. Wiese. Could large language models be conscious? a perspective from the Free Energy Principle. *Preprint*, 2023.

[49] B. Wile, J. Goss, and W. Roesner. *Comprehensive functional verification: The complete industry cycle*. Morgan Kaufmann, 2005.