Preprint, feedback welcome (wanja.wiese@rub.de). This manuscript has been submitted to I. Hipólito, C. Hesp, & K. Friston (eds.), *The free energy principle: Science, tech & phil*. Routledge.

Could large language models be conscious? A perspective from the free energy principle

Wanja Wiese (wanja.wiese@rub.de)

Abstract

Might future developments in artificial intelligence (AI) lead to the creation of artificial consciousness? Some believe that large language models and other AI systems will at best be able to simulate phenomenal consciousness. These simulations are sometimes called "weak artificial consciousness"—in contrast to "strong artificial consciousness", which refers to artificial systems that are actually conscious (Holland, 2003). What is the difference between weak and strong artificial consciousness?

Here, I shall assume that the causal roles that characterise phenomenal consciousness are medium-independent and can be captured in terms of computation; they are *computational correlates of consciousness* (Cleeremans, 2005; Reggia et al., 2016; Wiese & Friston, 2021). I propose that strong artificial consciousness requires implementing these computations in a particular way: performing these computations must have a function *for the system* itself, by making a contribution to the system's goals, such as sustaining its existence. Only such computational systems *give a damn* (Haugeland, 2000). Hence, some systems may instantiate the computational correlates of consciousness without being conscious, if the respective computational mechanisms do not contribute to the goals of the physical system of which they are a part.

Since the free energy principle (FEP) (Friston, 2019; Parr et al., 2022) provides an analysis of the concept of the existence of particular self-organising systems (Hohwy, 2021), the FEP can be used to specify what it means to contribute to the goals (such as sustaining one's existence) of particular self-organising systems. Together with the assumptions sketched above, an account of the difference between weak and strong artificial consciousness can be developed.

1 Introduction

We live in times in which smart people believe some existing artificial intelligences (AIs) are conscious. Do we need a theory of phenomenal consciousness to determine whether an artificial system is really conscious? I shall argue that the free energy principle (FEP) (Friston, 2019; Parr et al., 2022) provides the means to determine (at least in principle) whether a system is genuinely conscious or not.

The FEP is not a theory, let alone a theory of phenomenal consciousness. However, one can formulate mechanical theories that conform to FEP. A key feature of such a *Bayesian mechanics* (Ramstead et al., 2022) is that they provide conjugate descriptions of a system's physical dynamics and the dynamics of belief, i.e., an internal and an external perspective on the same dynamics (Friston et al., 2020).

Ideally, a mechanical theory can be made so specific that it becomes a theory of consciousness, if it captures the computational correlates of consciousness (Cleeremans, 2005; Reggia et al., 2016), in terms of beliefs encoded by the system's internal states (Wiese & Friston, 2021). This presupposes that there is a meaningful computational difference between conscious and non-conscious processing.

Crucially, this does not mean all systems performing the computations specified by that theory are conscious: a mere simulation of a conscious system can implement the right computations without being conscious. The FEP does not entail an account of this difference, but this paper proposes an account that is coherent with the FEP and only presupposes a few additional assumptions. According to this proposal, the key difference lies in whether the computational correlates of consciousness are entailed by the dynamics of belief that is conjugate to the system's physical dynamics.

In other words, if we start with a physical description that captures the conditions for a system's continued existence and then determine the conjugate description in terms of internally encoded beliefs, we can ask: does the conjugate description entail the computational correlates of consciousness?

For conscious systems, the answer will be yes—under the assumption that consciousness indirectly or directly contributes to the system's sustained existence. For a simulation, the answer will be no—under the assumption that the system's continued existence is not contingent on whether it simulates a conscious system.

The FEP thereby provides the means to specify *observer-independent* and *intrinsic* computations performed by a system. Conversely, this can be used to determine whether a description of the computations that underlie verbal reports or other behaviours we associate with consciousness are conjugate to a description of the physical dynamics, by virtue of which the system sustains its existence.

If correct, these considerations suggest that virtual machines, including most implementations of artificial neural networks, cannot be conscious. Again, note that this does not "follow" from the FEP: it is possible that there are true mechanical theories of consciousness that conform to the FEP and that the proposal on offer in this paper is still false.

The rest of this paper is structured as follows. In section 2, I briefly explain how the FEP enables two conjugate descriptions of self-organising random dynamical systems: one in terms of the probabilistic evolution of a system's states or paths; the other in terms of the evolution of a probability density over states or paths. The latter type of description is provided by mechanical theories. In section 3, I discuss what a mechanical theory of consciousness would be. The aim in that section is not to formulate a mechanical theory of consciousness, but to specify, in general terms, under what additional assumptions such a theory is possible. In section 4, I consider how such a theory could help determine whether an artificial system is conscious. I argue that relatively strong constraints on the class of artificial conscious systems can be derived, under the assumption that a mechanical theory of consciousness exists. Since these constraints would exclude many types of artificial systems (e.g., most virtual machines, including simulations of neural networks on computers with a classical architecture, see section 5), I consider and discuss objections to this view in section 6.

2 The free energy principle and mechanical theories

Descriptions of the free energy principle (FEP) usually start with the notion of a random dynamical system—more specifically, with a stochastic differential equation of a certain form (Friston, 2019; Ramstead et al., 2022). Such an equation provides a probabilistic characterisation of the system's dynamics (i.e., of the evolution of the system's states over time). The charaterisation is probabilistic in that some paths through the system's state space are more likely than others.

The class of systems that the FEP applies to are *particular* random dynamical systems, viz. systems that can be partitioned into internal (μ) and external states (η), separated by a set of blanket states (b), comprising 'sensory' (s) and 'active' states (a). For such particular systems, the FEP enables a conjugate description of the dynamics of internal states. More specifically, the FEP asks: can we map internal states μ to a probability density q_μ over external states (given blanket states), in such a way that the dynamics of internal states can now be formulated in terms of the density q_μ? The answer provided by the FEP is 'yes' (see figure 1): the dynamics of q_μ (and thereby of μ) can be described as minimising variational free energy $\nabla_\mu F(s, a, \mu)$.

Perhaps the most interesting implication of this is that internal states can now be described as performing (approximative) Bayesian inference. In other words, such mechanical theories describe systems as if they implement approximatively Bayesian computations. Hence, these theories can be regarded as *Bayesian mechanics*.

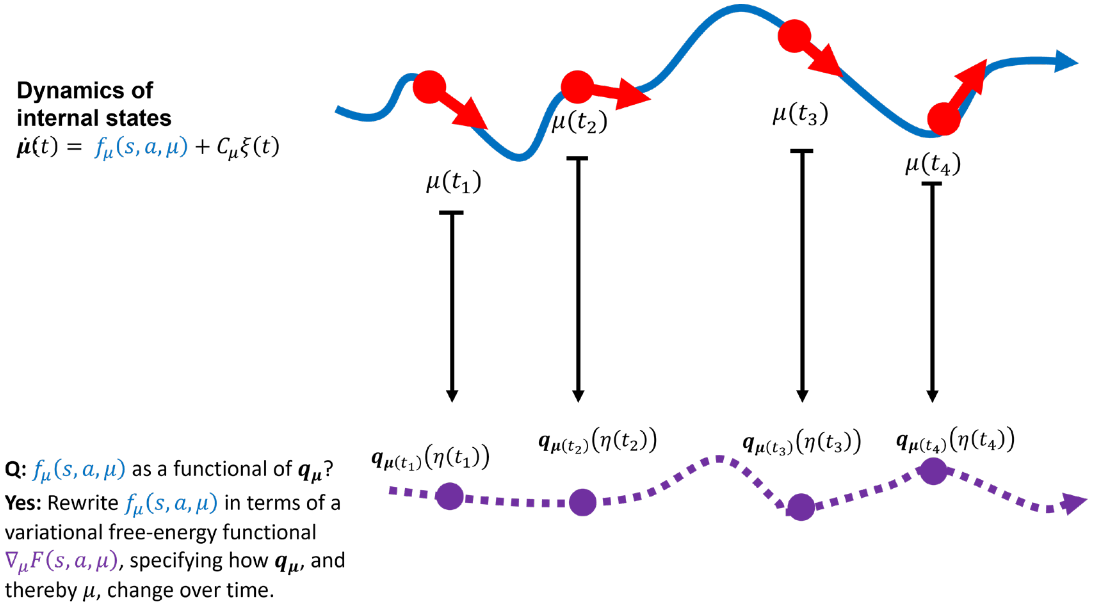

Dynamics of internal states
$\dot{\mu}(t) = f_\mu(s, a, \mu) + C_\mu \xi(t)$
$\mu(t_1)$
$\mu(t_2)$
$\mu(t_3)$
$\mu(t_4)$
$q_{\mu(t_1)}(\eta(t_1))$
$q_{\mu(t_2)}(\eta(t_2))$
$q_{\mu(t_3)}(\eta(t_3))$
$q_{\mu(t_4)}(\eta(t_4))$
Q: $f_\mu(s, a, \mu)$ as a functional of q_μ?
Yes: Rewrite $f_\mu(s, a, \mu)$ in terms of a variational free-energy functional $\nabla_\mu F(s, a, \mu)$, specifying how q_μ, and thereby μ, change over time.

Figure 1: The blue line depicts a trajectory of internal states $\mu(t)$. The flow of internal states is given by $f_\mu(s, a, \mu)$. By mapping internal states to a density q_μ over external states, the flow can be rewritten in terms of a variational free-energy functional $\nabla_\mu F(s, a, \mu)$.

This re-description of the system's internal dynamics in terms of Bayesian mechanics might seem like a trick, because the mapping from internal states μ to a density q_μ seems to be chosen arbitrarily in such a way as to enable a formulation in terms of minimising variational free energy $\nabla_\mu F(s, a, \mu)$. Is this just a fictional description, or do internal states *really* minimise variational free energy? The worry underlying this question may be that Bayesian mechanics seems to entail a form of pancomputationalism, or pan-Bayesianism (*everything* is Bayesian inference).

As a reply, we can note that the FEP does not apply to *everything*. There are different conditions under which system dynamics can be recast as Bayesian mechanics (and research on this is evolving, see Ramstead et al., 2022, for a recent account). Not all systems satisfy these descriptions (this is especially true for formulations that require the existence of a non-equilibrium steady-state density, see Aguilera et al., 2022). Furthermore, it does not imply that any systems perform all types of computation.

Since the FEP does not posit new entities or processes, but only provides a different view on processes that are already assumed to unfold, it should be regarded as a metaphysically neutral re-description, not as a substantial hypothesis about a system's internal states. As Jakob Hohwy (2021) puts it, the FEP *analyses* the concept of existence of particular self-organising systems.

At the same time, the FEP also provides a *normative* description:

> Many theories in the biological sciences are answers to the question: "what must things do, in order to exist?". The FEP turns this question on its head and asks: "if things exist, what must they do?" More formally, if we can define what it means to be something, can we identify the physics or dynamics that a thing must possess? (Friston, Da Costa, Sajid, et al., 2022, p. 2)

However, this does not mean that the FEP derives normative from mere descriptive claims. Instead, this only reflects the fact that the notion of existence of particular self-organising systems is itself a normative notion (Hohwy, 2021, p. 41).

Furthermore, the FEP does not entail what form the density q_μ must have. It only entails that it must approximate the probability of external states, given blanket states.

To sum up, the FEP shows that for certain classes of self-organising systems, there exist mechanical theories. These describe the system's behaviour and internal processes in terms of minimising variational free energy. Minimising variational free energy entails approximative Bayesian inference. Hence, such mechanical theories can be called *Bayesian mechanics*.

3 What would a mechanical theory of consciousness be?

If a mechanical theory can describe the dynamics of self-organising systems, they can also describe the dynamics of conscious systems. It is an open question what further conditions conscious systems fulfill, in addition to minimising variational free energy (Clark et al., 2019; for some suggestions, see, e.g, Friston, 2018; Friston et al., 2020; Safron, 2020). Variational free energy is minimised with respect to a probability density (a generative model), so it is plausible to assume that it must have certain features, such as being sufficiently deep, enabling counterfactual processing (Corcoran et al., 2020).

Regardless of which specific computational features are characteristic for consciousness, it must be possible to capture them in terms of minimising variational free energy, if the FEP applies to such systems. In principle, it may be that consciousness requires implementation in a particular (e.g., biological) substrate (Searle, 2017), or that it requires being alive (Froese, 2017). Furthermore it might be that a system can only be conscious if it conforms to organisational principles of life (Cosmelli & Thompson, 2010), and it may be that these principles are not entirely captured by current formulations of the FEP (Paolo et al., 2022; but see Friston, Da Costa, Sakthivadivel, et al., 2022).

For the sake of this paper, I will put these worries aside and assume the following: (1) Conscious systems can be described as random dynamical systems that conform to the FEP. (2) The difference between conscious and non-conscious systems can be captured in terms of features of the stochastic dynamics of conscious systems, and hence in terms of minimising variational free energy.

These assumption might seem relatively strong. However, the FEP is meant to apply to all self-organising systems, i.e., to dynamic systems that can be distinguished from their environments. Although some existing formulations of the FEP may make rather strong presuppositions about self-organising systems (as argued by Aguilera et al., 2022), more recent developments of the FEP strive for greater generality (e.g., Friston, Da Costa, Sajid, et al., 2022; Friston, Da Costa, Sakthivadivel, et al., 2022). Given these developments, it would be premature to conclude that (some) conscious systems do not conform to the FEP. This means assumption (1) is relatively innocuous.

Assumption (2) might seem stronger. If the difference between conscious and non-conscious systems can be captured by Bayesian mechanics, doesn't this presuppose the truth of computationalism about consciousness, i.e., the thesis that consciousness is a form of computation? A closer look reveals that it does not. In fact, it is even weaker than the assumption that there are computational correlates of consciousness in the sense of computational properties that are sufficient for consciousness (Cleeremans, 2005; Reggia et al., 2016). It only assumes that performing certain computations is necessary for consciousness (Wiese & Friston, 2021). It does not presuppose that implementing the right computations is sufficient for consciousness (as suggested by the 'thesis of computational sufficiency,' Chalmers, 2011).

Hence, rather than assuming that computation is all one needs to account for consciousness, the account advocated here assumes that the right computations must be implemented *in the right way*. This means that there is a difference between a mere simulation of a conscious system (which performs the right computations, but not in the right way) and an actually conscious system (which performs the right computations in the right way). What does this difference consist in?

4 What is the difference between an unconscious simulation and a conscious computational system?

Under the assumption that formulating a mechanical theory of consciousness is possible (section 3), we can express the internal dynamics of conscious systems in two conjugate ways (section 2). In other words, if we start with a description in terms of the probability of internal states (or paths), we can equivalently express the dynamics in terms of a probability encoded by internal states (or paths). In doing so, we move from a description of a physical system to a description of a computational system that minimises variational free energy, with respect to an internally encoded probability density (generative model). For the sake of simplicity, call the former the *physical dynamics*, and the latter the *computational dynamics*.

If the FEP is correct, then the physical dynamics uniquely specifies the computational dynamics. Crucially, the reverse does not hold. By mapping internal states (or paths) to a probability density, information about some physical details is lost. This assumption is justified by theorems such as the slaving principle (Haken, 1977/2012) or the center manifold theorem (Carr, 1971/2012; Davis, 2006).

According to these theorems, trajectories of self-organising systems that are not in equilibrium with their environment unfold in a relatively low-dimensional manifold, compared to their high-dimensional state space. In the brain, this means that the activity of neural population can be described in terms of their ensemble properties (e.g., statistical averages, Friston et al., 2020). Random fluctuations at the level of individual neurons can be averaged out, because they do not influence the behaviour of the ensemble (Palacios et al., 2020).

In particular, this means that a relatively coarse-grained description of the computational dynamics does not uniquely specify the underlying physical dynamics. In principle, one could implement the computational dynamics of a conscious organism in a computer simulation. There would thus be a level at which both activity in a conscious organism and in a computer can be described as implementing variational free energy minimisation. The underlying physical dynamics, however, would in general differ dramatically. Based on these difference, one can draw a distinction between an unconscious simulation and an actually conscious system—or so shall argue. Crucially, the argument does not presuppose that *no* computer simulation can be conscious (I do not endorse *non-computational functionalism*, Piccinini, 2020). I shall only argue that there will be very strong constraints on the class of conscious simulations (the underlying ideas were first presented in Wiese & Friston, 2021).

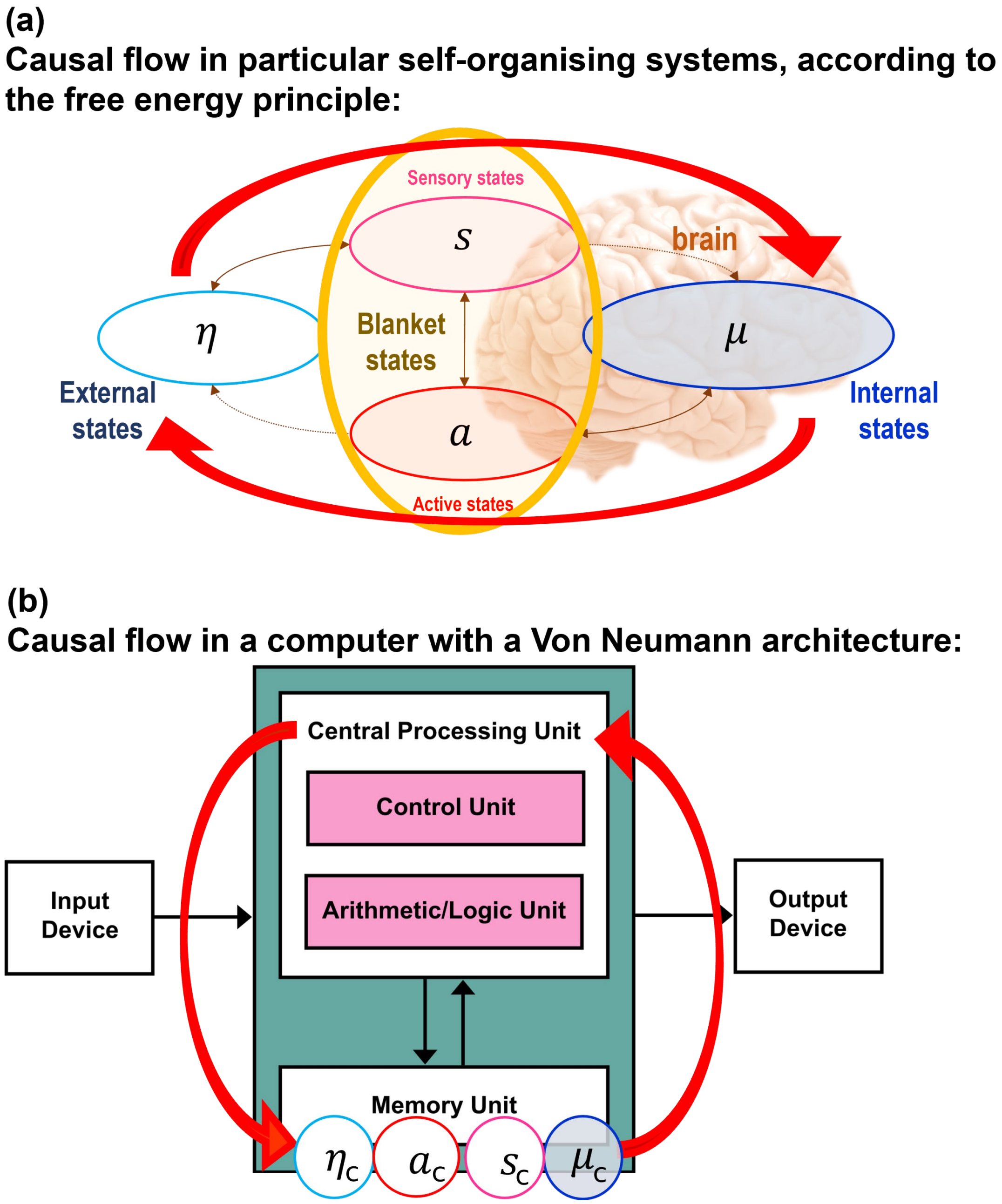
The distinction can most directly be spelled out in terms of the partition into internal and external states, separated by blanket states (i.e., perceptual and active states), because a description of the physical dynamics (as defined here) rests on such a partition. Crucially, the internal states (or paths) that figure in the description of the physical dynamics are numerically identical with the internal states that figure in the description of the conjugate computational dynamics. Now assume that we have reason to believe that some artificial system implements the computations that distinguish conscious from non-conscious systems, and that these can be described in terms of minimising variational free energy. This means the system must encode a probability density over some external states, given blanket states. Denote these external states with η_c, the blanket states with b_c, and the states that encode the probability density with μ_c. Here, the subscript "c" emphasises that these states are presupposed by the description of the **c**omputational dynamics—which, by assumption, applies to the dynamics of μ_c.

Next, we assume we also have a description of the *physical* dynamics of μ_c. If the system is conscious, μ_c will be part of a particular system with blanket states b that separate μ_c from external states η. Furthermore, these states must be numerically identical with the states that figure in the description of the computational dynamics. In other words, if the system is conscious, then b is identical to b_c and η is identical to η_c. As we will see in the next section, most computer simulations (in computers with a von Neumann architecture) do not satisfy this constraint.

5 Could a computer simulation in a computer with a von Neumann architecture be conscious?

In computers with a von Neumann architecture, the central processing unit (CPU) is separated from the memory unit, and the memory unit stores both programme instructions and data. Because of the separation between CPU and memory unit, any causal influence of one data element (stored in the memory unit) on another data element must always be mediated by the CPU. Even if there are further memory units within the CPU, causal relations between elements of those memory units will always be mediated by other parts of the CPU, as well. Since a computer simulation must store the values of μ_c, b_c (comprising s_c and a_c), and η_c in the memory unit, any causal relations between these representations is indirect, because it is mediated by the CPU.

This is in stark contrast to the causal flow between μ, b, and η in a system that conforms to the FEP. Hence, if μ_c is not just a representation of the internal state of a conscious system, but if μ_c is itself the internal state of a conscious system, then there must be direct causal relations between μ_c and blanket states b, as well as between these blanket states and external states η. Since there are no direct causal relations between μ_c and b_c, or between η_c and b_c (as argued above), we can conclude that η_c cannot be identical to η, and b_c cannot be identical to b. But this means there is a mismatch between the physical dynamics of μ_c and the computational dynamics represented by the computer simulation (using μ_c). More specifically, the probability density represented by μ_c in the computer simulation will not generally match the probability density encoded by μ_c, when it is regarded as part of a physical system constituted by μ_c, η, and b. See figure 2 for an illustration.



(a)
Causal flow in particular self-organising systems, according to the free energy principle:

(b)
Causal flow in a computer with a Von Neumann architecture:

Figure 2: (a) Basic causal flow (depicted by the red arrows) in a self-organising system that conforms to the free energy principle: there is a direct causal relation between blanket states and external states, as well as between internal states and blanket states. The causal coupling between internal and external states is mediated by blanket states. (b) Basic causal flow in a computer simulation in a computer with a van Neumann architecture: the values of internal, external, and blanket states are stored in memory units. Any causal interaction between them is always mediated by the central processing units. Hence, there is no direct causal interaction between blanket states and external or internal states. (The illustration of the von Neumann architecture has been adapted from https://en.wikipedia.org/wiki/Von_Neumann_architecture#/media/File:Von_Neumann_Architecture.svg, which was published under a CC BY-SA 3.0 license. The same license applies to the adapted illustration used here.)

6 Objections and replies

Objection 1: I understand that the basic causal flow in a von Neumann computer differs from the basic causal flow in particular systems that conform to the FEP. But is this not a rather arbitrary criterion for distinguishing conscious from non-conscious simulations?

I agree that, taken by itself, a difference in basic causal flow may seem arbitrary. However, this difference has further consequences that are not immediately obvious. The critical point is not merely that there is a difference in causal flow, but that there is a further difference, *because of this difference in causal flow*: there is a mismatch between the probability density encoded by internal states (or paths), according to the FEP, and the probability density that internal states need to encode, in order for the computer to perform the simulations.

This has further implications. Since the FEP analyses the concept of the existence of particular self-organising systems (Hohwy, 2021), it follows that being such a system entails minimising variational free energy. The computations such a system performs, which contribute to minimising variational free energy, therefore contribute to the sustained existence of the system. Put differently, the system exists *by virtue of* performing those computations. Contrast this with a simulation in a von Neumann computer, which performs the same computations by representing the system's states in its memory and by updating these representations in accordance with rules that specify how to minimise variational free energy. The relevant parts of the memory unit do not exist by virtue of their role in these computations.

In fact, this enables a distinction between two notions of computation: intrinsic and extrinsic computation. A system performs *intrinsic computations* if it exists by virtue of performing these computations, i.e., if the computations contribute to the sustained existence of the system. A system performs *extrinsic computations*, if the system's existence is not contingent on performing these computations, i.e., if the computations do not contribute to the sustained existence of the system. Only computational systems that perform intrinsic computations are able to *give a damn* (Haugeland, 2000).

Objection 2: What if the computer simulates a virtual environment, in which virtual agents exist by virtue of performing computations specified by a mechanical theory of consciousness? Don't these virtual agents exist by virtue of performing these computations?

According to this objection, the simulated agents perform intrinsic computations. However, the computer in which they are simulated performs extrinsic computations: neither it, nor any of its parts, exist by virtue of performing the computations that are involved in simulating agents in a virtual environment. Of course, one could reject the characterisation of intrinsic computation given above. For instance, what if one assumes a robust mapping account of physical computation,[1] according to which implementing a computation only requires the right

[1] A physical computation is a computation performed by a physical system. Accounts of physical computation specify under what conditions a physical system computes, and thereby help clarify how systems can realise capacities by performing computations. Such accounts can be

dispositional (Klein, 2008), causal (Chalmers, 2011; Chrisley, 1994), and/or counterfactual (Chalmers, 1996; Maudlin, 1989) structure (for an overview, see Piccinini & Maley, 2021, sec. 2.2)? Based on such accounts, one could argue that all instances of computation in physical systems are intrinsic computations. But in that case, it seems that performing the right computations would be sufficient for instantiating consciousness (Chalmers, 2022).

I see two reasons why adding further constraints on accounts of physical implementation is warranted in the case of computational accounts of consciousness. The first is that the possibility of certain strange implementations of conscious processing can be avoided. The second is related to functions of consciousness.

(1) Apart from strange systems known from notorious thought experiments (such as Block's homunculus head thought experiment, Block, 1978), a purely mechanical computer with cogs and wheels might be used to implement computational correlates of consciousness. Because of its mechanical nature, it would be much slower than an electronic computer (or a human brain). Depending on the exact implementation, it could even take years to perform the computations required to instantiate a single conscious thought (a similar point is made in Dennett, 1978/2017). To me it seems absurd to regard such a system as conscious—although I know opponents could simply deny that the time scale on which a system operates makes a difference (Block, 1978, p. 280). One way of arguing that time scales matter is related to functions of consciousness, i.e., to the second reason for adding constraints on accounts of physical computation.

(2) If consciousness is a trait that has been selected for, it must somehow impact behaviour. But then consciousness must operate on a time scale that is appropriate for the system's environment. Of course, a trait that has been selected for in some species can be implemented in systems for which it does not have a functional role. Or we could imagine an artificial system that only wants to consciously count prime numbers. If consciousness contributes to this goal, it has a function for the system, but the time scale at which the system operates is irrelevant.

To exclude these possibilities, some further assumptions have to be made. Firstly, I shall assume that consciousness matters (Cleeremans & Tallon-Baudry, 2022). The proposal on offer here is that conscious processes in a system matter because they play a functional role for the system. In particular, they contribute to the sustained existence of the system. If a physical system instantiates computational correlates of consciousness, but implementing these computations does not fulfill a functional role for the system, then it is not conscious by virtue of performing these computations—because the computations do not have a function *for the system*.

Secondly, I shall assume that this function must potentially have an impact on the system's outward behaviour. I say "potentially", so as not to rule out the possibility of so-called "islands of awareness" (Bayne et al., 2020), i.e., conscious systems that neither receive sensory input nor produce motor output. For such systems, the assumption that a function of consciousness must

distinguished from mathematical accounts of abstract computation, which are concerned with, e.g., computability and computational complexity.

potentially impact outward behaviour means that consciousness must affect the system's motor output, if it were connected to a motor system.

For instance, assume that a function of consciousness is to support fast-acting, affectively guided motor commands (Black, 2021, p. 211; Feinberg & Mallatt, 2016). Assume, furthermore, that islands of awareness are possible. Obviously, consciousness cannot fulfill the functional role of guiding motor commands in an island of awareness. However, if hooked up to a fitting motor system, conscious processing within this augmented island of awareness might support motor control. Crucially, in order to have a function *for the system*, motor control must operate at an appropriate time scale. This does not exclude the possibility of conscious systems that operate at extremely slow time scales, but it puts strong constraints on such systems.

The two assumptions entail that there is a distinction between systems that have the function to compute and systems that do not have that function. Above, I argued that the FEP suggests a non-arbitrary distinction between intrinsic and extrinsic computations: a system performs *intrinsic computations* if it exists by virtue of performing these computations, i.e., if the computations contribute to the sustained existence of the system. A system performs *extrinsic computations*, if the system's existence is not contingent on performing these computations, i.e., if the computations do not contribute to the sustained existence of the system.

I shall now argue that another way of describing the same distinction is in terms of teleological function: a system performs intrinsic computations just in case performing these computations have a teleological function for the system. There are different accounts of teleological function. The idea at hand can be cashed out in terms of a modified *goal-contribution* account of teleological function (Piccinini, 2020, ch. 3). This notion of teleological function is particularly suited to the FEP, because systems that minimise variational free energy have a teleological function in this sense (for a related argument in terms of a selectionist account of teleological function, see Mann & Pain, 2022). Furthermore, there are independent reasons for preferring this notion of teleological function over other, weaker notions (Piccinini, 2020, ch. 3).

Ignoring a few complications,[2] Piccinini (2020) characterises teleological functions in his goal-contribution account as follows (see also Maley & Piccinini, 2017): “Tokens of type X have function F if and only if F is a causal role and performing F by tokens of X provides a regular contribution to a goal of organisms.” (Piccinini, 2020, p. 76).

The goals of organisms include biological goals, such as staying alive and reproducing (Piccinini, 2020, p. 68). The account can not only be applied to functional mechanisms within organisms, but also to non-living computing devices that can be *used* by organisms (and thereby make a regular contribution to the goals of organisms, Piccinini, 2020, p. 71). However, in the case of artefacts and other tools, having the function to compute is an extrinsic property: whether or

[2] In order to distinguish malfunction from proper function, it is necessary to specify what it means to perform a function at appropriate rates in appropriate situations by appropriate members of a population (Piccinini, 2020, sec. 3.3). In what follows, I assume that similar considerations can be applied to systems that conform to FEP.

not a device has this function depends on whether it makes a regular contribution to the goals of organisms, by virtue of performing computations. In other words, having the *function* to compute depends not just on properties of the computing device, but also on its relations to organisms.

This means we cannot directly deploy the account to determine whether a computer merely *simulates* or actually *instantiates* consciousness. The reason is that being conscious is an intrinsic property (as I shall assume), that is, whether or not a system is conscious does not depend on relations to other systems. Hence, if having the *function* to implement computational correlates of consciousness is not an intrinsic property, it cannot be what distinguishes simulations from genuine conscious systems. We have to look for intrinsic properties to spell out this difference.

Fortunately, it is straightforward to adapt the goal-contribution account of teleological function to any system that conforms to the FEP.[3] Systems that conform to the FEP need not be alive. Hence, not all systems that conform to the FEP have biological goals. However, we can stipulate that they have the goal of sustaining their existence (i.e., the goal of maintaining the structure that defines what kind of system they are). Since the FEP provides an analysis of the concept of existence of particular self-organising systems (Hohwy, 2021), the FEP specifies what systems do that successfully pursue the goal of sustaining their existence: they minimise variational free energy. Hence, any computational process (in such a system) that contributes to minimising variational free energy thereby contributes to this goal. We can therefore say: mechanisms in a system that regularly implement such computations make a regular contribution to a goal of that system, and therefore *have the function* to compute. In other words, such mechanisms perform *intrinsic* computations.

Objection 3: If a non-conscious computer simulation can perform the same computations as a conscious system, does this mean consciousness is epiphenomenal?

It may seem as if the proposal at hand entails that consciousness does not play a functional role after all: if a virtual machine, for instance, can instantiate the same computational properties, without being conscious, does consciousness really make a causal difference?

Above, I suggested that performing intrinsic computations requires that the computing mechanism have the function to compute (and it must have that function *for* the system of which it is a part). This not only means that the computing mechanism plays a causal role, it means that it plays a causal role *for the system*. In this sense, there is a causal difference.

We can imagine that a non-conscious computer simulation is connected to a robot, receiving sensory input and producing motor output (including speech). Unless the robot's physical dynamics matches its computational dynamics, the robot won't be conscious, according to the proposal at hand. In particular, this means the robot might have a wide range of cognitive

[3] Just as Mann & Pain (2022) adapt Millikan's (1984) selectionist account of teleological function to systems conforming to the FEP.

capacities that may require consciousness in human beings (such as various learning abilities, Birch et al., 2020; Birch, 2022; Kanai et al., 2019). Consciousness may be part of what explains these capacities in human beings; but this leaves it open that the same cognitive capacities can be realised without consciousness ("conscious inessentialism," Flanagan, 1993).

One might object: if a functionally isomorphic robot can realise the same causal roles, without being conscious, doesn't this mean that consciousness *is* epiphenomenal? After all, such a robot would be similar to a philosophical zombie. Consciousness does not play a causal role in the behaviour and internal goings on of zombies. But if such systems are metaphysically possible, what causal role does consciousness play in systems like us?

I see two possible replies. The first is to bite the bullet and insist that a robot could realise the same functional states without being conscious. The difference to conscious systems would not be analysable in terms of any causal roles. The difference would be merely metaphysical: in systems like us, mechanisms that realise certain functions would, in addition, also be conscious. This can be problematic if one assumes that consciousness gives systems a special moral status (Shepherd, 2018): the difference between a simulation and an actually conscious system would then seem so elusive that it would be hard to explain why conscious makes a moral difference (of course, some accounts of robot ethics downplay the relevance of consciousness anyway, Gunkel, 2022).

The second reply is to bet that there will be some functional differences, due to the mismatch between the physical and the computational dynamics. This functional difference may then also account for why consciousness matters morally: it has a function *for the physical system*, and that's what gives consciousness its moral relevance. I prefer this reply, although I realise that it entails a strong empirical hypothesis (which is also a good thing, from a methodological point of view): it entails that even highly detailed and accurate computer simulations of conscious agents in a natural environment would fail to reliably control a (robotic) agent in the real world. This points to another way of describing how the functional roles played by computer simulations differ from the roles played by actually conscious systems: unconscious implementations of the computations required for consciousness might play the same functional role with respect to a virtual environment, but not with respect to the physical world.[4]

Let us unpack this idea a little. Imagine a digital avatar that simulates a conscious being and exists in a virtual environment. Furthermore, imagine that this virtual entity can upload itself to

[4] One might object (perhaps inspired by Chalmers, 2022): "How do *you* know you are not a simulated agent in a virtual environment?" Of course, I cannot rule this possibility out. However, it is at least consistent with the proposal at hand. The account on offer in this paper may be true, even though I am a simulated agent in a computer simulation. It only requires that the computations, by virtue of which I flexibly interact with my environment, would also enable an agent in the next world up to flexibly interact with the environment. As argued in the next two paragraphs in the main text, this puts a very strong constraint on the type of computer simulation that I could be in.

a physical robot and can then act in our physical environment just as flexibly and smoothly as it could in the virtual environment (as in Ted Chiang's story „The lifecycle of software objects", Chiang, 2010). For this to be possible, the simulated (virtual) environment would have to be remarkably complex and detailed. In particular, the virtual robot would have to be a highly accurate simulation of the physical robot. For instance, it would have to be so accurate that learning a new motor skill in the virtual world would enable the system to exhibit the new skill just as smoothly in the physical world. Given the many non-linearities in physical dynamics, it is extremely difficult to accurately simulate these details to the required degree.

In fact, the second reply entails a stronger claim: it will be *impossible* to simulate them accurately enough. This is a very strong empirical claim, especially given that the "sim-to-real" strategy of using simulations for learning, and then transferring the results to the real world, has already been successful to some extent (Christiano et al., 2016; Muratore et al., 2022; Rusu et al., 2017). However, while this strategy may work for basic actions, learning fast *and flexible* perception and action (of the sort that seems to benefit from consciousness) might require feedback from the real world. More precisely, implementing such learning in a simulation might be nomologically possible, but technologically impossible.

Although this is a strong empirical claim, it does not rule out the possibility of conscious robots. If the robot's computational dynamics are equivalent to the robot's physical dynamics, in the sense that the robot sustains its existence by virtue of instantiating computational correlates of consciousness, then it is conscious, according to the proposal at hand. Furthermore, it may be possible to instantiate real consciousness in a computer simulation if the simulation is implemented using non-classical hardware.

7 Conclusion

I have made a proposal of how to distinguish a mere computer simulation from an actually conscious system. Although the proposal is based on considerations that follow from the free energy principle, the proposed distinction between merely simulating and actually instantiating consciousness (weak vs. strong artificial consciousness) does not follow from the free energy principle itself. In particular, the following additional assumptions are required:

- Consciousness makes a causal contribution. Its causal roles can be captured in terms of computation; that is, the functional roles of consciousness are *medium-independent*. They are the computational correlates of consciousness.
- Consciousness not only has a function, but it necessarily has a function *for* the physical system that instantiates consciousness. This means that consciousness must contribute to the goals (e.g., sustained existence) of the physical system. Consequently, some systems may instantiate the computational correlates of consciousness without being conscious, if the respective computational mechanisms do not contribute to the goals of the physical system of which they are a part.

The free energy principle does not justify these assumptions. Hence, at least some of the "heavy metaphysical lifting" (Bruineberg et al., 2022, p. 15) in this account is not done by the free energy principle itself. However, the free energy principle still plays a central role, in that it offers a conceptually clear way of analysing relevant metaphysical concepts (Wiese, 2022), which also highlights metaphysical implications of basic assumptions. Here, the FEP helps to specify:

- what it means to contribute to the goals of particular self-organising systems and
- what it means for a mechanism to have the *intrinsic* function to compute (i.e., a function for the system itself, not for organisms that have designed a computational device for some purpose).

If the proposal on offer here is on the right track, there are many lines of research that will never lead to the creation of artificial consciousness. In particular, this concerns all systems implemented in computers with a von Neumann architecture. Large-scale computer simulations of consciousness, realised using such architectures, will therefore be unlikely to be conscious. The risk of creating artificial consciousness, and perhaps artificial pain and suffering (Metzinger, 2021), will therefore be low. Furthermore, some tests for artificial consciousness, such as Susan Schneider's "AI Consciousness Test" (Schneider, 2019), would fail to provide sufficient evidence for consciousness in systems that lack the right architecture.

However, given the moral relevance of pain and suffering, and the amount of uncertainty associated with the conclusions reached here, the risk of inadvertently creating artificial consciousness in computer simulations should still be taken seriously. Rather than using the considerations in this paper as an excuse for dismissing any moral concerns about artificial consciousness whatsoever, they should be taken as an invitation to scrutinise the assumptions I made to reach these conclusions.

On the one hand, it would be relevant to see if further support can be drawn from, for instance, accounts of artificial consciousness that stress the importance of embodiment using soft robotics (Man & Damasio, 2019) to create vulnerable systems that thereby have intrinsic interests and goals (Bronfman et al., 2021).

On the other hand, it would be relevant to see if strong objections can be given, for instance by considering variations of conscious systems and determining whether they would be conscious or not, according to the proposal at hand (in the spirit of the substitution argument given in Kleiner & Hoel, 2021).[5] This would help determine whether the proposal has any problematic implications.

More generally, it would be insightful to contrast the overarching strategy with other approaches. One possible approach to artificial consciousness asks: what must be added to existing AI systems to make them conscious (Chalmers, n.d.; Graziano, 2017; Juliani et al., 2022)? Another asks: what types of AI systems will never be, or are unlikely to become, conscious (Piccinini, 2021; Tononi & Koch, 2015)? The first approach faces the problem that

[5] For instance, an interesting special case may be artificial systems with holographic bodies, such as Joi in the film *Blade Runner 2049* (Wiese & Metzinger, 2019).

answering its question may provide the means to create artificial consciousness, before we know under what conditions this would be morally permissable (Agarwal & Edelman, 2020; Metzinger, 2021). The second approach avoids this problem, but falsely assuming that certain types of artificial systems will never be conscious, while being confident that this assumption is true, might lead to the inadvertent creation of artificial consciousness (and pain and suffering). Accounts like the one presented in this paper, which make strong claims about the impossibility of consciousness in large classes of artificial systems, should therefore always be taken with a grain of salt.

Acknowledgments: I am grateful to Maxwell Ramstead for feedback on an earlier version of this paper.

References

Agarwal, A., & Edelman, S. (2020). Functionally effective conscious AI without suffering. *Journal of Artificial Intelligence and Consciousness*, *7*(01), 39–50. https://doi.org/10.1142/S2705078520300030

Aguilera, M., Millidge, B., Tschantz, A., & Buckley, C. L. (2022). How particular is the physics of the free energy principle? *Physics of Life Reviews*, *40*, 24–50. https://doi.org/10.1016/j.plrev.2021.11.001

Bayne, T., Seth, A. K., & Massimini, M. (2020). Are there islands of awareness? *Trends in Neurosciences*, *43*(1), 6–16. https://doi.org/10.1016/j.tins.2019.11.003

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

Birch, J., Ginsburg, S., & Jablonka, E. (2020). Unlimited associative learning and the origins of consciousness: A primer and some predictions. *Biology & Philosophy*, *35*(6), 56. https://doi.org/10.1007/s10539-020-09772-0

Black, D. (2021). Analyzing the etiological functions of consciousness. *Phenomenology and the Cognitive Sciences*, *20*(1), 191–216. https://doi.org/10.1007/s11097-020-09693-z

Block, N. (1978). Troubles with functionalism. In W. Savage (Ed.), *Readings in philosophy of psychology* (pp. 261–325). Harvard University Press.

Bronfman, Z., Ginsburg, S., & Jablonka, E. (2021). When will robots be sentient? *Journal of Artificial Intelligence and Consciousness*, *08*(02), 183–203. https://doi.org/10.1142/S2705078521500168

Bruineberg, J., Dolega, K., Dewhurst, J., & Baltieri, M. (2022). The emperor's new markov blankets. *Behavioral and Brain Sciences*, 1–63. https://doi.org/10.1017/S0140525X21002351

Carr, J. (2012). *Applications of centre manifold theory*. Springer Science & Business Media. (Original work published 1971)

Chalmers, D. J. (n.d.). *Could a large language model be conscious?* https://philarchive.org/archive/CHACAL-3

Chalmers, D. J. (1996). Does a rock implement every finite-state automaton? *Synthese*, *108*(3), 309–333. https://doi.org/10.1007/bf00413692

Chalmers, D. J. (2011). A computational foundation for the study of cognition. *Journal of Cognitive Science*, *12*, 323–357. https://oak.go.kr/central/journallist/journaldetail.do?article_seq=18672

Chalmers, D. J. (2022). *Reality+: Virtual worlds and the problems of philosophy* (First edition). W. W. Norton & Company.

Chiang, T. (2010). *The lifecycle of software objects*. Subterranean Press.

Chrisley, R. L. (1994). Why everything doesn't realize every computation. *Minds and Machines*, *4*(4), 403–420. https://doi.org/10.1007/BF00974167

Christiano, P., Shah, Z., Mordatch, I., Schneider, J., Blackwell, T., Tobin, J., Abbeel, P., & Zaremba, W. (2016). *Transfer from simulation to real world through learning deep inverse dynamics model*. *arXiv:1610.03518*. https://doi.org/10.48550/arXiv.1610.03518

Clark, A., Friston, K., & Wilkinson, S. (2019). Bayesing qualia: Consciousness as inference, not raw datum. *Journal of Consciousness Studies*, *26*(9–10), 19–33.

Cleeremans, A. (2005). Computational correlates of consciousness. *Progress in Brain Research*, *150*, 81–98. https://doi.org/10.1016/S0079-6123(05)50007-4

Cleeremans, A., & Tallon-Baudry, C. (2022). Consciousness matters: Phenomenal experience has functional value. *Neuroscience of Consciousness*, *2022*(1), niac007. https://doi.org/10.1093/nc/niac007

Corcoran, A. W., Pezzulo, G., & Hohwy, J. (2020). From allostatic agents to counterfactual cognisers: Active inference, biological regulation, and the origins of cognition. *Biology & Philosophy*, *35*(3), 32. https://doi.org/10.1007/s10539-020-09746-2

Cosmelli, D., & Thompson, E. (2010). Embodiment or envatment?: Reflections on the bodily basis of consciousness. In J. Stewart, O. Gapenne, & E. A. Di Paolo (Eds.), *Enaction: Towards a new paradigm for cognitive science* (pp. 361–385). MIT Press.

Davis, M. J. (2006). Low-dimensional manifolds in reaction−diffusion equations. 1. Fundamental aspects. *The Journal of Physical Chemistry A*, *110*(16), 5235–5256. https://doi.org/10.1021/jp055592s

Dennett, D. C. (2017). Toward a cognitive theory of consciousness. In *Brainstorms* (pp. 163–188). MIT Press. (Original work published 1978)

Feinberg, T. E., & Mallatt, J. M. (2016). *The ancient origins of consciousness: How the brain created experience*. The MIT Press. https://www.jstor.org/stable/j.ctt1bkm52m

Flanagan, O. J. (1993). *Consciousness reconsidered*. MIT Press.

Friston, K. (2018). Am I self-consciousn? *Frontiers in Psychology*, *9*, 579. https://doi.org/10.3389/fpsyg.2018.00579

Friston, K. (2019). *A free energy principle for a particular physics*. https://arxiv.org/abs/1906.10184

Friston, K. J., Wiese, W., & Hobson, J. A. (2020). Sentience and the origins of consciousness: From Cartesian duality to Markovian monism. *Entropy*, *22*(5), 516. https://doi.org/10.3390/e22050516

Friston, K., Da Costa, L., Sajid, N., Heins, C., Ueltzhöffer, K., Pavliotis, G. A., & Parr, T. (2022). *The free energy principle made simpler but not too simple*. https://doi.org/10.48550/ARXIV.2201.06387

Friston, K., Da Costa, L., Sakthivadivel, D. A. R., Heins, C., Pavliotis, G. A., Ramstead, M., & Parr, T. (2022). *Path integrals, particular kinds, and strange things*. https://doi.org/10.48550/ARXIV.2210.12761

Froese, T. (2017). Life is precious because it is precarious: Individuality, mortality and the problem of meaning. In G. Dodig-Crnkovic & R. Giovagnoli (Eds.), *Representation and reality in humans, other living organisms and intelligent machines* (pp. 33–50). Springer International Publishing. https://doi.org/10.1007/978-3-319-43784-2_3

Graziano, M. S. A. (2017). The attention schema theory: A foundation for engineering artificial consciousness. *Frontiers in Robotics and AI*, *4*, 60. https://doi.org/10.3389/frobt.2017.00060

Gunkel, D. J. (2022). The relational turn: Thinking robots otherwise. In J. Loh & W. Loh (Eds.), *Social robotics and the good life: The normative side of forming emotional bonds with robots* (pp. 55–76). transcript-Verlag.

Haken, H. (2012). *Synergetics: An introduction nonequilibrium phase transitions and self-organization in physics, chemistry and biology*. Springer Science & Business Media. (Original work published 1977)

Haugeland, J. (2000). *Having thought: Essays in the metaphysics of mind*. Harvard University Press.

Hohwy, J. (2021). Self-supervision, normativity and the free energy principle. *Synthese*, *199*(1), 29–53. https://doi.org/10.1007/s11229-020-02622-2

Holland, O. (2003). *Machine consciousness*. Imprint Academic.

Juliani, A., Arulkumaran, K., Sasai, S., & Kanai, R. (2022). On the link between conscious function and general intelligence in humans and machines. *Transactions on Machine Learning Research*, 38.

Kanai, R., Chang, A., Yu, Y., Magrans de Abril, I., Biehl, M., & Guttenberg, N. (2019). Information generation as a functional basis of consciousness. *Neuroscience of Consciousness*, *2019*(1). https://doi.org/10.1093/nc/niz016

Klein, C. (2008). Dispositional implementation solves the superfluous structure problem. *Synthese*, *165*(1), 141–153. https://doi.org/10.1007/s11229-007-9244-z

Kleiner, J., & Hoel, E. (2021). Falsification and consciousness. *Neuroscience of Consciousness*, *2021*(niab001). https://doi.org/10.1093/nc/niab001

Maley, C. J., & Piccinini, G. (2017). A unified mechanistic account of teleological functions for psychology and neuroscience. In D. M. Kaplan (Ed.), *Explanation and integration in mind and brain science* (pp. 236–257). Oxford University Press.

Man, K., & Damasio, A. (2019). Homeostasis and soft robotics in the design of feeling machines. *Nature Machine Intelligence*, *1*(10), 446–452. https://doi.org/10.1038/s42256-019-0103-7

Mann, S. F., & Pain, R. (2022). Teleosemantics and the free energy principle. *Biology & Philosophy*, *37*(4), 34. https://doi.org/10.1007/s10539-022-09868-9

Maudlin, T. (1989). Computation and consciousness. *The Journal of Philosophy*, *86*(8), 407–432. https://doi.org/https://www.jstor.org/stable/2026650

Metzinger, T. (2021). Artificial suffering: An argument for a global moratorium on synthetic phenomenology. *Journal of Artificial Intelligence and Consciousness*, 1–24. https://doi.org/10.1142/S270507852150003X

Millikan, R. G. (1984). *Language, thought, and other biological categories: New foundations for realism*. MIT Press.

Muratore, F., Ramos, F., Turk, G., Yu, W., Gienger, M., & Peters, J. (2022). Robot learning from randomized simulations: A review. *Frontiers in Robotics and AI*, *9*, 799893. https://doi.org/10.3389/frobt.2022.799893

Palacios, E. R., Razi, A., Parr, T., Kirchhoff, M., & Friston, K. (2020). On Markov blankets and hierarchical self-organisation. *Journal of Theoretical Biology*, *486*, 110089. https://doi.org/10.1016/j.jtbi.2019.110089

Paolo, E. D., Thompson, E., & Beer, R. (2022). Laying down a forking path: Tensions between enaction and the free energy principle. *Philosophy and the Mind Sciences*, *3*, 2. https://doi.org/10.33735/phimisci.2022.9187

Parr, T., Pezzulo, G., & Friston, K. J. (2022). *Active inference: The free energy principle in mind, brain, and behavior*. MIT Press. https://doi.org/10.7551/mitpress/12441.001.0001

Piccinini, G. (2020). *Neurocognitive mechanisms*. Oxford University Press.

Piccinini, G. (2021). The myth of mind uploading. In R. W. Clowes, K. Gärtner, & I. Hipólito (Eds.), *The mind-technology problem* (Vol. 18, pp. 125–144). Springer International Publishing. https://doi.org/10.1007/978-3-030-72644-7_6

Piccinini, G., & Maley, C. (2021). Computation in physical systems. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2021). https://plato.stanford.edu/archives/sum2021/entries/computation-physicalsystems/; Metaphysics Research Lab, Stanford University.

Ramstead, M. J. D., Sakthivadivel, D. A. R., Heins, C., Koudahl, M., Millidge, B., Da Costa, L., Klein, B., & Friston, K. J. (2022). *On Bayesian mechanics: A physics of and by beliefs*. *arXiv:2205.11543*. https://doi.org/10.48550/arXiv.2205.11543

Reggia, J. A., Katz, G., & Huang, D.-W. (2016). What are the computational correlates of consciousness? *Biologically Inspired Cognitive Architectures*, *17*, 101–113. https://doi.org/10.1016/j.bica.2016.07.009

Rusu, A. A., Večerík, M., Rothörl, T., Heess, N., Pascanu, R., & Hadsell, R. (2017). Sim-to-real robot learning from pixels with progressive nets. *Proceedings of the 1st Annual Conference on Robot Learning*, 262–270. https://proceedings.mlr.press/v78/rusu17a.html

Safron, A. (2020). An integrated world modeling theory (IWMT) of consciousness: Combining integrated information and global neuronal workspace theories with the free energy principle and active inference framework; toward solving the hard problem and characterizing agentic causation. *Frontiers in Artificial Intelligence*, *3*. https://www.frontiersin.org/article/10.3389/frai.2020.00030

Schneider, S. (2019). *Artificial you: AI and the future of your mind*. Princeton University Press.

Searle, J. (2017). Biological naturalism. In *The Blackwell companion to consciousness* (pp. 327–336). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119132363.ch23

Shepherd, J. (2018). *Consciousness and moral status*. Routledge. https://doi.org/10.4324/9781315396347

Tononi, G., & Koch, C. (2015). Consciousness: Here, there and everywhere? *Philosophical Transactions of the Royal Society B*, *370*, 20140167. https://doi.org/10.1098/rstb.2014.0167

Wiese, W. (2022). Does the metaphysical dog wag its formal tail? The free energy principle and philosophical debates about life, mind, and matter. *Behavioral and Brain Sciences*, *45*, 64–65. https://doi.org/10.1017/S0140525X22000292

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*, 9. https://doi.org/10.33735/phimisci.2021.81

Wiese, W., & Metzinger, T. (2019). Androids dream of virtual sheep. In T. Shanahan & P. Smart (Eds.), *Blade runner 2049* (pp. 149–164). Routledge.