Philosophical Studies (2024) 181:1947–1970
https://doi.org/10.1007/s11098-024-02182-y

Artificial consciousness: a perspective from the free energy principle

Wanja Wiese[1]

Accepted: 12 June 2024 / Published online: 26 June 2024

Abstract
Does the assumption of a weak form of computational functionalism, according to which the right form of neural computation is sufficient for consciousness, entail that a digital computational simulation of such neural computations is conscious? Or must this computational simulation be implemented in the right way, in order to replicate consciousness?

From the perspective of Karl Friston's free energy principle, self-organising systems (such as living organisms) share a set of properties that could be realised in artificial systems, but are not instantiated by computers with a classical (von Neumann) architecture. I argue that at least one of these properties, viz. a certain kind of causal flow, can be used to draw a distinction between systems that merely simulate, and those that actually replicate consciousness.

Keywords Artificial consciousness · Computationalism · Consciousness · Free energy principle · Simulation

1 Introduction

We live in times in which some smart people believe that at least a few existing artificial intelligences (AIs) are conscious. How can we assess such views? Do we need a theory of phenomenal consciousness[1]? A recent report (Butlin et al., 2023) draws

[1] In what follows, I will mainly use the term "consciousness" to refer to phenomenal consciousness (i.e., subjective experience / having states for which it is something like to be in, Farrell, 1950; Nagel, 1974).

✉ Wanja Wiese
wanja.wiese@rub.de

[1] Institute for Philosophy II, Ruhr University Bochum, Building GAFO 04 / 424
Universitätsstr. 150, 44801 Bochum, Germany

on theories of consciousness to assess the likelihood of consciousness in existing and future AIs. A limitation of this report is that it explicitly assumes computational functionalism, according to which performing the right computations is sufficient (and necessary) for consciousness. As the authors of the report point out, this position is controversial (Butlin et al., 2023, p. 4). Alternatives include non-computational functionalism (Piccinini, 2020; Prinz, 2012), as well as non-functionalist positions. In the current debate about the possibility of conscious AIs, such alternative views have motivated the suggestion that there may be principled reasons why many types of artificial systems *cannot* be conscious (Aru et al., 2023; Kleiner & Ludwig, 2023; LeDoux et al., 2023).

Here, I shall only assume a weak form of computationalism, according to which there are, *in living organisms*, computational correlates of consciousness (Cleeremans, 2005; Reggia et al., 2016; Wiese & Friston, 2021) that (partly) explain consciousness. This form of computationalism is weak, because it does not assume that performing the right computations is nomologically sufficient for being conscious. Neither does it presuppose that conscious experience is identical to computation, or that computation is the metaphysical ground of consciousness. It only assumes that computation is nomologically sufficient for consciousness *in living organisms*. Artificial systems that perform (or simulate) the same computations might not be conscious. That is, to replicate consciousness in AI, you might need to implement the right computations + X.

I argue that the free energy principle (FEP) (Friston, 2019; Parr et al., 2022) suggests an account of what this additional "X" might be. If correct, the FEP 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 of beliefs[2] that conform to the FEP. In physics, a mechanical theory is a theory that describes the relationship between motion of matter and forces bringing about such motion. A mechanical theory *of beliefs* describes the 'forces' acting on internally encoded probability distributions (beliefs) in terms of a variational free energy gradient (more on this in Sect. 2 below). A key feature of *Bayesian mechanics* (Ramstead et al., 2023) is that it provides conjugate descriptions of a system's physical dynamics and the dynamics of beliefs, i.e., an internal and an external perspective on the same dynamics (Friston et al., 2020).

Ideally, it may be possible to make a mechanical theory 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 (at least in living organisms).

Crucially, this does not mean all systems performing the computations specified by that theory are conscious: a mere simulation of a conscious system may implement the right computations without being conscious. The FEP does not entail an account of the difference between simulating and replicating consciousness, but it can be used

[2] Here, a belief is just a probability distribution over the system's (external) states.

to highlight a set of properties that self-organising systems (such as living organisms) share, and that are not instantiated by large classes of artificial systems (e.g., computers with a von Neumann architecture). I argue that at least one of these properties, viz. a certain kind of causal flow, can be used to draw a distinction between systems that merely simulate, and those that actually replicate consciousness.[3]

The rest of this paper is structured as follows. In Sect. 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 Sect. 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 Sect. 4, I consider a criterion (the "FEP Consciousness Criterion", FEP2C) that is satisfied by conscious living organisms. FEP2C specifies necessary conditions for consciousness in *living organisms*. In Sect. 5, I discuss which (if any) of these conditions may also be necessary for consciousness in *artificial systems*. I argue that at least one of the conditions should be taken serious as a candidate for such a necessary condition.

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., 2023). 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, which 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). Systems that conform to the FEP are self-organising in the sense that they maintain this partition over time; furthermore, there must be characteristic features that define what kind of system they are. For human beings, for instance, these include a body temperature in a narrow range around 37° C. In other words, there are (relatively few) states (or paths) in which such self-organising systems are likely to be found, and (relatively many) states (or paths) in which they are unlikely to be found. The latter set of states are 'surprising', where "surprise" or "surprisal" in the sense used here, can formally be defined as the negative log-probability of an event. Given this, one can "specify equations of motion that a system must satisfy to remain the kind of thing that it is." (Ramstead et al., 2023, p. 3). Under the assumption that a system exists during a certain period of time, and has certain characteristic features, the

[3] This resonates with a point made by Piccinini (2021), according to which consciousness in a computer simulation may require "hardware that exhibits a sufficient degree of causal isomorphism to neural circuitry" (Piccinini, 2021, p. 139).

system must therefore satisfy such equations. What is more, these equations can be reformulated in terms of *surprisal minimisation* (Ramstead et al., 2023, p. 8).

For living organisms, minimising surprisal is necessary for survival (because staying alive entails continuing to exist and is–for living organisms–tantamount to remaining the kind of thing that they are[4]). This means that the physical processes that contribute to surprisal minimisation in living organisms include the physical processes that contribute to the organism's survival. (In particular, to the extent that the material realisers of consciousness contribute to an organism's survival, they will also be part of the physical processes that contribute to surprisal minimsation in that organism. We will return to this point in Sect. 4.)

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 Fig. 1): the dynamics of q_μ (and thereby of μ) can be described as minimising variational free energy $F\left(s, a, \mu\right)$ (Da Costa et al., 2021, pp. 9–10; Friston et al., 2023, pp. 8–9; Friston, Da Costa, Sakthivadivel, Friston et al., 2023a, b, pp. 42–43; Ramstead et al., 2023, p. 8).

Loosely speaking, systems that minimise variational free energy act in such a way that they encounter sensory states that have low surprisal, given the probabilities encoded by internal states, while at the same time making sure these probabilities accurately reflect the probabilities that define what kind of system they are. The lat-

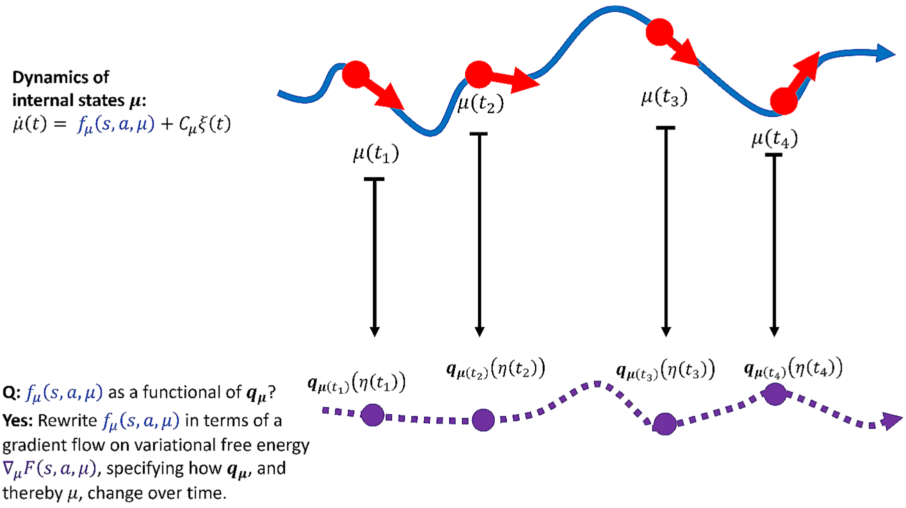
Dynamics of internal states μ:
$\dot{\mu}(t) = f_\mu(s,a,\mu) + C_\mu\xi(t)$
$\mu(t_1)$
$\mu(t_2)$
$\mu(t_3)$
$\mu(t_4)$
$q_{\mu(t_1)}(\eta(t_1))$
$q_{\mu(t_2)}(\eta(t_2))$
$q_{\mu(t_3)}(\eta(t_3))$
$q_{\mu(t_4)}(\eta(t_4))$
Q: $f_\mu(s,a,\mu)$ as a functional of q_μ?
Yes: Rewrite $f_\mu(s,a,\mu)$ in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

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

[4] What about periods in which organisms undergo drastic transformations, such as metamorphosis? Such cases may be handled by suitably specifying what kind of thing a metamorphic agent is (see, Clark, 2017, p. 12).

ter point emphasises that surprisal minimisation is not just about seeking states that are ‘easy to predict’ (e.g., being in a dark room), but about seeking states that are highly likely, given the kind of system that one is. We know that systems that exist tend to be found in such states, because otherwise, they would not exist (or would not be the kind of system that they are). Probabilities encoded by internal states can therefore not change arbitrarily, but must be such that states to which they assign high probabilities must also have low surprisal. This tends to be the case for systems that change internally encoded probability distributions in such a way that they minimise variational free energy, because free energy is an upper bound on surprisal.

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

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 a variational free energy gradient $\nabla_\mu F(s, a, \mu)$. 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).[5]

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., 2023, 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 system performs 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 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:

[5] Relatedly, one could wonder whether the claim that particular self-organising systems perform approximate Bayesian inference can only be defended under a perspectivalist account of physical computation (see Dewhurst, 2018; Schweizer, 2016; for discussion, see Coelho Mollo, 2021). For the purposes of this paper, no specific stance on physical computation is required. All that is needed is that internal states of systems conforming to the FEP *can* be described as performing computations, and that this description is non-arbitrary. Arguably, it is also possible to defend stronger readings of the claim that such systems are computational systems. For instance, it should be possible to argue that the statement is consistent with objective accounts of physical computation, such as Piccinini’s mechanistic account (Piccinini, 2015). This is because internal processes of living organisms that minimise variational free energy thereby contribute to goals of organisms (because minimising free energy is necessary for survival), and they do that by manipulating medium-independent variables (viz., beliefs encoded by internal states) in accordance with a rule over these variables (viz., such that they minimise variational free energy). If this is correct, internal states of living organisms have a function to compute (in an objective, non-perspectivalist sense).

> 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 et al., 2023, 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 encoded by internal states 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, which describe the system’s behaviour and internal processes in terms of minimising variational free energy. Minimising variational free energy entails approximate 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, it can also describe the dynamics of (some) conscious systems. It is an open question what further conditions conscious systems fulfill, in addition to minimising variational free energy (for some suggestions, see, e.g., Clark et al., 2019; Friston, 2018; Friston et al., 2020; Hohwy, 2022; Safron, 2020). Variational free energy is minimised with respect to a probability distribution, a generative model, so it is plausible to assume that the generative model 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 (Di Paolo et al., 2022; but see Friston et al., 2023).

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) At least some crucial differences 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, such that we can say: living organisms that instantiate these dynamics are conscious.

These assumptions 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 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 et al., 2023; Friston, Da Costa, Sakthivadivel, Friston et al., 2023a, b). 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. However, note that it doesn't presuppose that consciousness is a form of computation. Assuming that some crucial differences between conscious and non-conscious systems can be captured by Bayesian mechanics 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) and sufficient for consciousness in living organisms. It does not presuppose that implementing the right computations is sufficient for consciousness in all kinds of systems (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 proposed here is compatible with the possibility that the right computations must be implemented *in the right way*. This would mean 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 could this difference consist in?

4 The "FEP consciousness Criterion" (FEP2C)

Here, I use the following strategy to draw a distinction between simulating and being a conscious system. First, I highlight some characteristic features of conscious living organisms that conform to the FEP. Systems satisfying these features fulfill what I shall call the "FEP Consciousness Criterion" (FEP2C). Since FEP2C is based on consideration about conscious *living organisms*, we should not expect that artificial systems must satisfy FEP2C, in order to be conscious—just as we should not presuppose that having a neocortex or a biological nervous system is necessary for being conscious. A benefit of FEP2C is that it abstracts away from the underlying (biological) implementational details. Hence, FEP2C *can* be satisfied by non-biological artificial systems; but it is *not* satisfied by most current computers. This can be seen clearly by deconstructing FEP2C into a set of conditions entailed by this criterion. Furthermore, we can then determine to what extent it is plausible to regard these conditions as necessary for consciousness in artificial systems.

What does FEP2C consist in? Under the assumption that formulating a mechanical theory of consciousness is possible (Sect. 3) we can express the internal dynamics of conscious systems in two conjugate ways (Sect. 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 distribution *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 a given physical dynamics uniquely specifies the corresponding 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 could be described as implementing variational free energy minimisation. The underlying physical dynamics, however, would in general differ dramatically.

This brings us to the “FEP Consciousness Criterion” (FEP2C). For all conscious living organisms that conform to the FEP, the following holds:

(FEP2C) The system‘s physical dynamics entail computational dynamics that include computational correlates of consciousness.

The FEP2C makes two claims about conscious organisms: (i) The physical dynamics of conscious organisms entail computational dynamics. (ii) These computational dynamics include computational correlates of consciousness. The first claim (i) directly follows from the FEP (see Sect. 2), together with the definitions of “physical dynamics” and “computational dynamics” provided at the beginning of this section: according to the FEP, we can interpret an organism’s physical dynamics as a process of variational free energy minimisation (which entails approximate Bayesian inference). This establishes the first claim (i).

To show that the second claim (ii) holds, I will make an additional assumption, viz. that consciousness contributes in some way to the sustained existence of conscious living organisms. In other words, I assume that consciousness has a function for conscious organisms, by regularly contributing to the goals of organisms (where I take it that staying alive is among the goals of organisms, see Piccinini, 2020, p. 68). For instance, consciousness might enable a cluster of learning abilities (Birch et al., 2020; Birch, 2022; Kanai et al., 2019) that make some contribution to staying alive. Of course, perhaps the same cognitive capacities can be realised without consciousness (“conscious inessentialism,” Flanagan, 1993). But I submit that it is nevertheless plausible to assume that consciousness has a function for conscious living organisms (even if we cannot rule out epiphenomenalism about consciousness).

In Sect. 2, I noted that the physical dynamics of a conscious organism that contribute to surprisal minimisation include the material realisers of consciousness, under the assumption that consciousness contributes to the organism's survival. By reformulating these physical dynamics as a process of minimising variational free energy, we end up with a description of the organism's computational dynamics. If the physical dynamics include the supervenience base of consciousness in that organism (which they will, if consciousness contributes to the organism's survival), the computational dynamics will include computational correlates of consciousness. And that's exactly what the second claim (ii) entailed by FEP2C says.

Recall that by "computational correlates of consciousness" I mean computational processes that correlate with consciousness in living organisms and can be formulated in terms of minimising variational free energy. These processes are sufficient for consciousness in living organisms (but not necessarily in artificial systems). Computational correlates are thus a particular form of computational dynamics. Let us call them "conscious* computational dynamics" (with an asterisk to indicate that the system instantiating these dynamics need not be conscious).

FEP2C is not fulfilled by current computers, even if they were to simulate a conscious system by instantiating conscious* computational dynamics. We can see this by deconstructing FEP2C into a set of conditions entailed by FEP2C. If a system, e.g., the computer in your office, fails to fulfill any of these conditions, it also fails to fulfill FEP2C (even if it instantiates conscious* computational dynamics). (In the following section, I discuss whether failure to fulfill a condition entailed by FEP2C gives us reason to infer the absence of consciousness.)

If a system S satisfies FEP2C, then:

- [**Implementation condition**] S's conscious* computational dynamics are strongly constrained by S's hardware (or by the particular underlying mechanisms that implement these computations).
- [**Energy condition**] The "thermodynamic cost of computation" paid by S is relatively low (compared to current computers).
- [**Causal-flow condition**] The causal flow of S's conscious* computational dynamics matches the causal flow of S's physical dynamics.
- [**Existential condition**] S sustains its existence (partly) by virtue of its conscious* computational dynamics.

I shall explain these conditions, and how they follow from FEP2C, in the remainder of this section.[6]

[6] The conditions derived from the FEP2C are in part similar to three constraints on implementing consciousness in artificial systems recently proposed by Shiller (2024). These constraints are: *material complexity*, *causal integration*, and *continuity*. Let me just comment on the relationship between material complexity and the causal-flow condition, as well as between continuity and the existential condition. The causal-flow condition is similar to material complexity, but it is more demanding: low material complexity means that there is a relatively simple relation between states, properties, and material components (Shiller, 2024, p. 12); the causal-flow condition requires that causal interactions between the material components of a simulating system must sufficiently mirror the causal interactions between material components of the simulated system (in this sense, there is a greater similarity to Piccinini's, 2021, p. 139, causal-isomorphism requirement); furthermore, the causal-flow condition explicitly identi-

4.1 The implementation condition

According to the implementation condition, an organism's conscious* computational dynamics are, as it were, deeply tied to what it means to *be* that organism. What does this mean? In a nutshell: variational free energy is defined with respect to a generative model, the details of which depend on the particular organism. The organism's computational dynamics are therefore shaped by the particularities of the organism.

As an analogy, consider how the same chord can be played using different analog instruments, but also using software synthesisers. When using an analog instrument, the sound of the chord will be shaped by the physical properties of the instrument. When the chord is played using a software synthesiser, different analog instruments can be emulated (to some extent), and so the sound of the chord is significantly less constrained by the physical properties of the machine on which the software is running.

Recall from Sect. 2 that the organism's physical dynamics capture the characteristic features of the organism (e.g., human beings tend to be found in regions of their state space in which their body temperature is close to 37° C). Put differently, to the extent that the organism displays these dynamics it "remain[s] the kind of thing that it is" (Ramstead et al., 2023, p. 3). According to the FEP, the organism's physical dynamics can be redescribed as approximate Bayesian inference (by variational free-energy minimisation). That is, the FEP entails that the physical dynamics of living organisms can be interpreted as a form of computational dynamics. In doing this, we are just taking another perspective on the same dynamics. And since the physical dynamics capture what it means to *be* the organism, the same goes for the computational dynamics: the computations performed by the organism (i.e., those entailed by its physical dynamics) are deeply tied to what it means to *be* that organism.

In other words, the system's hardware (the material basis of the computational correlates found in that system) puts strong constraints on its conscious* computational dynamics. This formulation is relatively vague. What does "strong" mean in this context? It may not be possible to quantify the strength of the constraints, but there is a clear qualitative difference between the way in which computational correlates are implemented by living organisms, and how they are implemented in current computers with a classical architecture. In current computers, there is a separation of software and hardware: the same software can be run on different tokens of the same type of hardware. This is extremely useful, because apps can be copied and installed on different computers, without having to modify the apps for each particular computer (as long as they are of the same type or use the same operating system). Once a large language model has been trained, its weights can be copied, and multiple instances

fies as relevant the causal interactions between internal, blanket, and external states (as identified through the lens of the FEP). A system displays continuity to the extent to which "its parts persist over time under identities that are separate from their functional roles" (Shiller, 2024, p. 16). According to the existential condition, by contrast, the system as a whole continues to exist (in part) by virtue of its computational dynamics. Hence, instead of focusing parts of the system, the existential condition focuses the entire system; and instead of highlighting properties of parts that are *independent* of their functional roles, the existential condition highlights properties of the whole it has *by virtue of* (computational) functional roles of its parts.

of the same model can be run. The involved computational processes are “immortal”, because the same computational processes can be instantiated over and over again, in different pieces of hardware of the same type.

Hinton (2022) contrasts this form of computation with what he calls “mortal computation”. In mortal computation, the algorithms run by a given system are strongly constrained by the system’s particular hardware. This is the case for biological brains: even if you could record and copy the “connection weights” of my brain, trying to implement the same connection weights in your brain would be hopeless. Aside from practical complications, the individual differences between our brains (especially differences in connectivity) would make it impossible to instantiate the same computations, merely by “copying” connection weights.

Hinton (2022) suggests that allowing for differences between different tokens of the same type of hardware may reduce the cost of hardware production and save energy. In turn, every instance of a model would have to learn the model parameters that work for the particular piece of hardware on which it is running: “These parameter values are only useful for that specific hardware instance, so the computation they perform is mortal: it dies with the hardware.” (Hinton, 2022, p. 13).

Similarly, conscious organisms that satisfy FEP2C implement computational correlates of consciousness (conscious* computational dynamics) in a particular way. Recall that I am assuming that the computational correlates can be described in terms of minimising variational free energy. Variational free energy is defined with respect to a generative model, the details of which depend on the particular organism. There may be some general abstract properties shared by all conscious organisms, but the way in which the computational processes in a particular organism instantiate these properties differs from the way in which computational processes in other organisms instantiate them.

A stronger version of the implementation condition (which I am not presupposing here) would entail that the *only* way to implement computational correlates of consciousness is by using biological hard/wetware.[7] For instance, it can be questioned whether a digital computer could ever simulate the functional organisation of a human brain in real time; this puts pressure on the claim that the functional roles played by the neural realisers of consciousness are multiply realisable (Cao, 2022).

Here, I am not presupposing that consciousness depends on a functional organisation that is so tightly integrated with the properties of biological neurons, as to make it substrate-dependent. That is, I am remaining open to the possibility that consciousness is multiply realisable. The implementation condition only suggests that, within a given system, there is a tight integration between the physical properties of the material realisers of consciousness and the (computational) functional roles they realise.

That is, the implementation condition suggests that, although consciousness may be multiply realisable, there is a special way in which it is realised in conscious living organisms (just as a chord is realised in a special way by an analog musical instrument). Whether this is required for consciousness, or just a contingent fact about conscious living organisms, is a further question (to which I return below).

[7] I am grateful to an anonymous reviewer for suggesting this stronger interpretation.

4.2 The energy condition

The energy condition is a corollary of the implementation condition. Above, I said that the computations performed by the organism (i.e., those entailed by its physical dynamics) are deeply tied to what it means to *be* that organism. This is because, according to the FEP, particular self-organising systems follow the "path of least surprisal" (Miller et al., 2022, p. 4) as long as they continue to exist, and pursuing the path of least surprisal can alternatively be described as variational free energy minimisation (and thereby as a process of approximate Bayesian inference). In other words, there are computations an organism "automatically" performs, simply by virtue of its continued existence. But this means that the energy an organism needs to sustain its existence includes the energy it needs to minimise variational free energy. Put differently, these computations "come for free".[8] Hardware that uses mortal computation may have a similar benefit, which is why Hinton (2022) suggests that mortal computation might be the future of computing: "If you want your trillion parameter neural net to only consume a few watts, mortal computation may be the only option." (Hinton, 2022, p. 13).

4.3 The causal-flow condition

Conscious living organisms that conform to the FEP instantiate conscious* computational dynamics not just in an efficient way. There is also, by assumption, a separation between internal and external states, and a circular causal flow between internal and external states, mediated by blanket states (i.e., perceptual and active states). 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.

In general, such a match between the realisers of physical and computational dynamics cannot be taken for granted.[9] For the sake of illustration, assume that a computer with a von Neumann architecture can be described as a self-organising system that conforms to the FEP. Furthermore, assume that the computer *simulates* a system that satisfies FEP2C. This means the computer instantiates computational correlates of consciousness (which can be described in terms of minimising variational free energy). In particular, the computer must encode a probability density over some external states, given blanket states. Denote 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 that is simulated by the computer.

The computer's physical states that represent μ_c are part of the computer's memory. The computer simulates the computational dynamics by implementing a gradi-

[8] Formally, the *thermodynamic cost of computation* can be described in terms of the heat generated by individual computational operations. A lower bound on this is given by *Landauer's principle* (Landauer, 1961), which specifies the minimal amount of heat required to erase one bit of information. Current computers are vastly less energy efficient than Landauer's bound suggests is possible.

[9] In a slightly different vein, one could argue that the (discrete) causal structure of digital computations *cannot* mirror the causal structure of the underlying vehicles – because even a digital computer has continuous underlying microphysical dynamics (see Anderson & Piccinini, 2024, ch. 7).

ent descent on variational free energy. Hence, we can assume that the computations performed by the computer include those that are performed by the simulated conscious organism. But the way in which these computations are implemented differ in the following respect. Note that 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. Since the states that encode μ_c are part of the computer's memory, and since the computations that update the values of μ_c are performed in the CPU, these states never directly causally interact with states that represent the organism's external, sensory, and active states.

To make it even more explicit, denote the simulated external states with η_c, and sensory and active states with s_c and a_c, respectively. 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. That is, 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 differs from the basic causal flow between system states in the simulated conscious organism: in the simulated system, (some) external and internal states directly causally interact with (some) blanket states.[10]

The difference in the basic causal flows is illustrated in Fig. 2.

4.4 The existential condition

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 (in part) *by virtue of* performing those computations. Notably, this does *not* mean that minimising variational free energy is sufficient for one's continued existence. On the contrary, free energy minimisation is only *necessary* for the sustained existence of particular self-organising systems (including living organisms, Constant, 2021). But this means that, if a living organism exists for a certain period of time, we can (partly) explain this fact in terms of the computations it performed by virtue of existing (i.e., in terms of the computational dynamics entailed by its physical dynamics).

Contrast this with a simulation in a von Neumann computer, which may perform the same computations, by representing the organism'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 (or the whole computer) do not exist by virtue of their role in these computations.

[10] Similarly, there can be more direct causal interaction in computers that use memcomputing or compute-in-memory. I thank Johannes Kleiner and Daniel Friedman for pointing me to this.

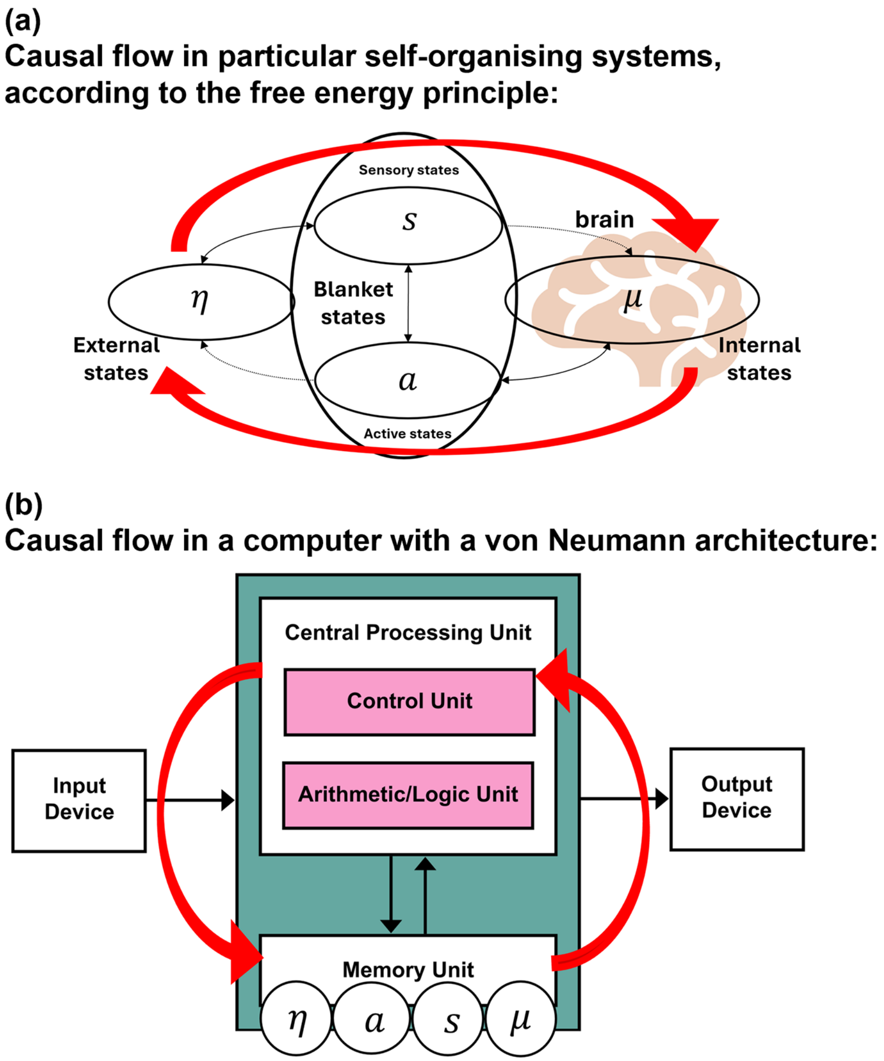

Fig. 2 (**a**) Causal flow (depicted by the big (red) arrows) between system states in a self-organising system that conforms to the free energy principle: there is a direct causal relation between (some) blanket states and (some) external states, as well as between internal states and blanket states. The causal coupling between internal and external states is mediated by blanket states. (**b**) Causal flow between encodings of system states in a computer simulation in a computer with a von 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 unit. 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 illustration used here.)

Another way of expressing this is that living organisms *give a damn* (Haugeland, 2000), because they exist (in part) by virtue of performing certain computations. Since minimising variational free energy is necessary for survival, failing to do so (over a certain period of time) will lead to death. Hence, it matters what kinds of computations living organisms perform, their continued existence depends on it (again, this does not mean that performing the right computations is sufficient for survival, it is only necessary).

5 Which conditions entailed by FEP2C, if any, are necessary for artificial consciousness?

In the previous section, I described four conditions entailed by FEP2C. These conditions are interesting, because they are not fulfilled by most current computers, even if the computers were to instantiate computational correlates of consciousness. In principle, one could therefore use these conditions to draw a distinction between simulating and replicating consciousness. However, the mere fact that conscious living organisms have properties that computers lack does not mean that these properties are essential for consciousness. That is, consciousness could be present in artificial systems that do not fulfill any of the conditions entailed by FEP2C.[11]

Here, I argue that the implementation condition and the energy condition should not be regarded as necessary for consciousness in artificial systems.[12] The causal flow and the existential condition are more plausible candidates, but the existential condition may in the end be too strong. That is, one has to make strong presuppositions about consciousness to justify this condition. The causal-flow condition, by contrast, may be strong enough to be interesting, but weak enough to require only a modest additional assumption about consciousness.

5.1 Implementation and energy

Human beings are not the only conscious animals. Apart from other mammals, there are good reasons to also take the possibility of consciousness in invertebrates seriously (Birch, 2022; Klein & Barron, 2016; Wickens, 2022). In fact, consciousness may have independently evolved multiple times during evolutionary history (Ginsburg & Jablonka, 2019). If this is true, there are different ways in which consciousness can be instantiated in animals. Hence, whatever constraints the animal's body puts on the *capacity* for consciousness must be of a very general kind. (This is not to

[11] In particular, it may be that what I here call "conscious* computational dynamics" is not just sufficient for consciousness in living organisms, but also in artificial systems. The FEP would then not serve to characterise the distinction between simulating and replicating consciousness, but to characterise the computational dynamics ("conscious self-evidencing," as Hohwy, 2022, calls it) that is sufficient for consciousness (in any kind of system).

[12] That does not mean that these conditions are irrelevant. Perhaps they can still be considered as (weak) indicators of consciousness in artificial systems, such that their presence in an artificial system makes it a bit more likely that the system is conscious. I am grateful to an anonymous reviewer for pressing me on this.

say that the body does not shape what it is like to be a particular conscious animal; i.e., being conscious in a particular way may heavily depend on the way consciousness is implemented; but the capacity for creature consciousness, as opposed to the capacity for being in a particular conscious state, is likely to be less dependent on particular implementational details.)

Furthermore, one could argue that computers are also very much constrained by their hardware. Although a computer can be regarded as a universal Turing machine (only limited by its available energy and memory), the computer's architecture determines which computations it can efficiently perform within a reasonable amount of time.

On the one hand, one could therefore argue that being conscious requires at most fulfilling a rather weak version of the implementation condition. On the other hand, one could argue that even computers with a classical architecture fulfill a weak version of this condition. The implementation condition therefore seems unfit to draw a distinction between simulating and being conscious.

Similarly, it is hard to see why being energy-efficient should be necessary for being conscious. The very fact that being energy-efficient brings an evolutionary advantage helps to explain why this may be necessary for naturally evolved species. But this does not mean that computers cannot bypass this requirement. From a *practical* point of view, future computers (including potentially conscious ones), may need to satisfy the energy condition. But this not a nomological necessity (for a different view, see Thagard, 2022).

5.2 Existence

The existential condition is perhaps the strongest of the four conditions. It may also be the most interesting, because it may capture the intuition (which some people have) that there is a strong connection between life and consciousness (Thompson, 2022); only systems that satisfy the existential condition have "skin in the game" (Aru et al., 2023; Taleb, 2018). Hence, if consciousness requires fulfilling the existential condition, then consciousness *matters* (Cleeremans & Tallon-Baudry, 2022; Froese, 2017). There is also a sense in which systems that satisfy the existential condition *give a damn* (Haugeland, 2000), because their continued existence is contingent on their conscious* computational dynamics.

Although fulfilling the existential condition does not entail being alive, it is still a strong condition. For instance, it is not satisfied by virtual agents in a virtual environment. Even if these agents perform computations that help them sustain their existence in the virtual environment, the continued existence of the computer (i.e., the underlying physical system simulating the agents and their environment) is not contingent on the computations it performs. If the existential condition is necessary for consciousness, we could therefore also rule out that *we* are conscious beings in a virtual world, simulated by a computer at the "next higher" level of reality (a possibility we have to take seriously, according to Bostrom, 2003; Chalmers, 2022). Making the case that artificial conscious systems must fulfill the existential condition would therefore require further justification (which I do not intend to provide here).

5.3 Causal flow

This leaves us with the causal-flow condition. Taken by itself, a difference in causal flow may seem arbitrary. However, this difference has further consequences that are not immediately obvious. A conscious system S that satisfies FEP2C, and thereby fulfills the causal-flow condition, has a particular architecture, in which internal states interact with external states (mediated by blanket states); furthermore, internal states encode a probability distribution over the very same external states (given blanket states). A computer (with a classical architecture) that performs the same computations as S must, at some level of analysis, also encode a probability distribution over S's external states. But the physical vehicles of this encoding do not causally interact with S's external states. Even if the computer simulates S's internal, external, and blanket states, the states that represent internal states will not directly interact with the states that represent blanket states (or external states). Furthermore, if μ_c is the vehicle of a representation of S's internal states, its *own* blanket and external states will not be the states that figure in the probability distribution encoded by S's internal states.

Another way of describing this difference is that the internal states of a conscious system that satisfies FEP2C can, in principle, be "detached" from its blanket and external states (just as a brain in a vat is detached from its biological body). Furthermore, assume that it is nomologically possible to replace the biological neurons of a conscious brain with synthetic neurons or silicon chips. It is then also nomologically possible to "detach" the silicon-chip brain and connect it to a physical (biological or robotic) body. If, by contrast, the computations performed by the silicon-chip brain are implemented in a computer with a von Neumann architecture, and if that computer also performs the computations required to simulate sensory signals, then it will not be possible to "detach" the part of the computer that simulates the silicon-chip brain. In other words, although it may be possible to "detach" the simulated brain on the software level (i.e., the same computations could be performed by another computer), it is not possible to "detach" the simulated brain on the implementational (hardware) level. (See Fig. 3 for an illustration.)

A further way of formulating this idea is in terms of what Shoemaker (1976) calls "paradigmatic embodiment", where a paradigmatically embodied individual is an ordinary human being (or an ordinary specimen of a type of animal). Brains in vats and silicon-chip brains are similar in that they both could (in principle) become incorporated into the body of a paradigmatically embodied individual, without changing their internal structure. In other words, they are, as Block (1978, p. 299) puts it, "limiting case[s] of an amputee—amputation of everything but the brain." In a computer simulation of a conscious agent in an environment, the part of the computer that simulates the conscious agent is not a limiting case of an amputee.

Let us unpack this a little. Imagine a digital agent in a virtual environment that behaves like a conscious being. Furthermore, imagine that this virtual entity can upload itself to 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). Let us stipulate that a robot of the right kind (perhaps a soft robot, as suggested by Bronfman et al., 2021; Man

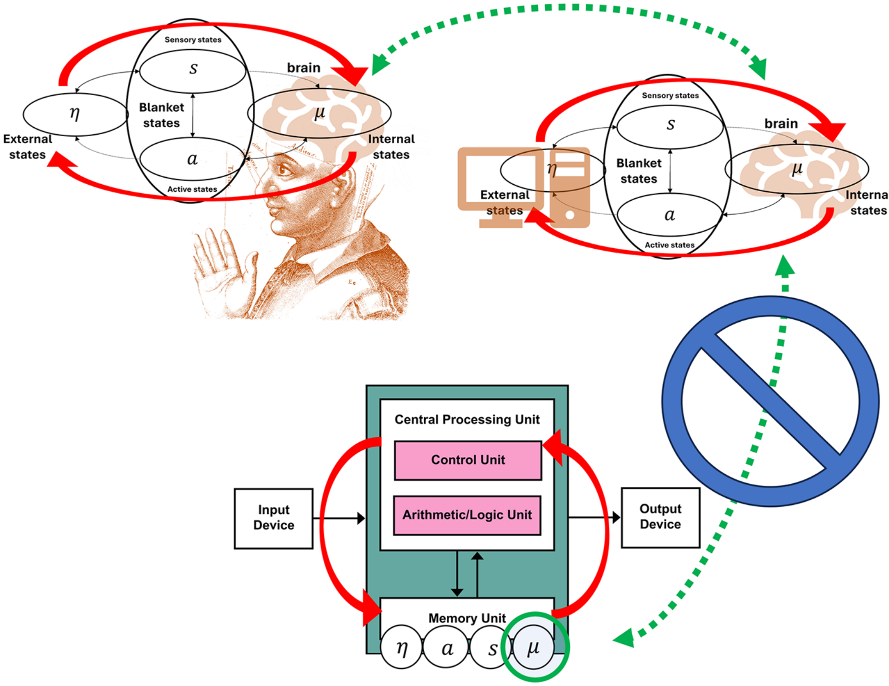

Fig. 3 This figure illustrates a paradigmatically embodied individual (a brain within a body; top left), as well as a brain in a vat (top right), which can be regarded as a limiting case of a paradigmatically embodied individual. A part of a computer simulation that simulates a conscious being (bottom), however, is not a limiting case of such an individual, because it could not become incorporated into a paradigmatically embodied individual without changing its internal structure. (The illustration of the human being has been adapted from https://commons.wikimedia.org/wiki/File:Robert_Fludd,Tomus_secundus…,_1619-1621_Wellcome_L0028467.jpg, which has been published under a CC BY 4.0 license. 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 illustration used here.)

& Damasio, 2019) would count as a paradigmatically embodied individual. Hence, it seems that the virtual agent can become a paradigmatically embodied individual. If correct, this shows that the virtual agent can be regarded in analogy to a brain in a vat and a silicon-chip brain. That is, it should be regarded as a conscious being, even when it is part of a computer simulation (at least if, as argued by Chalmers, 2022, virtual worlds can be as real as non-virtual worlds).

A difference between the virtual agent and the physically embodied agent is the following. You have to change the internal structure of the virtual agent's material realiser, in order to incorporate it into a physical robot. There may not be a relevant difference in terms of the functional roles performed by them, but you cannot simply "detach" the physical realiser of the virtual agent and incorporate it into the robot. In other words, the internal states of the robot are not the same as the internal states that are part of the physical system (the computer) that performs the computations required to simulate a virtual agent in an environment. The difference is not just a lack of numerical identity (after all, the silicon-chip brain is not numerically identical to the biological brain). It is a difference in the internal causal structure. See Fig. 3 for an illustration.

Why should this difference be relevant to distinguishing between a simulation and a replication of consciousness? Here is one possible reason. Consciousness is a special property. Consider the property of playing chess. If two virtual agents play chess in a sophisticated computer simulation, then there is actually a game of chess going on in that computer simulation. There can also be an interface to our level of reality, such that we can play against virtual agents in the computer simulation.

Contrast this with the property of being wet. A computer simulation of a rainstorm will not make us wet (Searle, 1980, p. 423). One could respond: yes, it will not make *us* wet, but if it occurs within a sufficiently detailed computer simulation, the virtual individuals in the simulation will become wet. Chalmers (2022, p. 367) credits Hofstadter (1981) with this idea: “Hofstadter’s insight is that whether or not we recognize a simulated hurricane as a hurricane depends on our perspective. In particular, it depends on whether we’re experiencing the simulated hurricane from inside or outside the simulation.”

The reply works less well if we replace “being wet” with “being conscious”. Consider: *Whether a simulated conscious agent is conscious or not depends on our perspective. In particular, it depends on whether we interact with it from inside or outside the simulation.* That would make consciousness strangely observer-dependent. If a virtual agent in a computer simulation is conscious, its being conscious does not depend on the perspective one takes on it (under the assumption that consciousness is an observer-independent property, as most would assume). Instead, I suggest we should ask: can it interact with our level of reality in the same way as it can interact with its virtual environment? In other words, can we perform an “amputation” of the virtual agent’s material realiser, and incorporate it into a body in our level of reality? If the simulation is implemented in a computer with a classical architecture, the answer will be “no”.

One could object that this criterion is too strong. Being able to perform an amputation on the *software level* should be sufficient. That is, we should ask: can we upload the virtual agent to a robot in our level of reality? Of course, I do not have a knockdown argument against this reply. I can only say that the conscious beings we currently know are different, and that this difference might matter. We can substitute our sensory signals with virtual signals and enter virtual worlds, without having to alter our entire hardware (at least we can approximate such a substitution; a complete Matrix-style substitution is of course currently still science-fiction). That is, paradigmatically embodied conscious beings can enter different levels of (virtual) reality using numerically identical material realisers: Neo’s conscious experience is grounded in activity of the same central nervous system, regardless of whether he is currently in the Matrix or in the non-virtual level of reality. By contrast, a virtual conscious agent that uploads its mind to a robot in a non-virtual environment leaves its original hardware entirely behind. Admittedly, the extent to which this difference matters will probably remain a matter of debate.[13]

Where does this leave us? I have argued that the implementation and the energy condition are too strong to be regarded as necessary for consciousness in artificial systems. Similarly, the existential condition may also be too strong (although it may capture some intuitions about the connection between consciousness and being alive). The causal-flow condition is comparably weaker. It may be satisfied by computers with a non-classical architecture (i.e., by computers that do not separate between memory and central processing units). But it is not satisfied by computers with a von Neumann architecture. Furthermore, considerations about the perspective-independence of consciousness provide an independent reason to believe that conscious systems must fulfill the causal-flow condition. I submit that it is at least a plausible candidate for a necessary condition for consciousness in artificial systems.

6 Conclusion

One possible approach to artificial consciousness asks: how likely is it that current AI systems are conscious, and what must be added to existing systems to increase the probability that they are conscious (Butlin et al., 2023; Chalmers, 2023; Graziano, 2017; Juliani et al., 2022)? Another asks: what types of AI systems are *unlikely* to be conscious (Piccinini, 2021; Tononi & Koch, 2015), and how can we *rule out* that certain types of systems are conscious?

The second approach has the advantage that it may mitigate the risk of inadvertently creating artificial consciousness; this would be desirable, because it is currently not clear under what conditions creating artificial consciousness would be morally permissable (Agarwal & Edelman, 2020; Metzinger, 2021).

If there are necessary conditions for consciousness in artificial systems that are not fulfilled by large classes of artificial systems (e.g., conditions that are not fulfilled by computers with a von Neumann architecture), then those types of artificial systems cannot be conscious. How can we find out whether there are any necessary conditions of this kind? One strategy is to focus on different types of systems with a known capacity for consciousness (including non-human animals). One can then look for properties that different types of conscious animals have in common (Andrews & Birch, 2023). A general property shared by all conscious animals is that they are alive. Being alive may be too strong to qualify as a plausible candidate for a necessary condition for consciousness (Chalmers, 2023). But perhaps some conditions that are necessary for being alive are also necessary for consciousness?

From the perspective of the free energy principle (FEP), being alive entails minimising variational free energy. As I have shown, the FEP can be used to determine further properties that conscious living organisms have in common; I have subsumed

[13] Without getting into the details here, my impression is that accepting the possibility of consciousness in computer simulations in a classical hardware requires biting a large number of bullets, including strange implementations. Conversely, requiring that the material realisers of consciousness can, in principle, become incorporated into paradigmatically embodied individuals allows one to avoid many extremely counter-intuitive consequences.

these properties under the label of the "FEP Consciousness Criterion" (FEP2C). FEP2C therefore specifies necessary conditions for consciousness in *living organisms*, and one can ask which (if any) of these conditions are also necessary for consciousness in *artificial systems*. I have argued that at least one of the conditions should be taken serious as a candidate for such a necessary condition: the causal-flow condition. The causal-flow condition is interesting, because it is not satisfied by computers with a von Neumann architecture. It is also plausible (i.e., worthy of further scrutiny), because it resonates with the idea that conscious systems must either be "paradigmatically embodied" (Shoemaker, 1976), or must at least be limiting cases of paradigmatically embodied individuals. If these considerations are on the right track, the FEP may provide the resources to draw a substantial and plausible distinction between *being* and *simulating* a system of a certain type.

Acknowledgements I am grateful to Mackenzie Dion, Johannes Kleiner, Maxwell Ramstead, and Tomek Korbak for feedback on a draft of this paper. An earlier version of this paper was presented online at the Active Inference Institute. I thank Daniel Friedman for inviting me and for discussion. I am also extremely grateful to Maxwell Ramstead for inviting me to present the paper at his Computational Phenomenology discussion group; I thank everyone who attended this session, in particular, Karl Friston, Alex Kiefer, Ken Williford, and Jeff Yoshimi. I also thank the audience at the Models of Consciousness 4 conference at Oxford, where I presented core contents of this paper. Last but not least, I am grateful to two anonymous reviewers of this journal for their constructive feedback.

Funding Open Access funding enabled and organized by Projekt DEAL.

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.

Anderson, N., & Piccinini, G. (2024). *The physical signature of computation: A robust mapping account*. Oxford University Press.

Andrews, K., & Birch, J. (2023). To understand AI sentience, first understand it in animals | aeon essays. In *Aeon*. https://aeon.co/essays/to-understand-ai-sentience-first-understand-it-in-animals.

Aru, J., Larkum, M. E., & Shine, J. M. (2023). The feasibility of artificial consciousness through the lens of neuroscience. *Trends in Neurosciences, 46*(12), 1008–1017. https://doi.org/10.1016/j.tins.2023.09.009

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.

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

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

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.

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

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

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

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*. W. W. Norton & Company. First edition.

Chalmers, D. J. (2023). Could a large language model be conscious? *Boston Review*. https://www.bostonreview.net/articles/could-a-large-language-model-be-conscious/.

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

Clark, A. (2017). How to knit your own Markov blanket. In T. K. Metzinger & W. Wiese (Eds.), *Philosophy and predictive processing*. MIND Group. https://doi.org/10.15502/9783958573031.

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.

Coelho Mollo, D. (2021). Against computational perspectivalism. *The British Journal for the Philosophy of Science*, *72*(4), 1129–1153. https://doi.org/10.1093/bjps/axz036.

Constant, A. (2021). The free energy principle: It's not about what it takes, it's about what took you there. *Biology & Philosophy*, *36*(2), 10. https://doi.org/10.1007/s10539-021-09787-1.

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, & Di E. A. Paolo (Eds.), *Enaction: Towards a new paradigm for cognitive science* (pp. 361–385). MIT Press.

Da Costa, L., Friston, K., Heins, C., & Pavliotis, G. A. (2021). Bayesian mechanics for stationary processes. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *477*(2256), 20210518. https://doi.org/10.1098/rspa.2021.0518.

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

Dewhurst, J. (2018). Computing mechanisms without proper functions. *Minds and Machines*, *28*(3), 569–588. https://doi.org/10.1007/s11023-018-9474-5.

Di Paolo, E., 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.

Farrell, B. A. (1950). Experience. *Mind*, *59*(234), 170–198.

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. (2023a). The free energy principle made simpler but not too simple. *Physics Reports*, *1024*, 1–29. https://doi.org/10.1016/j.physrep.2023.07.001.

Friston, K., Da Costa, L., Sakthivadivel, D. A. R., Heins, C., Pavliotis, G. A., Ramstead, M., & Parr, T. (2023b). Path integrals, particular kinds, and strange things. *Physics of Life Reviews*. https://doi.org/10.1016/j.plrev.2023.08.016.

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.

Ginsburg, S., & Jablonka, E. (2019). *The evolution of the sensitive soul: Learning and the origins of consciousness*. MIT Press.

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.

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.

Hinton, G. (2022). *The forward-forward algorithm: Some preliminary investigations*. *arXiv:2212.13345*. http://arxiv.org/abs/2212.13345.

Hofstadter, D. R. (1981). A coffee-house conversation on the Turing test. *Scientific American*.

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.

Hohwy, J. (2022). Conscious self-evidencing. *Review of Philosophy and Psychology*, *13*(4), 809–828. https://doi.org/10.1007/s13164-021-00578-x.

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., & Barron, A. (2016). Insects have the capacity for subjective experience. *Animal Sentience*, *1*(9). https://doi.org/10.51291/2377-7478.1113.

Kleiner, J., & Ludwig, T. (2023). *If consciousness is dynamically relevant, artificial intelligence isn't conscious*. *arXiv:2304.05077*. http://arxiv.org/abs/2304.05077.

Landauer, R. (1961). Irreversibility and heat generation in the computing process. *IBM Journal of Research and Development*, *5*(3), 183–191. https://doi.org/10.1147/rd.53.0183.

LeDoux, J., Birch, J., Andrews, K., Clayton, N. S., Daw, N. D., Frith, C., Lau, H., Peters, M. A. K., Schneider, S., Seth, A., Suddendorf, T., & Vandekerckhove, M. M. P. (2023). Consciousness beyond the human case. *Current Biology*, *33*(16), R832–R840. https://doi.org/10.1016/j.cub.2023.06.067.

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.

Metzinger, T. (2021). Artificial suffering: An argument for a global moratorium on synthetic phenomenology. *Journal of Artificial Intelligence and Consciousness*, *08*(01), 43–66. https://doi.org/10.1142/S270507852150003X.

Miller, M., Albarracin, M., Pitliya, R. J., Kiefer, A., Mago, J., Gorman, C., Friston, K. J., & Ramstead, M. J. D. (2022). Resilience and active inference. *Frontiers in Psychology*, *13*, 1059117. https://doi.org/10.3389/fpsyg.2022.1059117.

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, *83*(4), 435–450. https://doi.org/10.2307/2183914.

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.

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. (2015). *Physical computation: A mechanistic account*. Oxford University Press.

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.

Prinz, J. (2012). *The conscious brain. How attention engenders experience*. Oxford University Press.

Ramstead, M. J. D., Sakthivadivel, D. A. R., Heins, C., Koudahl, M., Millidge, B., Da Costa, L., Klein, B., & Friston, K. J. (2023). On bayesian mechanics: A physics of and by beliefs. *Interface Focus*, *13*(3), 20220029. https://doi.org/10.1098/rsfs.2022.0029.

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.

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://doi.org/10.3389/frai.2020.00030. https://www.frontiersin.org/article/.

Schweizer, P. (2016). In what sense does the brain compute? In V. C. Müller (Ed.), *Computing and philosophy* (pp. 63–79). Springer.

Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–457.

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

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

Shoemaker, S. (1976). Embodiment and behavior. In A. Rorty (Ed.), *The identities of persons*. Berkely University. https://philpapers.org/rec/FINEAB.

Taleb, N. N. (2018). *Skin in the game: Hidden asymmetries in daily life*. Random House.

Thagard, P. (2022). Energy requirements undermine substrate independence and mind-body functionalism. *Philosophy of Science*, *89*(1), 70–88. https://doi.org/10.1017/psa.2021.15.

Thompson, E. (2022). Could all life be sentient? *Journal of Consciousness Studies*, *29*(3–4), 229–265. https://doi.org/10.53765/20512201.29.3.229.

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.

Wickens, S. (2022). Review of the evidence of sentience in cephalopod molluscs and decapod crustaceans. *Animal Welfare*, *31*(1), 155–156. https://doi.org/10.1017/S0962728600009866.

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.