PhiMiSci
Philosophy and the Mind Sciences

The neural correlates of consciousness under the free energy principle

From computational correlates to computational explanation

Wanja Wiese[a] (wanja.wiese@rub.de)
Karl J. Friston[b] (k.friston@ucl.ac.uk)

Abstract

How can the free energy principle contribute to research on neural correlates of consciousness, and to the scientific study of consciousness more generally? Under the free energy principle, neural correlates should be defined in terms of neural *dynamics*, not neural states, and should be complemented by research on *computational* correlates of consciousness – defined in terms of probabilities encoded by neural states.

We argue that these restrictions brighten the prospects of a computational explanation of consciousness, by addressing two central problems. The first is to account for consciousness in the absence of sensory stimulation and behaviour. The second is to allow for the possibility of systems that implement computations associated with consciousness, without being conscious, which requires differentiating between computational systems that merely simulate conscious beings and computational systems that are conscious in and of themselves.

Given the notion of computation entailed by the free energy principle, we derive constraints on the ascription of consciousness in controversial cases (e.g., in the absence of sensory stimulation and behaviour). We show that this also has implications for what it means to *be*, as opposed to merely *simulate* a conscious system.

Keywords

Active inference · Computational correlates of consciousness · Computational explanation · Consciousness · Free energy principle · Islands of awareness · Minimal unifying model · Neural correlates of consciousness

This article is part of a special issue on "The Neural Correlates of Consciousness," edited by Sascha Benjamin Fink and Ying-Tung Lin.

[a]Institute of Philosophy II, Ruhr-Universität Bochum
[b]Wellcome Centre for Human Neuroimaging, UCL

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

1 Introduction

The free energy principle (FEP, Friston, 2010) provides an information-theoretic analysis of the concept of existence of self-organising systems (Hohwy, 2020). It entails that self-organising systems at non-equilibrium steady state[1] bound the entropy of their sensory signals by minimising – on average – a quantity known as variational free energy. As such, FEP is not a theory of consciousness. Consequently, it is not obvious that FEP has any relevance to research on neural correlates of consciousness (NCCs, see Chalmers, 2000). Furthermore, FEP is a computational principle;[2] hence, it may seem even less relevant to investigating the *neural* structures associated with consciousness.

Here, we will argue that FEP provides constraints on *computational* correlates of consciousness (CCCs, see Atkinson et al., 2000; Cleeremans, 2005; Mathis & Mozer, 1996; Reggia et al., 2016, 2019). CCCs specify computational properties instantiated by NCCs, and thereby provide some insights as to *why* a particular neural structure or type of activity is associated with consciousness.

More specifically, we will argue that FEP supports the following three observations with respect to the debate on NCCs and CCCs. (i) According to FEP, NCCs must be defined in terms of neural *dynamics*, not neural states. (ii) According to FEP, there is a relevant distinction to be made between the probabilities *of* neural states (or of trajectories) and the probabilities *encoded by* neural states.[3] These distinct families of probability distributions can be regarded as points on statistical manifolds, with corresponding information geometries. In Friston, Wiese, et al. (2020), the information geometries – associated with the probabilities of states and with probabilities encoded by states – are called *intrinsic* and *extrinsic information geometries*, respectively. In line with this distinction, neural *dynamics* (with NCCs as a special case) pertain to movements on the intrinsic statistical manifold, whereas neural *computations* (with CCCs as a special case) pertain to movements on the extrinsic manifold. Furthermore, the computations ascribed to movements on the extrinsic manifold can be regarded as inferences or, more poetically, self-evidencing (Hohwy, 2016).[4] (iii) Some candidates for CCCs that have been pro-

[1] A system is at non-equilibrium with its environment if it is in exchange with its environment. It is in steady state if it has characteristic features that remain invariant during this exchange, i.e., if it continues to exist.

[2] Technically, the FEP is a variational principle of stationary action. The "action" in question here is a path integral of a functional of probabilistic beliefs encoded by the internal states of a system. As such, FEP becomes a computational principle; in that the probabilistic beliefs in question are *about* something (i.e., the external states of a system). For a discussion of the relationship between FEP, computation, and representation, see Wiese & Friston (2021).

[3] A general, lucid analysis of the two kinds of information processing associated with probabilities of states (representational vehicles) and with probabilities represented by states (probabilities in represented content), respectively, is provided by Sprevak (2020). For a related, excellent discussion with respect to FEP, see Kiefer (2020).

[4] This is because every point on the extrinsic manifold is equipped with a free energy functional, namely a function of the probabilistic belief corresponding to that point. Crucially, this functional

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

posed in the literature are – from the point of view of FEP – just mathematical descriptions of NCCs. (iv) If a goal of research on correlates of consciousness is to infer the presence or absence of consciousness in controversial cases, then one should regard these correlates as necessary (not sufficient) conditions for consciousness. In particular, this holds for computational principles entailed by FEP.

A key contribution to the debate that follows from these observations is that computational explanations of consciousness (or "explanatory correlates of consciousness," Seth, 2009) must be specified with respect to the extrinsic information geometry. That is, a computational explanation of consciousness requires more than a formal description of neural dynamics. It must also specify computations instantiated by these dynamics. Moreover, FEP has implications for the very idea of a computational explanation of mental phenomena. FEP not only specifies computations that have to be performed by self-organising systems at non-equilibrium steady state, but also entails that descriptions of these computations are equivalent to descriptions of physical systems interacting with their environment. Therefore, it provides additional constraints on what it means to *be*, as opposed to merely *simulate* a member of a certain class of computational systems.[5] This enables a non-trivial role for computational correlates in explanations of consciousness, without having to accept what David Chalmers (2011) calls the *thesis of computational sufficiency* (according to which the right computational structure is sufficient for mind and consciousness). In particular, our account is compatible with the possibility of synthetic phenomenology, but does not entail that every computer that performs the right computations is conscious.

Although FEP itself does not provide an explanation of consciousness, it provides a quantitative framework within which models and measures of consciousness can be developed. More specifically, process theories conforming to FEP, such as active inference (Friston et al., 2017), can facilitate formulating and testing hypotheses about necessary computational mechanisms that may underpin different cognitive capacities, or different measures of consciousness. Because of the broad applicability of the framework (for a recent overview, see Costa, Parr, et al., 2020), completely different cognitive or formal properties associated with consciousness can be modelled and analysed. Since all models that conform to FEP describe the phenomena that are modelled as ways of minimising free energy, FEP provides a unifying perspective (Hohwy & Seth, 2020) from which it becomes tractable to determine what necessary computational mechanisms the different phenomena have in common (Wiese, 2018).

(function of a function) is an upper bound on surprisal. Equivalently, the negative free energy constitutes a lower bound on model evidence (a.k.a. marginal likelihood) in Bayesian statistics.

[5]This means FEP not only provides the foundation of a framework for consciousness, but also building blocks for a theory of consciousness. This contrasts with the proposal in Hohwy & Seth (2020), according to which predictive processing (PP) only provides a framework for research on consciousness (and NCCs), without itself being a theory of consciousness. (Although Hohwy and Seth suggest that "aspects of PP may themselves coalesce into a theory of consciousness in its own right," Hohwy & Seth, 2020, p. 24.)

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

The rest of this paper is structured as follows. In section 2, we review the canonical notion of an NCC, as defined by Chalmers (2000). We highlight three challenges for research on *neural* correlates and suggest that these challenges can be overcome by research on *computational* correlates of consciousness (CCCs). In section 3, we discuss research on CCCs (section 3.1) and argue that some computational principles associated with consciousness (Cleeremans, 2005) are already entailed by FEP (or by FEP and further, general assumptions; section 3.2). This is one reason why CCCs should be regarded as necessary conditions for consciousness (not sufficient conditions; section 3.3). Furthermore, we take up a special case of a challenge identified in section 2, viz. the challenge of "islands of awareness" (Bayne et al., 2020). We argue that FEP furnishes a constraint on the ascription of consciousness in such unusual cases: according to FEP, such systems must minimise the statistical complexity of their internal model (section 3.4). Since empirical evidence suggests that activity in conscious systems has high dynamical complexity, we briefly discuss how these notions relate to one another and suggest that it should be possible to define dynamical complexity in terms of a system's extrinsic information geometry (section 3.5). The benefit of such a definition is that it would provide not just a fundamental theoretical motivation for associating consciousness with dynamical complexity, but also the basis for a computational *explanation* of consciousness. The explanatory value of FEP is a matter of controversy. Hence, in section 4, we explain what contribution FEP would make to a computational explanation of consciousness. Finally, in section 5, we discuss its potential role in developing a minimal unifying model of consciousness, following a suggestion made in Wiese (2020).

2 From neural correlates to computational correlates of consciousness

According to the canonical notion of an NCC, as defined by Chalmers (2000), "[a]n NCC is a minimal neural system N such that there is a mapping from states of N to states of consciousness, where a given state of N is sufficient under conditions C, for the corresponding state of consciousness" (Chalmers, 2000, p. 31). A classic example of an NCC is provided by Crick & Koch (1990), who hypothesized that synchronised oscillations (in a range around 40–70 Hz) in parts of visual cortex may underpin visual consciousness. The neural system N would be visual cortex (or parts of it). States of – or, rather, activity in – visual cortex corresponding to visual experiences would be synchronised oscillations within a certain frequency.

Note that the definition cited above is a general definition for neural correlates of states of consciousness. That is, it can be applied to neural differences between consciousness and the absence of consciousness, as well as to neural differences between background states, such as wakefulness, dreaming, or hypnosis, or between particular conscious experiences, such as smelling coffee, hearing a voice,

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

seeing a face, or seeing a house. In addition to this general notion, Chalmers (2000) proposes a more specific definition for neural correlates of individual *contents* of consciousness. In what follows, we shall focus on the general definition, because some parts of our discussion mainly apply to neural correlates of global states of consciousness (e.g., being conscious vs. unconscious, being awake vs. dreaming, etc.).

We entertain three challenges for Chalmers' general definition of an NCC. The aim of this discussion is not to argue that the definition is problematic.[6] Instead, we wish to highlight the virtues of computational correlates of consciousness (CCCs, Cleeremans, 2005; Reggia et al., 2016). CCCs, as understood here, are not defined in terms of neural structures, but in terms of computational properties that can be instantiated by different substrates. CCCs are thus more general and promise to be more explanatory than NCCs. However, they also pose challenges of their own.

2.1 Challenge 1: Global dynamics

The definition of an NCC cited above is meant to pick out core correlates of consciousness. An implicit hypothesis is that the difference between neural activity that is associated with consciousness, and neural activity that is not, can be specified in terms of a difference in the neural systems involved, as well as by differences in the type of activity ("states") instantiated by these systems. This hypothesis is challenged by theoretical considerations, according to which *states* cannot be mapped to conscious experiences, but only trajectories (Fekete & Edelman, 2011); furthermore, it is in tension with considerations that privilege global (as opposed to local) processes (Mashour et al., 2020; Tononi et al., 2016).[7]

Recent empirical approaches support such theoretical considerations, by highlighting *global dynamics* that are associated with consciousness. Examples include differences in functional connectivity (Huang et al., 2020; Luppi et al., 2019) or in the power spectral density of resting EEG (Colombo et al., 2019; He et al., 2010). A further example is given by research on differences between global neural dynamics and state transitions during wakefulness and non-REM sleep (Stevner et al., 2019), or global dynamics during wakefulness and different types of unrespon-

[6] Some of these challenges are explicitly mentioned in Chalmers (2000) already; for discussions of Chalmers' definition, and NCC research more generally, see Noë & Thompson (2004); Block (2005); Miller (2007); Tononi & Koch (2008); Hohwy (2009); Seth (2009); Aru et al. (2012); Tsuchiya et al. (2015); Koch et al. (2016); Fink (2016); Mashour (2018).

[7] Chalmers' definition of an NCC is compatible with this possibility: it may turn out that the minimal neural system that is sufficient for states of consciousness is the entire brain. However, the possibility also illustrates that NCCs, defined in terms of neural systems, may be less informative if consciousness mainly depends on how neural activity evolves over a certain period of time, and depends less on where in the brain the neural activity takes place. The virtue of computational correlates is that they can be informative regardless of whether they describe properties of the entire brain or of small regions in the brain.

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

siveness, such as anaesthesia, unresponsive wakefulness syndrome, or minimally conscious states (Demertzi et al., 2019).

Computational correlates are not defined in terms of neural structures, and hence are neutral with respect to the question whether conscious experiences correlate with global neural activity or not. Furthermore, they are compatible with the possibility that neural *state transitions* are more important than the states themselves. Hence, computational correlates promise progress on the challenge of accounting for global dynamics associated with consciousness.

2.2 Challenge 2: Non-arbitrary mappings

A further challenge consists in determining *why* a particular neural structure N is an NCC of consciousness (see Hohwy & Seth, 2020). Chalmers' (2000) definition only requires that there be a mapping between neural states and states of consciousness. This mapping can, as Chalmers himself points out, be "seemingly arbitrary" (Chalmers, 2000, p. 23). The mapping can be arbitrary in the sense that there seems to be no epistemically necessary connection between certain neural states and states of consciousness; however, such mappings *will be* non-arbitrary in the sense that the connection must be general and lawful: states of an NCC, as defined by Chalmers, must be *sufficient* for states of consciousness, given certain *general conditions*. Finding arbitrary mappings between neural states and conscious experiences can help illuminate how consciousness arises in human beings, but it will not lead to a deep understanding of consciousness.

Chalmers (2000, p. 23) suggests that a non-arbitrary mapping could be established if not only neural states themselves are mapped to states of consciousness, but if also relations between neural states are mapped to relations between conscious states (a recent approach pursuing this idea using category theory is presented in Northoff et al., 2019). Drawing on such a mapping, one can extrapolate from existing NCCs to predict how certain changes in neural activity will change the correlated conscious experience (a similar idea, though not using the notion of an NCC, is explored in Churchland, 2005).

CCCs provide another way of increasing explanatory and predictive power. If it is shown that neural activity associated with consciousness implements certain computations, then this provides an explanation of *why* that activity is associated with consciousness. In principle, this can then be used to predict the presence (or absence) of consciousness in novel (or controversial) cases. This brings us to the final challenge we will consider here.

2.3 Challenge 3: Unusual conditions

A problem for research on NCCs consists in specifying the conditions under which a correlation between neural and conscious states exists. Chalmers' (2000) definition circumvents this problem to some extent, in that his definition only requires

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

that the states of an NCC be *sufficient* for conscious states. It does not require that they be necessary. If, by contrast, a perfect positive correlation between states of an NCC and states of consciousness were required, then states of an NCC would also have to be necessary.

However, even a (mere) sufficiency requirement is challenged by unusual cases. Perhaps a certain type of neural activity goes along with consciousness in ordinary cases, but does not when brain functioning is affected by lesions? Chalmers anticipates this challenge and restricts the required mapping between neural states and conscious states (i) to ordinary functioning brains in ordinary environments, (ii) with unusual inputs, and (iii) under local brain stimulation.

Although this restriction creates a safe methodological footing for research on NCCs, it also comes with a limitation. Ideally, we would like to go beyond usual cases, and make inferences about unusual cases. For instance, we would like to know whether an unresponsive patient is non-conscious, or whether they are just unable to report their conscious experience. Failing to find neural activity that is sufficient for consciousness cannot rule out that the patient is conscious. This would be different if an NCC provided a necessary condition for consciousness.

An extreme case of an unusual condition is exemplified by the possibility of consciousness in the complete absence of both sensory stimulation and observable behaviour. Bayne et al. (2020) call such cases "islands of awareness" (IOA). The authors define an IOA as a "conscious stream (or system) whose contents are not shaped by sensory input from either the external world or the body and which cannot be expressed via motor output" (Bayne et al., 2020, p. 7). Cases such as dreaming or locked-in syndrome provide an approximation to IOAs (but not complete IOAs), in which some connections to the environment (via sensory input or motor output) remain. Genuine IOAs could, as the authors point out, be found in *ex cranio* brains, hemispherotomy, or cerebral organoids (Bayne et al., 2020, pp. 6–11). An *ex cranio* brain is a brain that has been extracted from the cranium and is kept alive outside of the body (post-mortem). An example can be found in a study by Vrselja et al. (2019), who extracted pig brains and observed "spontaneous synaptic activity" (Vrselja et al., 2019, p. 336) in them. In hemispherotomy (De Ribaupierre & Delalande, 2008), a cortical hemisphere is more or less disconnected from the rest of the brain, leaving only vascular connectivity intact. A cerebral organoid is a neural structure grown in the lab (Lancaster et al., 2013).

How would one determine whether conscious experience is sustained in such isolated systems? NCCs can provide some guideline, at least if activity that has been found to be sufficient for consciousness in ordinary cases is measured in IOAs, as well. However, it would be desirable to have additional criteria for the ascription of consciousness, ideally ones that are also necessary. Bayne et al. (2020) suggest using measures of dynamical complexity, which have been found to be reliable proxies for consciousness (and its absence, respectively), across a variety of conditions (Demertzi et al., 2019; Li & Mashour, 2019; Schartner et al., 2015; Schartner et al., 2017; Sitt et al., 2014).

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

For this reason, we shall take a closer look at the concept of dynamical complexity in the following section. In particular, we will discuss whether it can be regarded as a computational correlate of consciousness. We will also revisit the challenge of IOAs, as this challenge is especially helpful for discussing the relevance of FEP to research on NCCs and CCCs: an application of FEP to IOAs seems to suggest that IOAs must minimise complexity, whereas empirical research suggests that conscious systems (and hence also IOAs) display high complexity. We shall argue that this apparent tension is resolved upon a closer look; in fact, free energy minimising systems are *ipso facto* capable of producing activity with dynamical complexity – and certain kinds of systems are stipulatively defined by their complex dynamics.[8]

3 Computational correlates of consciousness and the free energy principle

A computational correlate of consciousness (CCC) is a set of computational properties that are associated with consciousness. Such properties are specified by computational models. Below we will argue that CCCs should be construed as necessary conditions for consciousness. This constitutes a major departure from the concept of an NCCs, which is usually regarded as a (minimally) sufficient condition for consciousness. As we point out below (in section 3.2), this conforms to how the notion of a CCC is construed in the literature. Furthermore, necessary conditions for consciousness have some advantages that merely sufficient conditions do not have (in particular, they can be used to infer the absence of consciousness, viz., if a necessary condition is not fulfilled in a given case, see Fink, 2016).

Although characterising CCCs in terms of "computational properties associated with consciousness" is relatively general and vague, it already shows that the first two challenges identified above can be overcome by CCCs. A CCC can be specified by a computational model of global neural dynamics, thereby meeting the first challenge. In addition, CCCs are non-arbitrary in the sense that they describe computational functions that are realised by neural activity, which can help explain cognitive capacities associated with consciousness (see section 5 below for more details). The second challenge can therefore also be met by CCCs.[9] In order to enrich the general characterisation of a CCC just given, we first review an influential paper by Cleeremans (2005) on CCCs. We then connect some of the ideas presented in that paper to FEP and explain how the notion of a CCC would be conceived from the point of view of FEP.

[8] There can still be large differences in the *level of complexity* displayed by different systems. That is, differences in the level of complexity may still account for differences in consciousness.

[9] Of course, CCCs may still seem arbitrary in the sense that they do not directly address the hard problem (Chalmers, 1995). A complete computational explanation of consciousness, based on CCCs, must therefore also address the meta-problem (Chalmers, 2018).

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

3.1 CCCs and computational principles

Cleeremans (2005) takes the "contrastive approach to consciousness" as a starting point and argues that the study of the correlates of consciousness must go beyond differences in neural activity, by including differences in computations (and behaviour) that are associated with consciousness (see Cleeremans, 2005, p. 84). Following Mathis & Mozer (1996), he refers to these computations as the *computational correlates of consciousness*. Based on this idea, Cleeremans suggests two relatively general "computational principles" associated with consciousness.

The first principle, "quality of representation," is defined in terms of a representation's "stability in time, strength, and distinctiveness" (Cleeremans, 2005, pp. 91–92). The stability of a representation corresponds to the amount of time for which its contents are available. The strength of a representation is a property of its vehicles and corresponds to the number of "processing units" of which it consists, as well as of their activation strength (Cleeremans, 2005, p. 92). Distinctiveness is primarily a content property, namely its specificity. According to Cleeremans' second principle, consciousness involves meta-representations.

Why are these two computational principles important? One reason is that they refer to properties that can help explain how presumed functions of consciousness are realised. The functions mentioned by Cleeremans (2005) include: "flexible, adaptive control over action," simulating possible actions and their consequences, and "error-correcting functions" (Cleeremans, 2005, p. 85). High-quality representations, as characterised by the first principle, are required for control over action and planning, and error correction requires meta-representation (because representations of error are meta-representational, see Shea, 2012b).

The benefit of these general computational principles is that they make relatively little assumptions about consciousness but can still provide some guidance for specific computational models of consciousness. Furthermore, they can suggest why a given type of activity correlates with consciousness: if the neural populations involved can be interpreted as a representation having the properties described by the two principles, and if this helps explain a given type of behaviour (a behavioural correlate of consciousness), then we understand (to some extent), why the measured neural activity is associated with consciousness.

The generality of these computational principles also has a disadvantage: the properties described by them may be necessary for many types of conscious experience, but probably not sufficient. In particular, stability, strength, and meta-representation may be individually necessary, but it is unclear whether any combination of (individually necessary) CCCs would be sufficient for consciousness. In other words, a complete explanation of consciousness would require more (Cleeremans et al., 2020). At the same time, it could even be questioned whether the properties are necessary. In order to address this question, we show to what extent, and under what assumptions, some computational principles can be derived from first principles.

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

3.2 Deriving computational principles from first principles

Here, we shall succinctly summarise the gist of results presented in more detail elsewhere (see especially Friston, 2019b; Friston, Wiese, et al., 2020). This will show that some properties highlighted by the two computational principles given above are already instantiated by a very general class of systems, comprising single-cell organisms, but also more complex systems, such as human beings. This underscores that the computational principles reviewed above are not sufficient for consciousness (unless one is willing to ascribe consciousness to cellular organisms and similarly simple types of system).

A general (coarse-grained) form in which a random dynamical system can be described is in terms of Langevin dynamics (Ao, 2008; Seifert, 2012; Sekimoto, 1998):

$$\dot{x}(\tau) = f(x, \tau) + \omega.$$

Here, the state of the system $x(\tau)$ at time τ is constituted by slowly-changing macroscopic variables, which are grounded in microscopic variables with faster dynamics. This is why the result is a *stochastic* differential equation, comprising not only the state-dependent flow f, but also a stochastic term ω (which is a Gaussian stochastic term with mean equal to zero and covariance 2Γ – so Γ is the amplitude of the random fluctuation).

The fundamental assumption we shall make here is that things that persist must have measurable properties that remain invariant over a certain amount of time. Formally, this means that the system's long-term behaviour can be described with reference to a global random attractor (or *pullback* attractor). Put differently, there is a (random) set of states to which the system will converge. Since the attractor is a random set, it can be described in terms of a density, its non-equilibrium steady-state (NESS) density. The NESS density is a solution to the Fokker-Planck equation (for details, see Friston, 2019b):

$$\dot{p}(x, \tau) = \nabla \cdot (\Gamma \nabla - f) p(x, \tau)$$

The NESS density will be a density $p(x, \tau)$ that does not change over time, i.e.

$$\dot{p}(x, \tau) = \nabla \cdot (\Gamma \nabla - f) p(x, \tau) = 0.$$

A solution to the Fokker-Planck equation satisfying $\dot{p}(x, \tau) = 0$ is

$$p(x, \tau) = \exp(-\Im(x, \tau)),$$

where $\Im(x)$ is *self-information* (or surprisal):

$$\Im(x) = -\ln p(x).$$

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

This means we can express the flow f in terms of surprisal (and hence, in terms of the NESS density):

$$\begin{aligned} f(x) &= (Q(x) - \Gamma) \cdot \nabla \mathfrak{I}(x) \\ &= (\Gamma - Q(x)) \cdot \nabla \ln p(x). \end{aligned}$$

Here, Q denotes solenoidal flow. Furthermore, following a variational principle of least action, the ensuing dynamics are a gradient flow on surprisal. In other words – on average – the dynamics minimise self-information (surprisal) $\mathfrak{I}(x)$ (Friston, 2019b, pp. 12–13).

To recap, anything that exists will have characteristic features that do not change over time. Formally, we express this by assuming that the system is a random dynamical system in non-equilibrium steady state (NESS). This means that the system is in exchange with its environment and that its dynamics can be described in terms of its NESS probability density, i.e., a probability distribution that does not change during the time over which the system is said to exist. Such a system has a random attractor, which can be thought of in two ways: first, it can be considered as the trajectory of systemic states as they evolve over time. Here, the key aspect is that the system will, after sufficient time, revisit particular regions of its state space, i.e., those regions that constitute the attracting set. The second interpretation considers the random attractor as subtending a probability density p over the states that the system will be found in, when sampled at random. The density dynamics are described by the Fokker-Planck equation. In nonequilibrium steady state, this density does not change with time. The solution (i.e., setting the equation to zero) shows that there is a lawful relationship between the flow of states f and the probability density p.

A further assumption we shall make is that the system possesses a Markov blanket. A Markov blanket can be conceived as a boundary between a system and its environment. It is defined as that set of states that statistically separate a system's internal states from external states: the system's internal states μ are conditionally independent of external states η, given the Markov blanket b. Here, the Markov blanket $b = \{s, a\}$ comprises both sensory states s (which are not directly influenced by internal states) and active states a (which are not directly influenced by external states). Note that internal, external, and blanket states jointly constitute the systemic states $x = \{\eta, s, a, \mu\}$.

The term "Markov blanket" was originally coined by Pearl (1988), in the context of Bayesian networks, i.e., directed acyclic graphs. However, the notion can also be applied to directed cyclic graphs (i.e., dependency networks) and undirected graphs (i.e., Markov random fields). For examples of applications of the notion, see Clark (2017), Palacios et al. (2017), Parr et al. (2020), Pellet & Elisseeff (2008).[10]

[10] In fact, there is an ongoing discussion about how the Markov blanket construct, as used in Friston (2013), Friston (2019b), Friston, Wiese, et al. (2020), and related works, differs from the original definition presented in Pearl (1988). The original definition only considers synchronic properties,

Here, what matters is that the existence of a Markov blanket has subtle, but profound implications.

The fact that internal states are conditionally independent of external states, given blanket states, allows us to disentangle the system's dynamics. In other words, we can write down separate equations of motion for internal states and active states – that only depend on blanket states:

$$\begin{aligned} f_\alpha(\pi) &= (Q_{\alpha\alpha} - \Gamma_{\alpha\alpha})\nabla_\alpha \mathfrak{I}(\pi) \\ \alpha &= \{a, \mu\} \quad \text{(autonomous states)} \\ \pi &= \{s, \alpha\} \quad \text{(particular states)} \end{aligned}$$

Here, $f_\alpha(\pi)$ is the flow of internal and active states α, to which we refer as *autonomous states*. Note that the flow only depends on particular (non-external) states $\pi = \{s, \alpha\}$. This in turn enables us to express the dynamics of internal states in terms of a probability distribution over external states – encoded by internal states. That is, for any blanket state b, we can map internal states μ to a probability distribution $q_\mu(\eta)$ over external states, given blanket states.

Strictly speaking, we can only map expected internal states, because there need not be a unique internal state, associated with a given blanket state. In practice, one must therefore consider averages of internal states (such as average activity of neural populations), which approximate their expected value.

If the probability distribution $q_\mu(\eta)$ is sufficiently similar to the actual conditional distribution $p(\eta|\pi)$ over external states (where p is the NESS density), then we can approximate the flow of autonomous states using $q_\mu(\eta)$. Equivalently, we can say that the gradient flow of autonomous states can be considered as a gradient flow on variational free energy F:

$$\begin{aligned} f_\alpha(\pi) &\approx (Q_{\alpha\alpha} - \Gamma_{\alpha\alpha})\nabla_\alpha F(\pi), \text{ with} \\ F(\pi) &\triangleq E_q[\mathfrak{I}(\eta, \pi)] - H[q_\mu(\eta)] \\ &= \mathfrak{I}(\pi) + D[q_\mu(\eta)||p(\eta|\pi)] \\ &= E_q[\mathfrak{I}(\pi|\eta)] + D[q_\mu(\eta)||p(\eta)] \geq \mathfrak{I}(\pi). \end{aligned}$$

According to this way of expressing $f_\alpha(\pi)$, the system's autonomous states α must change in such a way that the variational free energy $F(\pi)$ is minimised – where variational free energy is a functional of the probability distribution $q_\mu(\eta)$ encoded by internal states. Hence, as internal states change, $q_\mu(\eta)$ changes, as well; more specifically, it changes in a way that makes it more similar to $p(\eta|\pi)$ (as measured by the Kullback-Leibler divergence $D[q_\mu(\eta)||p(\eta|\pi)]$).

Changes of internal states therefore go along with changes in the probability distribution encoded by internal states. We can thus map a given internal state

whereas its application in the context of FEP involves diachronic properties (i.e., it is meant to formalise dependencies in an action-perception loop). For discussion, see Rosas et al. (2020); see also Friston, Costa, et al. (2020); Friston et al. (2021).

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

to a point on a statistical manifold, and associate changes of internal states with movement on the statistical manifold. This manifold has an information geometry, which we shall call the *extrinsic* information geometry (because it is the geometry of probabilistic beliefs about *external* states). The extrinsic information geometry is thus associated with the probability distributions *encoded by* internal states. Now note that the NESS density also specifies a probability *of* internal states. We can regard it as a point on another statistical manifold, which has a geometry we shall call the *intrinsic*[11] information geometry.

If we parametrise the density over internal states with time, we can associate the progression of time with a movement on the system's intrinsic manifold. However, since the NESS is constant, movement on the intrinsic manifold will eventually stop (once the system has reached non-equilibrium steady state). In other words, if – as physicists like to say – we "prepare" the internal states in some initial configuration, and then "push" the probabilistic configuration forwards in time, the internal states pass through a series of probabilistic configurations until they reach the point on the intrinsic statistical manifold corresponding to the NESS. Furthermore, if we assume some of the internal states are exchangeable (e.g., one neuron in a neuronal population can be treated as the same as another from the same population), we equip internal states with a statistical mechanics or thermodynamics; i.e., the initial configuration will, over time, converge to the NESS and *thermodynamic* free energy will fall progressively. In short, the intrinsic geometry underwrites the dynamical complexity and thermodynamics of neuronal populations.

Conversely, on the extrinsic manifold, the internal states minimise *variational* free energy, when conditioned on the blanket states. In other words, whenever the blanket states change, the expected internal state changes and there is movement on the extrinsic manifold that – by construction – can be cast as Bayesian belief updating. This is because the (expected) internal states encode probabilistic (i.e., Bayesian) beliefs about external states. The blanket states here can be construed as the sensory impressions of external states on the Markov blanket that contains the internal states. In short, it may be more useful to consider the system from the point of view of the extrinsic information geometry (probabilities encoded by internal states) than from the point of view of the intrinsic information geometry (probabilities of internal states).

We can now relate some of the computational principles of conscious processing, as suggested by Cleeremans (2005), to the probability density dynamics just sketched. In particular, we shall consider the "strength" of neural representations (the number of "processing units" that constitute the representational vehicle) and meta-representations. Firstly, recall that only expected internal states can be said

[11] It is not just intrinsic in the sense that it is about internal states, it is also intrinsic in the sense that it does not depend on the manifold associated with the probabilities encoded by internal states. The extrinsic information geometry, by contrast, presupposes the existence of the manifold associated with probabilities of internal states.

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

to encode a probability distribution over external states. If we want to interpret actual neural activity as encoding a probability distribution, we therefore have to consider population averages. Any neural representation will therefore have high "strength", from the point of view of FEP. Secondly, note that – under Gaussian assumptions about the encoded probabilistic beliefs – minimising variational free energy becomes prediction error minimisation (Friston & Kiebel, 2009). Prediction error signals are meta-representational (Shea, 2012b), so the feature of meta-representation is easily fulfilled by many systems.[12]

We draw the following lessons from these observations. (i) The activity of a large class of systems conforms to at least some of the computational principles that are identified as computational correlates of consciousness in Cleeremans (2005). In particular, this is true for many non-conscious systems (see also Reggia et al., 2016, p. 111). (ii) This suggests, more generally, that computational correlates should not be regarded as minimally sufficient conditions for consciousness (in contrast to Chalmers' NCCs, as presented in his 2000). In particular, computational correlates need not be associated with the difference between conscious and unconscious processing (*pace* Cleeremans, 2005). (iii) Instead, we should expect computational correlates to specify *necessary* conditions for consciousness.[13] For instance, Reggia et al. (2019) present a model of working memory and propose that CCCs are identified by the "underlying computational mechanisms [that] are critically needed to realize such a model and distinguish it from other contemporary neural networks in general" (Reggia et al., 2019, pp. 265–266). This is a useful feature, as not fulfilling a necessary condition can license an inference to the absence of consciousness (Fink, 2016). Furthermore, existing evidence for the presence of consciousness in controversial cases can be strengthened, if necessary conditions for consciousness are fulfilled.

3.3 Computational correlates as necessary, sufficient, or necessary *and* sufficient conditions for consciousness?

We noted, almost in passing, that some of the principles identified as computational principles by Cleeremans should be regarded as merely necessary conditions for

[12] In fact, one could argue that all systems that exist – in the above sense – are meta-representational on this definition. This follows because the gradients of variational free energy can always be expressed as a form of prediction error. One could object that this only entails (meta)representations in an instrumentalist sense, i.e., we can describe internal states of the system *as if* they were representations (Ramstead et al., 2020). In Wiese & Friston (2021), we argue that at least living organisms that embody a hierarchical generative model are representational systems in a realist sense. However, we are aware that this is a controversial claim (see Bruineberg et al., 2018; Downey, 2018; Hutto, 2018; Kirchhoff & Robertson, 2018; Van Es, 2019).

[13] Some of them may be trivial (e.g., single-cell organisms may fulfill them). A challenge is to find computational correlates that specify non-trivial necessary conditions. One strategy is to focus on computational properties associated with functions of consciousness (see Cleeremans, 2005, and section 5).

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

consciousness. Even if one agrees with this remark (since, otherwise, one would have to accept that many very simple systems are conscious), one could argue that computational correlates *should* be sufficient conditions for consciousness. That is, Cleeremans' principle should then only be regarded as part of a bundle of computational principles that (together) yield sufficient conditions for consciousness. In particular, according to this line of reasoning, the search for computational correlates of consciousness should consist in searching for necessary and sufficient conditions for consciousness.

We disagree. We argue that even finding merely necessary conditions for consciousness can be useful and important in the science of consciousness. First of all, necessary conditions for consciousness can complement "minimally sufficient" conditions for consciousness. That is, the search for necessary conditions need not replace the search for minimally sufficient conditions. Secondly, whether sufficient or necessary conditions are more useful depends on the domain of application. For instance, if one already knows that a creature is conscious (e.g., a human being with a normally functioning brain in alert wakefulness), minimally sufficient conditions for consciousness *in this type of creature* will be most informative. However, if it is unclear whether a creature (or machine) is conscious or not, determining whether necessary conditions for consciousness are fulfilled can be immensely informative. If a necessary conditions is not fulfilled, the question ("Is it conscious or not?") can even be settled.[14] Thirdly, necessary conditions may provide unification. For instance, if CCCs are investigated by modelling different cognitive capacities that are associated with consciousness, necessary computational mechanisms of these different capacities can reveal what they have in common. (And it is to be expected that they have something in common if they are all enabled, or facilitated, by the same phenomenon, i.e., consciousness. See Birch, 2020, pp. 8–9.) We shall elaborate on this point in section 5 below.

3.4 The challenge of islands of awareness

We can now return to a special case of challenge 3 (identified in section 2 above), viz. the possibility of what Bayne et al. (2020) call "islands of awareness" (IOA): conscious experiences that are instantiated in the absence of sensory stimulation and observable behaviour. Most measures of consciousness rely on some kind of neural response to sensory stimulation, which is precluded in the case of IOAs. From the point of view of FEP, the challenge may seem even harder. FEP has

[14] In a related vein, Fink (2016) discusses four problems with Chalmers' definition of an NCC, one of which is that sufficient conditions do not enable one to rule out that a system is conscious. Fink argues: "in order to overcome some of the problems of the Chalmers-NCC, it seems that we have to introduce necessity somewhere" (Fink, 2016, p. 9). Fink suggests identifying a feature bundle $\mathbb{F}$ that is shared by different neural state tokens that are (each) sufficient for the same phenomenal state type. Crucially, $\mathbb{F}$ must be a necessary condition for consciousness. If the feature bundle consists of neurofunctional and neurocomputational features, then it may also be applicable to many non-human animals, or even machines (Fink, 2016, p. 11).

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

been developed for systems that sense and act, not for systems that do not receive sensory input and are incapable of producing motor output. In response to this challenge, we will propose to treat IOAs in analogy to brains in altered states of consciousness, such as hallucination or sleep and dreaming. As previously argued (Friston, Wiese, et al., 2020; Hobson & Friston, 2012; Hobson & Friston, 2014), neuronal dynamics during sleep minimise the complexity of the brain's generative model (i.e., by removing redundant model parameters). This can be seen by noting that variational free energy can be expressed as follows:

$$F(\pi) = E_q[\mathfrak{I}(\pi|\eta)] + D[q_\mu(\eta)||p(\eta)].$$

Here, $D[q_\mu(\eta)||p(\eta)]$ describes the statistical complexity of the internal model ("How much do I have to change my mind to accommodate current changes in sensory signals?"), whereas $-E_q[\mathfrak{I}(\pi|\eta)]$ describes accuracy (so minimising free energy entails minimising the difference between accuracy and complexity). Note that accuracy depends on the particular states π (comprising blanket states and internal states). During a period of (partial) disconnection from blanket states, such as dreaming, the variational free energy gradient in the following equation will mainly be driven by the complexity part (which does not depend on blanket states):

$$f_\alpha(\pi) \approx (Q_{\alpha\alpha} - \Gamma_{\alpha\alpha})\nabla_\alpha F(\pi).$$

This entails that neural processes can perform the computations required to minimise free energy, even when the coupling with the environment is temporarily suspended. Using sleep and dreaming as a model, FEP can be applied to IOAs and yields the same result: isolated systems will minimise free energy by reducing the complexity of the generative model.

This creates a seeming tension between FEP and the suggestion by Bayne et al. (2020) that high complexity could serve as a proxy for consciousness in IOAs. Measures used to infer the presence of consciousness in altered states (Rohaut et al., 2017; Sarasso et al., 2014; Sarasso et al., 2015) are measures of entropic complexity such as Lempel-Ziv (LZ) complexity (Lempel & Ziv, 1976) that provide an upper bound on algorithmic complexity (Ruffini, 2017) – where the algorithmic complexity of a signal is defined as the length of the shortest algorithm that produces the signal. This is exactly the same quantity that underwrites variational free energy.

In fact, variational free energy minimization in machine learning was predicated on minimizing algorithmic (computational) complexity (MacKay, 1995; Wallace & Dowe, 1999), which is the basis of universal computation (Hutter, 2005). Consciousness researchers assume complexity has to be large and yet the imperatives for universal computation, in general – and the free energy principle, in particular – say exactly the opposite.

This apparent paradox can be resolved easily by noting that free energy is a bound on marginal likelihood (a.k.a. model evidence) in the same way that the

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

LZ complexity is an upper bound on algorithmic complexity. The logarithm of evidence can always be expressed as accuracy minus complexity. In brief, if consciousness entails belief updating (technically, movement on a statistical manifold) then the imperative for this movement is to minimise the difference between accuracy and complexity. On this view, complexity provides an incomplete metric for any CCC because it fails to account for accuracy, i.e., the marginal likelihood of the sensorium. In the rhetoric of, e.g., integrated information theory (Tononi et al., 2016) it is the "matching" that matters not the complexity (where matching is scored by model evidence). In a deeply structured world, an accurate account of the sensorium can only be supplied by a complex (Bayesian) belief about the causes of sensations. This means that complexity is always chasing accuracy, so that it will look as if complexity is being maximized. Technically, however, it is the difference between complexity and accuracy that is being minimized.

Sleep and related IOA present an interesting scenario. Because there is a transient suspension of exchange with the sensorium (or other parts of the brain) it would appear that there is no accuracy (because there are no inputs to provide an accurate account of). However, these IOA are themselves accountable to a larger context; namely, all the inputs that have been experienced – and will be experienced. On this view, it is then no surprise to see that a good model of a complex world will show a high degree of algorithmic or statistical complexity, even when observed in temporary isolation.

This perspective provides fundamental constraints on (the utility of) complexity measures of consciousness. High complexity (as measured by LZ complexity) is necessary for consciousness in a complex world, but it is not sufficient. Under FEP, the minimisation of free energy constitutes a necessary condition for consciousness. This second condition entails the first condition when, and only when, a high degree of complexity is necessary to explain the sensorium. Hence, evidence for the presence of consciousness in IOAs can be garnered from at least two sources: First, from measurements of apparent algorithmic complexity (such as LZ complexity). Secondly, from models of IOAs that show such systems minimise free energy when their insular nature disappears.

3.5 Dynamical complexity as a CCC?

The proposed solution to the problem posed by potential IOAs is slightly inelegant. While it is pleasing that it does not constitute a problem for FEP per se, it may still seem that FEP can only make a tiny contribution to understanding such cases (or to "measuring consciousness"). This is because measures of dynamical complexity (such as LZ complexity of neural time series) may seem to provide much stronger evidence than any criteria derived from FEP: existing empirical results already suggest that dynamical complexity is a correlate of consciousness (Demertzi et al., 2019; Li & Mashour, 2019; Schartner et al., 2015; Schartner et al., 2017; Sitt et

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

al., 2014). Hence, it is not unreasonable to extrapolate from this to unusual cases (Bayne et al., 2020), perhaps even without requiring any additional evidence.

Let us take a step back and ask: "What is dynamical complexity?" – rather than: "How can it be measured?" In dynamical systems theory, three types of complex behaviour can be distinguished: chaotic itinerancy, heteroclinic cycling, and the switching under multistability (Friston, 2019a). All three types of behaviour involve meandering movements between regions that are revisited time and time again, but always following slightly different, unpredictable trajectories. A difference between these types of complex behaviour rests on how they are brought about. Either the system switches between basins of attraction (due to noise, as in multistability, or due to the proximity of basins of attraction, as in chaotic itinerancy), or the system's state space contains a circle of connected saddles with both attracting and repelling manifolds.

FEP provides an insight into the conditions under which stochastic dynamical systems display behaviour that is characteristic of dynamical complexity (i.e., behaviour such as chaotic itineracy, heteroclinic cycling, or multi-stability and switching). As we saw above, any system with a Markov blanket is trying to minimise its self-information, i.e. surprisal. One can show that this brings about a tendency to self-organised criticality and dynamical complexity (Friston, 2019a; Friston et al., 2012). At the same time, there is a relation between minimising surprisal of blanket states and minimising statistical (and algorithmic) complexity[15]: the very process by which a system minimises surprisal of blanket states can, equivalently, be described as the process of minimising variational free energy with respect to the internally encoded recognition density. Recall that variational free energy can be described as statistical complexity minus accuracy. Hence, if a system minimises surprisal of blanket states by minimising variational free energy, it will also tend to keep statistical complexity within bounds.

Developing a measure of dynamical complexity within the framework of FEP, i.e., a measure that invokes a system's extrinsic information geometry, would help determine under what conditions systems (that minimise variational free energy) display high dynamical complexity, and under what conditions they display low dynamical complexity. Furthermore, existing measures of dynamical complexity cannot be said to measure computational correlates of consciousness, from the

[15]The relation between surprisal minimisation and minimising algorithmic complexity (Kolmogorov complexity) is not straightforward. Heuristically, it can be shown that maximising the posterior probability of external states, given blanket states, $p(\eta|b)$, is equivalent to minimising the algorithmic complexity of blanket states (Wallace & Dowe, 1999). Since minimising variational free energy entails that the recognition density $q_\mu(\eta)$ approximates the posterior $p(\eta|b)$, there is at least an indirect link to algorithmic complexity. Furthermore, in AIXI, a formalism for artificial general intelligence, a reinforcement learning agent optimises its hypotheses about the environment by minimising their algorithmic complexity (Hutter, 2000, 2005). Since active inference (which involves minimising expected free energy) subsumes reinforcement learning as a special case (Costa, Sajid, et al., 2020), it should in principle be possible to describe an AIXI agent as an agent that minimises (expected) variational free energy, under a particular set of priors – this remains an important task for future work.

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

point of view of FEP. This is because the computations performed by a system that minimises variational free energy are *inferences*, which correspond to movements on the system's *extrinsic* statistical manifold. Existing measures of dynamical complexity, however, do not measure complexity in terms of probabilities encoded *by* neural activity, but in terms of statistical properties *of* neural activity, that is, properties of the *intrinsic* statistical manifold. Hence, dynamical complexity, as measured by existing approaches, should be associated with NCCs (at least from the point of view of FEP).

Interestingly, FEP suggests an alternative. An FEP-inspired measure of dynamical complexity would most naturally be based on the *information length* of neurally encoded probability distributions (i.e., Bayesian beliefs): instead of considering the statistical properties of a neural time series as such, one would have to consider the probability distributions encoded by this activity – or, equivalently, trajectories on the extrinsic statistical manifold. The information length of a trajectory then simply corresponds to the distance travelled on the manifold (which is measured at any point on the path by the Fisher information metric).

Even if a measure based on information length should turn out to be equivalent to existing measures of dynamical complexity, it would be advantageous, because it could unify existing approaches under a single overarching principle (i.e., the FEP). Note that, although many existing measures of dynamical complexity are based on LZ complexity (Demertzi et al., 2019; Li & Mashour, 2019; Sitt et al., 2014), some also use amplitude coalition entropy and synchrony coalition entropy (Schartner et al., 2015; Schartner et al., 2017), or the spectral exponent of resting EEG (Colombo et al., 2019) as a proxy. In fact, there is no obviously "correct" way of measuring dynamical complexity. To the extent that proposed measures of dynamical complexity have empirical validity (as measure of consciousness), they may be regarded as empirically adequate. But it would still be desirable to provide a fundamental justification because this could further support their application to controversial cases. In addition, it could help resolve an uncertainty about whether existing measures of dynamical complexity also track algorithmic complexity. It is well-known that LZ complexity, for instance, constitutes an upper bound on algorithmic complexity. However, this entertains the possibility that conscious activity maximises LZ complexity (and related measures of dynamical complexity), but minimises algorithmic complexity (this hypothesis has been stated by Ruffini, 2017.). Finally, developing a measure of dynamical complexity within FEP could help to provide a *computational explanation* of consciousness (or "explanatory correlates" of consciousness, see Seth, 2009; Seth & Edelman, 2009).

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

4 From computational correlates to a computational explanation of consciousness

Measuring neural activity in conscious creatures yields insights into correlates of consciousness. Here, we shall probe the idea that CCCs, construed as necessary conditions for consciousness, may actually yield a computational *explanation* of consciousness.

Do computational explanations of consciousness imply that implementing the right computations is sufficient for being conscious?[16] Or is it possible that some simulations of conscious systems perform all the computations performed by actual conscious systems, without being conscious? Here, we will argue that implementing the right computations is not sufficient for instantiating consciousness. But what would then count as a computational explanation of consciousness? We shall argue that the computations must be implemented by the right kind of system, and that FEP puts constraints on what the right kind of system is.

To illustrate, consider the following example.[17] A digital assistant (say, a sufficiently sophisticated successor of Siri or Alexa) may be able to pass a version of the Turing test, if it implements the right computations. But implementing these computations will not ensure its continued existence. In particular, it may be able to provide sensible answers to any questions it is asked – but it would also continue to exist if it were to respond with nonsensical answers. Such a system computes and may represent the world, but it is not committed to the existence of anything that corresponds to the way it represents the world (Smith, 2019). This is different for systems like us. Our continued existence is systematically related to the way we register (and interact with) the world. Similarly, a digital computer may implement all computations underpinning conscious experience. But it must also (have the potential to[18]) successfully interact with the world *by virtue of these computations*, in order to be conscious. We elaborate on this point below (section 4.2). Before that, we discuss a more fundamental problem associated with computational explanations of consciousness (section 4.1)

4.1 The scope of computational explanations

Computational explanations seek to explain a target phenomenon characteristic of a system (e.g., a particular type of behaviour) in terms of computations performed by that system. If no computational properties are sufficient for consciousness, then it seems there is something that is left out by computational explanations.[19]

[16] This is a special case of what Chalmers (2011) calls the *thesis of computational sufficiency*.

[17] This example was brought up in a discussion with Karen Konkoly's reading group; we are grateful for this suggestion.

[18] This clause is necessary to accommodate the possibility of islands of awareness.

[19] Note that this is not simply the explanatory gap (Levine, 1983). An explanatory gap arises if there is no necessary epistemic connection between the explanans and the explanandum. The absence

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

Let us consider this problem in a bit more detail. A notorious challenge for computational explanations is that they depend on an account of physical computation, i.e., an account of what it means for a concrete, physical system to perform a given computation (Milkowski, 2013; Piccinini, 2015). Such an account should rule out trivial implementations (Sprevak, 2019). For instance, if a computational explanation of consciousness specifies internal state transitions that are also instantiated by a rock or a piece of Swiss cheese, then it is doubtful whether describing a system as computing can explain why it is conscious. In other words, if a computational explanation focuses on abstract computational properties (disregarding how they are implemented in the system), then it must be complemented by an account of physical implementation, showing that the same computational properties cannot be instantiated by systems we would not regard as computational. There is, however, no consensus on whether any account of physical computation can avoid trivial implementations (Sprevak, 2019), nor on whether computational properties are intrinsic properties of physical systems (Dewhurst, 2018; Fresco, 2015).

For this reason, Schweizer (2019) argues that a computational explanation needs to specify more than just internal state transitions that account for abstractly described input-output patterns. Rather, a computational explanation should specify inputs and outputs relative to a particular system. For instance, a computational explanation of mental arithmetic, as performed by actual human beings, would have to specify in which sensory modality and in which format the input is presented to the subject (e.g., whether in terms of verbal questions), as well as how the output is produced (verbally, by way of written response, etc.; see Schweizer, 2019, p. 294). Internal state transitions posited by a computational explanation must then not only be mapped to internal states of the system in question, but these internal states must also be causally connected to the sensory input and behaviour figuring in the description of the input-output patterns that are to be explained. For instance, in a computational explanation of mental arithmetic, as performed by a human being, one would ideally specify neuronal states corresponding to the postulated computational steps, i.e., concrete internal state transitions, which are causally implicated in bringing about behaviour associated with the capacity to perform mental arithmetic (Schweizer, 2019, p. 303).

This account of computational explanation is applicable to behavioural explananda. It is not obvious how to apply it to the problem of consciousness.[20]

of a necessary epistemic connection is compatible with the presence of a necessary ontological connection. For instance, if pain experiences are identical with the firing of certain neural populations, then there is a necessary ontological connection, but it may still be reasonable to ask *why* this type of neural activity is identical with pain experiences. By contrast, if all computations performed by a conscious being can be performed by an unconscious computer, then there is not just an epistemic gap, but also an ontological gap between computational properties and consciousness.

[20]Schweizer makes this point: "The only non-abstract effects that instantiated formalisms are required to preserve are defined in terms of their input/output profiles, and thus *internal* experi-

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

Indeed, it seems it poses a dilemma for computational explanations of consciousness. One option is that the computational explanation targets *behaviour* that is associated with consciousness. But this would always beg the question whether the explanation really explains consciousness – even assuming there are types of behaviour that cannot be performed without consciousness (by human beings at least). On the face of it, such an explanation would only explain behaviour. The other option is that a computational explanation of consciousness targets internal processes in a conscious system (during conscious processing). In fact, given that the first option seemingly only explains behaviour, it seems this is what a computational explanation must account for. However, if one believes that implementing the right kind of computational system is insufficient for instantiating consciousness, then there will be non-conscious systems that still display all the computational properties instantiated by conscious beings. The properties that account for a system's being conscious must therefore also comprise non-computational properties. Can FEP bring anything more to the table?

FEP is not an account of physical computation. However, it can provide constraints that may complement existing accounts of computation. In other words, if the FEP applies to a given physical process, then that process has to possess certain properties. An account of physical computation specifies what it means to perform a computation (such as variational free-energy minimisation). Furthermore, it is likely that some computing devices will never be conscious, regardless of which computations they perform (e.g., a desktop PC). Such systems might simulate consciousness, but will never be conscious.

4.2 Distinguishing simulation from instantiation

A fundamental constraint provided by FEP is that it only applies to active systems that engage with their environment (but see the remarks about islands of awareness in section 3 above). A computer simulation of a system that minimises variational free energy may perform computations that are also instantiated by a conscious agent that interacts with its environment. But if these computations are realised by a piece of hardware that cannot be regarded as an agent (e.g., they cannot move or change external states), then they can at most lead to a simulation of consciousness (islands of awareness would constitute a limit case, i.e., they must at least have the potential to interact with the environment, if the causal connection to the environment via sensory and motor systems is restored; see footnote 26 below). Let us unpack this idea.

A system in NESS with a Markov blanket comprises internal, external, and blanket states. As shown in section 3, such a system can also be described as performing certain computations (inferences), by virtue of causal relations between

ences, qua actual events, are in principle omitted" (Schweizer, 2002, p. 144; see also 2019, footnote 11).

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

internal, external, and blanket states. This is different for a mere simulation.[21] To be sure, a simulating system can perform the same computations, and it will perform them because of causal relations between its parts. However, it will not perform them because of causal relations between its internal, external, and blanket states. The simulation may instantiate a *virtual* machine, which may have internal, external, and blanket states, such that its internal states encode a probability distribution over external states, given blanket states. But the internal states of the virtual machine will not be the internal states of the *physical* machine on which the virtual machine is running. In particular, consider a computer simulation on a *von Neumann computer* that stores the values of a system's states in its memory registers. There must be some (indirect) causal links between these memory states, but the interaction between them will always be mediated by the CPU. The causal flow in the virtual machine is thus different from the causal flow in the physical machine.

Put differently, not all virtual machines that perform approximate Bayesian inference by encoding a probability distribution over external states, given blanket states, are realised by physical machines with the same Markov blanket partition. If we keep this in mind, we can maintain a distinction between simulating and instantiating consciousness, but still retain the hypothesis that the right computational properties are sufficient for consciousness (Chalmers, 2011), if they are *instantiated by the right kind of system.*[22] Moreover, the "right kind of system" is not simply defined in terms of its physical properties (i.e., it is not about silicon- vs. carbon-based systems). Rather, the right kind of system is one in which the internal, external, and blanket states that figure in a computational description of the system are also physical states of the system, such that the system's internal dynamics can equivalently be described (a) with respect to the system's NESS density, or (b) with respect to the probability density (i.e., Bayesian beliefs) encoded by internal states.[23]

One might worry that this constraint still does not suffice to distinguish simulation from duplication. A computer simulating a system that performs variational inference over external states must have physical states encoding the probability distributions over external states that figure in those inferences. Define these physical states as internal. There will then be a unique Markov blanket for these internal states. Hence, the physical system's internal states will be numerically identical with the internal states that encode the probability distributions figuring in the computational system implemented by the physical system.

[21] Technically, in dynamical systems theory, this kind of "simulation" corresponds to a *skew-product* or *master-slave* system, in which the influence of external states on internal states (mediated vicariously by blanket states) was not complemented by an influence of internal states on external states. In other words, the circular causality – that renders the system autonomous – is precluded.

[22] In other words, our account is compatible with a slightly modified version of the thesis of computational sufficiency. However, it is also compatible with the possibility that computational explanations cannot explain all aspects of consciousness.

[23] A related argument is presented in Kiefer (2020).

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

This consideration overlooks that a system's Markov blanket must be specified with respect to the system's dynamics. These dynamics depend on the structure of the system. For instance, the dynamics of a human organism are quite different from the dynamics of a digital computer, even if both have internal states that encode the same probability distributions.[24] Moreover, a computer can simulate a vast number of different types of system (with different Markov blankets), but the computer's Markov blanket will not change when it simulates different systems (or at least it will not change in such a way as to match the Markov blanket of the simulated system). This means the computer's continued existence does not depend on the computations it performs in simulating another system. Conversely, digital computers with different causal topologies (and different Markov blankets) may be able to simulate the same virtual machine (Gamez, 2014, p. 180).

5 Computational explanations of consciousness and minimal unifying models

So far, we have argued that FEP can provide the basis for a computational explanation of consciousness. A computational explanation of consciousness must specify computations that are sufficient for at least some features of consciousness, under the assumption that the computations are performed by the right kind of system. If the computations can be described as inferences over external states, then FEP tells us what the right kind of system is: a physical system with internal states that encode the probability distributions figuring in the computational explanation, such that the physical system's internal dynamics can equivalently be described a) with respect to the system's NESS density, or b) with respect to the probability distribution encoded by internal states.

If correct, this only shows that FEP can provide a framework for computational explanations of consciousness (similarly to the proposal by Hohwy & Seth, 2020). It does not directly provide computational explanations. How should one expect to get to an explanation (or a theory) of consciousness, starting from a framework that itself is not a theory of consciousness (because it applies to basically everything)?

The account on offer here is that a computational explanation of consciousness should explain teleological functions of consciousness, i.e., cognitive capacities associated with consciousness (see also Cleeremans, 2005). Explanations of cognitive capacities cannot be provided by FEP itself, but by process theories such as active inference (which are theories of how minimising variational free energy is realised). The FEP can fulfil a unificatory role: using FEP as a framework, one can

[24] A special case of the possibility described in the paragraph above is a computer simulation of an island of awareness. In a mere simulation, blanket states are virtual, not just (temporally) disconnected from internal states; a real artificial island of awareness would have to be able to continue to function (i.e., to continue minimising variational free energy) if it were connected to *physical* sensors and actuators – not just virtual ones.

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

formulate and test hypotheses about necessary computational mechanisms that may underpin different cognitive capacities,[25] or different measures of consciousness.

This could also be used as a defense against the following charge: operationalisations of consciousness employed by consciousness researchers are so heterogeneous that they do not serve to unequivocally pick out a unique phenomenon (Irvine, 2013, 2017). Hence, it has been argued that the concept of consciousness should be eliminated from scientific discourse (Irvine, 2012; Irvine & Sprevak, 2020). As a unifying framework, FEP (and process theories such as active inference) could enable insights into what the phenomena picked out by different measures of consciousness have in common (see also Wiese, 2018). We shall unpack this idea in what follows.

Wiese (2020) argues that theoretical work in consciousness research should not focus on creating new theories of consciousness, but on creating a minimal unifying model, which embodies assumptions that existing theories have in common, and are thus more likely to be true than any entire theory. In doing this, only *necessary* conditions for consciousness are picked out (instead of sufficient conditions) – this is the main sense in which such a model will be minimal (see also Metzinger, 2020). This dovetails with the results presented here, in that CCCs should, from the point of view of FEP, be regarded only as necessary properties of consciousness, not as (minimally) sufficient properties (as suggested by Chalmers' NCC notion).

Minimising variational free energy could count as a minimal unifying model: variational free energy must be minimised by every self-organising system that persists, and hence also by any conscious system. Minimising variational free energy is thus a necessary condition, trivially presupposed by any theory of consciousness that does not explicitly reject FEP. Unfortunately, this necessary condition is not really informative, because it is trivial.

However, recall that slightly more informative constraints directly follow from FEP. CCCs should be specified in terms of probabilities encoded by internal states, not in terms of probabilities of internal states. Conscious (and unconscious) systems must minimise the difference between accuracy and complexity – and systems with (transiently absent) connection to sensory inputs and motor outputs will tend to minimise complexity.

Apart from these implications, more informative minimal unifying models can be developed under the FEP by taking results from consciousness research into account. For instance, a promising approach is to focus on *cognitive capacities* associated with consciousness, such as trace conditioning, complex types of learning (Birch, 2020; Ginsburg & Jablonka, 2019; Kanai et al., 2019), or perhaps working memory (Reggia et al., 2019). It is unlikely that any particular cognitive capacity by itself is sufficient for consciousness (Birch, 2020; Wiese, 2020). Instead, the

[25] A related idea, put forward by Ginsburg & Jablonka (2019), is that there is a single learning capacitiy that underpins various characteristic features of consciousness.

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

hypothesis is that consciousness enables a number of different cognitive abilities. Furthermore, if consciousness is a natural kind (Shea, 2012a; Shea & Bayne, 2010), then it is to be expected that there is a single underlying mechanism for these cognitive abilities. Whether consciousness is indeed a natural kind is an empirical question: finding a positive correlation between different cognitive capacities associated with consciousness would support the assumption.

FEP-based models can contribute to this approach from a theoretical perspective. For instance, active inference models (Friston et al., 2017) could establish that minimising (expected) variational free energy in deep models is sufficient for a broad range of cognitive capacities associated with consciousness (at least if certain further, to-be-specified conditions are fulfilled; see Wiese, 2018). More specifically, assume there is a cluster of consciousness-linked cognitive capacities $c_1, c_2, c_3, \ldots$ Strong support for the natural-kind hypothesis would be provided if a minimal model of c_1 were *ipso facto* a minimal model of c_2 (and of c_3, ...). This is unlikely, but it may turn out that minimal models of c_1, c_2, c_3, ..., respectively, share at least some core structural features (Wiese, 2020) that play a central role in accounting for these different cognitive abilities. A model with these core structural features would then point to a common mechanism underlying these capacities. The computational description afforded by such a unifying model would provide an essential part of a computational explanation of consciousness.

The explanatory "heavy lifting" would not be done by FEP itself, but by specific models of the respective cognitive capacities (Schlicht & Dołega, 2021). However, finding common core structural features of these models (Fink et al., 2021; Song, 2021) will be facilitated by (or will even be conditional on) expressing the models within a single formal framework, as provided by FEP (and active inference).

6 Conclusion

How can research on neural correlates of consciousness (NCCs) contribute to an explanation of consciousness, and what role could the free energy principle (FEP) possibly play in this endeavour? We have suggested that research on neural correlates should be complemented by research on computational correlates of consciousness (CCCs), in order to yield a computational explanation of consciousness. Moreover, according to FEP, NCCs must be defined in terms of neural *dynamics*, not neural states, and CCCs should be defined in terms of probabilities *encoded by* neural states.

Consequently, a computational explanation of consciousness requires more than a formal description of neural dynamics. It must also specify computations implemented by these dynamics. A key contribution by FEP is that it provides additional constraints on what it means to *be*, as opposed to merely *simulate* a conscious system. Put differently, the possibility of a computational explanation of consciousness is compatible with the assumption that some systems instantiate all computational properties associated with consciousness, without being conscious.

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

There are many open problems and questions we have barely mentioned, or even outright ignored, such as the hard problem (Chalmers, 1995), the meta problem (Chalmers, 2018), or the computational explanatory gap (Reggia et al., 2017; Reggia et al., 2014). A computational explanation of consciousness can only be achieved if these problems are addressed as well. Furthermore, the possibility of a computational explanation of consciousness also raises the possibility of machine consciousness (Reggia, 2013), as well as ethical concerns that should be addressed (Agarwal & Edelman, 2020; Metzinger, 2013, 2017, 2021). However, our goal has been more modest. All we hope to have shown is that a computational explanation of consciousness is possible, and that the free energy principle can make a significant contribution to this project.

Acknowledgments
We are extremely grateful to all attendants of a virtual theoretical neurobiology meeting at the Wellcome Centre for Human Neuroimaging, at which an early version of this paper was discussed. WW is grateful to Joe Dewhurst and Mark Sprevak for correspondence on issues related to ideas presented in this paper, as well as to audiences at the Tokyo Consciousness Club, the Active Inference Lab, the Mathematical Consciousness Science Seminar, and Karen Konkoly's reading group. We also thank two anonymous reviewers and Ying-Tung Lin and Sascha Fink for their critical and constructive comments. This research was supported by a Wellcome Trust Principal Research Fellowship (KF; Ref: 088130/Z/09/Z).

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

Ao, P. (2008). Emerging of stochastic dynamical equalities and steady state thermodynamics. *Communications in Theoretical Physics*, *49*(5), 1073–1090. https://doi.org/10.1088/0253-6102/49/5/01

Aru, J., Bachmann, T., Singer, W., & Melloni, L. (2012). Distilling the neural correlates of consciousness. *Neuroscience & Biobehavioral Reviews*, *36*(2), 737–746. https://doi.org/10.1016/j.neubiorev.2011.12.003

Atkinson, A. P., Thomas, M. S. C., & Cleeremans, A. (2000). Consciousness: Mapping the theoretical landscape. *Trends in Cognitive Sciences*, *4*(10), 372–382. https://doi.org/10.1016/S1364-6613(00)01533-3

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. (2020). The search for invertebrate consciousness. *Noûs*, 1–21. https://doi.org/10.1111/nous.12351

Block, N. (2005). Two neural correlates of consciousness. *Trends in Cognitive Sciences*, *9*(2), 46-52. https://doi.org/10.1016/j.tics.2004.12.006

Bruineberg, J., Kiverstein, J., & Rietveld, E. (2018). The anticipating brain is not a scientist: The free-energy principle from an ecological-enactive perspective. *Synthese*, *195*, 2417–2444. https://doi.org/10.1007/s11229-016-1239-1

Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, *2*(3), 200–219. https://www.ingentaconnect.com/content/imp/jcs/1995/00000002/00000003/653

Chalmers, D. J. (2000). What is a neural correlate of consciousness? In T. Metzinger (Ed.), *Neural correlates of consciousness: Empirical and conceptual questions* (pp. 17–40). MIT Press.

Chalmers, D. J. (2011). A computational foundation for the study of cognition. *Journal of Cognitive Science*, *12*, 323–357.

Chalmers, D. J. (2018). The meta-problem of consciousness. *Journal of Consciousness Studies*, *25*(9-10), 6–61. https://www.ingentaconnect.com/content/imp/jcs/2018/00000025/f0020009/art00001

Churchland, P. (2005). Chimerical colors: Some phenomenological predictions from cognitive neuroscience. *Philosophical Psychology*, *18*(5), 527–560. https://doi.org/10.1080/09515080500264115

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

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., Achoui, D., Beauny, A., Keuninckx, L., Martin, J.-R., Muñoz-Moldes, S., Vuillaume, L., & Heering, A. de. (2020). Learning to be conscious. *Trends in Cognitive Sciences*, *24*(2), 112–123. https://doi.org/10.1016/j.tics.2019.11.011

Colombo, M. A., Napolitani, M., Boly, M., Gosseries, O., Casarotto, S., Rosanova, M., Brichant, J. F., Boveroux, P., Rex, S., Laureys, S., Massimini, M., Chieregato, A., & Sarasso, S. (2019). The spectral exponent of the resting EEG indexes the presence of consciousness during unresponsiveness induced by propofol, xenon, and ketamine. *Neuroimage*, *189*, 631–644. https://doi.org/10.1016/j.neuroimage.2019.01.024

Costa, L. D., Parr, T., Sajid, N., Veselic, S., Neacsu, V., & Friston, K. (2020). *Active inference on discrete state-spaces: A synthesis*. https://arxiv.org/abs/2001.07203

Costa, L. D., Sajid, N., Parr, T., Friston, K., & Smith, R. (2020). *The relationship between dynamic programming and active inference: The discrete, finite-horizon case*. https://arxiv.org/abs/2009.08111

Crick, F., & Koch, C. (1990). Towards a neurobiological theory of consciousness. *Seminars in the Neurosciences*, *2*, 263–275. https://doi.org/https://resolver.caltech.edu/CaltechAUTHORS:20130816-103136937

De Ribaupierre, S., & Delalande, O. (2008). Hemispherotomy and other disconnective techniques. *Neurosurgical Focus*, *25*(3), E14. https://doi.org/10.3171/FOC/2008/25/9/E14

Demertzi, A., Tagliazucchi, E., Dehaene, S., Deco, G., Barttfeld, P., Raimondo, F., Martial, C., Fernández-Espejo, D., Rohaut, B., Voss, H., & others. (2019). Human consciousness is supported by dynamic complex patterns of brain signal coordination. *Science Advances*, *5*(2), eaat7603. https://doi.org/10.1126/sciadv.aat7603

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

Downey, A. (2018). Predictive processing and the representation wars: A victory for the eliminativist (via fictionalism). *Synthese*, *195*(12), 5115–5139. https://doi.org/10.1007/s11229-017-1442-8

Fekete, T., & Edelman, S. (2011). Towards a computational theory of experience. *Consciousness and Cognition*, *20*(3), 807–827. https://doi.org/10.1016/j.concog.2011.02.010

Fink, S. B. (2016). A deeper look at the "neural correlate of consciousness." *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01044

Fink, S. B., Lyre, H., & Kob, L. (2021). A structural constraint on neural correlates of consciousness. *Philosophy and the Mind Sciences*, *2*, 7. https://doi.org/10.33735/phimisci.2021.79

Fresco, N. (2015). Objective computation versus subjective computation. *Erkenntnis*, *80*(5), 1031–1053. https://doi.org/10.1007/s10670-014-9696-8

Friston, J. K. (2019a). Complexity and computation in the brain. The knowns and the known unknowns. In W. Singer, T. J. Sejnowski, & P. Rakic (Eds.), *The neocortex* (pp. 269–291). MIT Press.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787

Friston, K. (2013). Life as we know it. *Journal of The Royal Society Interface*, *10*(86). https://doi.org/10.1098/rsif.2013.0475

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

Friston, K., Breakspear, M., & Deco, G. (2012). Perception and self-organized instability. *Frontiers in Computational Neuroscience*, *6*, 44. https://doi.org/10.3389/fncom.2012.00044

Friston, K., Costa, L. D., & Parr, T. (2020). *Some interesting observations on the free energy principle*. https://arxiv.org/abs/2002.04501

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: A process theory. *Neural Computation*, *29*(1), 1–49. https://doi.org/10.1162/NECO_a_00912

Friston, K. J., Fagerholm, E. D., Zarghami, T. S., Parr, T., Hipólito, I., Magrou, L., & Razi, A. (2021). Parcels and particles: Markov blankets in the brain. In *Network Neuroscience* (No. 1; Vol. 5, pp. 211–251). https://doi.org/10.1162/netn_a_00175

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., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1211–1221. https://doi.org/10.1098/rstb.2008.0300

Gamez, D. (2014). Can we prove that there are computational correlates of consciousness in the brain? *Journal of Cognitive Science*, *15*(2), 149–186. https://doi.org/10.17791/jcs.2014.15.2.149

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

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

He, B. J., Zempel, J. M., Snyder, A. Z., & Raichle, M. E. (2010). The temporal structures and functional significance of scale-free brain activity. *Neuron*, *66*(3), 353–369. https://doi.org/10.1016/j.neuron.2010.04.020

Hobson, J. A., & Friston, K. J. (2012). Waking and dreaming consciousness: Neurobiological and functional considerations. *Progress in Neurobiology*, *98*(1), 82–98. https://doi.org/10.1016/j.pneurobio.2012.05.003

Hobson, J. A., & Friston, K. J. (2014). Consciousness, dreams, and inference: The Cartesian theatre revisited. *Journal of Consciousness Studies*, *21*(1-2), 6–32. https://www.ingentaconnect.com/content/imp/jcs/2014/00000021/F0020001/art00001

Hohwy, J. (2009). The neural correlates of consciousness: New experimental approaches needed? *Consciousness and Cognition*, *18*(2), 428–438. https://doi.org/10.1016/j.concog.2009.02.006

Hohwy, J. (2016). The self-evidencing brain. *Noûs*, *50*(2), 259–285. https://doi.org/10.1111/nous.12062

Hohwy, J. (2020). Self-supervision, normativity and the free energy principle. *Synthese*, 1–25. https://doi.org/10.1007/s11229-020-02622-2

Hohwy, J., & Seth, A. (2020). Predictive processing as a systematic basis for identifying the neural correlates of consciousness. *Philosophy and the Mind Sciences*, *1*(II), 3. https://doi.org/10.33735/phimisci.2020.II.64

Huang, Z., Zhang, J., Wu, J., Mashour, G. A., & Hudetz, A. G. (2020). Temporal circuit of macroscale dynamic brain activity supports human consciousness. *Science Advances*, *6*(11). https://doi.org/10.1126/sciadv.aaz0087

Hutter, M. (2000). A theory of universal artificial intelligence based on algorithmic complexity. *CoRR*, *cs.AI/0004001*. https://arxiv.org/abs/cs/0004001

Hutter, M. (2005). *Universal artificial intelligence: Sequential decisions based on algorithmic probability*. Springer Science & Business Media.

Hutto, D. D. (2018). Getting into predictive processing's great guessing game: Bootstrap heaven or hell? *Synthese*, *195*(6), 2445–2458. https://doi.org/10.1007/s11229-017-1385-0

Irvine, E. (2012). *Consciousness as a scientific concept: A philosophy of science perspective*. Springer Science & Business Media.

Irvine, E. (2013). Measures of consciousness. *Philosophy Compass*, *8*(3), 285–297. https://doi.org/10.1111/phc3.12016

Irvine, E. (2017). Explaining what? *Topoi*, *36*(1), 95–106. https://doi.org/10.1007/s11245-014-9273-4

Irvine, E., & Sprevak, M. (2020). Eliminativism about consciousness. In U. Kriegel (Ed.), *The Oxford handbook of the philosophy of consciousness* (pp. 348–370). Oxford University Press.

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

Kiefer, A. B. (2020). Psychophysical identity and free energy. *Journal of The Royal Society Interface*, *17*(169), 20200370. https://doi.org/10.1098/rsif.2020.0370

Kirchhoff, M. D., & Robertson, I. (2018). Enactivism and predictive processing: A non-representational view. *Philosophical Explorations*, *21*(2), 264–281. https://doi.org/10.1080/13869795.2018.1477983

Koch, C., Massimini, M., Boly, M., & Tononi, G. (2016). Neural correlates of consciousness: Progress and problems. *Nature Reviews Neuroscience*, *17*(5), 307–321. https://doi.org/10.1038/nrn.2016.22

Lancaster, M. A., Renner, M., Martin, C.-A., Wenzel, D., Bicknell, L. S., Hurles, M. E., Homfray, T., Penninger, J. M., Jackson, A. P., & Knoblich, J. A. (2013). Cerebral organoids model human brain development and microcephaly. *Nature*, *501*(7467), 373–379. https://doi.org/10.1038/nature12517

Lempel, A., & Ziv, J. (1976). On the complexity of finite sequences. *IEEE Transactions on Information Theory*, *22*(1), 75–81.

Levine, J. (1983). Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly*, *64*(9), 354–361.

Li, D., & Mashour, G. A. (2019). Cortical dynamics during psychedelic and anesthetized states induced by ketamine. *NeuroImage*, *196*, 32–40. https://doi.org/https://doi.org/10.1016/j.neuroimage.2019.03.076

Luppi, A. I., Craig, M. M., Pappas, I., Finoia, P., Williams, G. B., Allanson, J., Pickard, J. D., Owen, A. M., Naci, L., Menon, D. K., & Stamatakis, E. A. (2019). Consciousness-specific dynamic interactions of brain integration and functional diversity. *Nature Communications*, *10*(1), 4616. https://doi.org/10.1038/s41467-019-12658-9

MacKay, D. J. (1995). Free energy minimisation algorithm for decoding and cryptanalysis. *Electronics Letters*, *31*(6), 446–447. https://doi.org/10.1049/el:19950331

Mashour, G. A. (2018). The controversial correlates of consciousness. *Science*, *360*(6388), 493–494. https://doi.org/10.1126/science.aat5616

Mashour, G. A., Roelfsema, P., Changeux, J. P., & Dehaene, S. (2020). Conscious processing and the global neuronal workspace hypothesis. *Neuron*, *105*(5), 776–798. https://doi.org/10.1016/j.neuron.2020.01.026

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

Mathis, D., & Mozer, M. C. (1996). Conscious and unconscious perception: A computational theory. In G. Cottrell (Ed.), *Proceedings of the eighteenth annual conference of the cognitive science society* (pp. 324–328). Erlbaum.

Metzinger, T. (2013). Two principles for robot ethics. In E. Hilgendorf & J.-P. Günther (Eds.), *Robotik und Gesetzgebung. Beiträge der Tagung vom 7. bis 9. Mai 2012 in Bielefeld.* (pp. 263–302). Nomos.

Metzinger, T. (2017). Suffering, the cognitive scotoma. In K. Almqvist & A. Haag (Eds.), *The return of consciousness. A new science on old questions* (pp. 237–262). Axel and Margaret Axson Johnson Foundation.

Metzinger, T. (2020). Minimal phenomenal experience: Meditation, tonic alertness, and the phenomenology of "pure" consciousness. *Philosophy and the Mind Sciences*, *1*(I), 7. https://doi.org/10.33735/phimisci.2020.I.46

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

Milkowski, M. (2013). *Explaining the computational mind.* MIT Press.

Miller, S. M. (2007). On the correlation/constitution distinction problem (and other hard problems) in the scientific study of consciousness. *Acta Neuropsychiatrica*, *19*(3), 159–176. https://doi.org/10.1111/j.1601-5215.2007.00207.x

Noë, A., & Thompson, E. (2004). Are there neural correlates of consciousness? *Journal of Consciousness Studies*, *11*(1), 3–28. https://www.ingentaconnect.com/content/imp/jcs/2004/00000011/00000001/1400

Northoff, G., Tsuchiya, N., & Saigo, H. (2019). Mathematics and the brain: A category theoretical approach to go beyond the neural correlates of consciousness. *Entropy*, *21*(12), 1234. https://doi.org/10.3390/e21121234

Palacios, E. R., Razi, A., Parr, T., Kirchhoff, M., & Friston, K. (2017). Biological self-organisation and Markov blankets. *bioRxiv*. https://doi.org/10.1101/227181

Parr, T., Da Costa, L., & Friston, K. (2020). Markov blankets, information geometry and stochastic thermodynamics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *378*(2164), 20190159. https://doi.org/10.1098/rsta.2019.0159

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference.* Morgan Kaufmann Publishers.

Pellet, J. P., & Elisseeff, A. (2008). Using Markov blankets for causal structure learning. *Journal of Machine Learning Research*, *9*, 1295–1342. http://jmlr.org/papers/v9/pellet08a.html

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

Ramstead, M. J. D., Friston, K. J., & Hipólito, I. (2020). Is the free-energy principle a formal theory of semantics? From variational density dynamics to neural and phenotypic representations. *Entropy*, *22*(8), 889. https://doi.org/10.3390/e22080889

Reggia, J. A. (2013). The rise of machine consciousness: Studying consciousness with computational models. *Neural Networks*, *44*, 112–131. https://doi.org/10.1016/j.neunet.2013.03.011

Reggia, J. A., Katz, G., & Davis, G. P. (2019). Modeling working memory to identify computational correlates of consciousness. *Open Philosophy*, *2*, 252–269. https://doi.org/10.1515/opphil-2019-0022

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

Reggia, J. A., Monner, D., & Sylvester, J. (2014). The computational explanatory gap. *Journal of Consciousness Studies*, *21*(9-10), 153–178. https://www.ingentaconnect.com/content/imp/jcs/2014/00000021/F0020009/art00007

Reggia, J., Huang, D.-W., & Katz, G. (2017). Exploring the computational explanatory gap. *Philosophies*, *2*(1), 5. https://doi.org/10.3390/philosophies2010005

Rohaut, B., Raimondo, F., Galanaud, D., Valente, M., Sitt, J. D., & Naccache, L. (2017). Probing consciousness in a sensory-disconnected paralyzed patient. *Brain Injury*, *31*(10), 1398–1403. https://doi.org/10.1080/02699052.2017.1327673

Rosas, F. E., Mediano, P. A. M., Biehl, M., Chandaria, S., & Polani, D. (2020). *Causal blankets: Theory and algorithmic framework.* https://arxiv.org/abs/2008.12568

Ruffini, G. (2017). An algorithmic information theory of consciousness. *Neuroscience of Consciousness*, *3*(1). https://doi.org/10.1093/nc/nix019

Sarasso, S., Boly, M., Napolitani, M., Gosseries, O., Charland-Verville, V., Casarotto, S., Rosanova, M., Casali, A. G., Brichant, J.-F., Boveroux, P., & others. (2015). Consciousness and complexity during unresponsiveness induced by propofol, xenon, and ketamine. *Current Biology*, *25*(23), 3099–3105. https://doi.org/10.1016/j.cub.2015.10.014

Sarasso, S., Rosanova, M., Casali, A. G., Casarotto, S., Fecchio, M., Boly, M., Gosseries, O., Tononi, G., Laureys, S., & Massimini, M. (2014). Quantifying cortical EEG responses to TMS in (un)consciousness. *Clinical EEG and Neuroscience*, *45*(1), 40–49. https://doi.org/10.1177/1550059413513723

Schartner, M. M., Pigorini, A., Gibbs, S. A., Arnulfo, G., Sarasso, S., Barnett, L., Nobili, L., Massimini, M., Seth, A. K., & Barrett, A. B. (2017). Global and local complexity of intracranial EEG decreases during NREM sleep. *Neuroscience of Consciousness*, *2017*(1), niw022. https://doi.org/10.1093/nc/niw022

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

Schartner, M., Seth, A., Noirhomme, Q., Boly, M., Bruno, M.-A., Laureys, S., & Barrett, A. (2015). Complexity of multi-dimensional spontaneous EEG decreases during propofol induced general anaesthesia. *PloS One*, *10*(8), e0133532. https://doi.org/10.1371/journal.pone.0133532

Schlicht, T., & Dołega, K. (2021). You can't always get what you want: Predictive processing and consciousness. *Philosophy and the Mind Sciences*, *2*, 8. https://doi.org/10.33735/phimisci.2021.80

Schweizer, P. (2002). Consciousness and computation. *Minds and Machines*, *12*(1), 143–144. https://doi.org/10.1023/A:1013741414324

Schweizer, P. (2019). Triviality arguments reconsidered. *Minds and Machines*, *29*, 287–308. https://doi.org/10.1007/s11023-019-09501-x

Seifert, U. (2012). Stochastic thermodynamics, fluctuation theorems and molecular machines. *Reports on Progress in Physics*, *75*(12), 126001. https://doi.org/10.1088/0034-4885/75/12/126001

Sekimoto, K. (1998). Langevin equation and thermodynamics. *Progress of Theoretical Physics Supplement*, *130*, 17–27. https://doi.org/10.1143/PTPS.130.17

Seth, A. K. (2009). Explanatory correlates of consciousness: Theoretical and computational challenges. *Cognitive Computation*, *1*(1), 50–63. https://doi.org/10.1007/s12559-009-9007-x

Seth, A. K., & Edelman, G. M. (2009). Consciousness and complexity. In R. A. Meyers (Ed.), *Springer encyclopedia of complexity and systems science* (pp. 1424–1443). Springer.

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

Shea, N. (2012b). Reward prediction error signals are meta-representational. *Noûs*, *48*(2), 314–341. https://doi.org/10.1111/j.1468-0068.2012.00863.x

Shea, N., & Bayne, T. (2010). The vegetative state and the science of consciousness. *The British Journal for the Philosophy of Science*, *61*(3), 459–484. https://doi.org/10.1093/bjps/axp046

Sitt, J. D., King, J. R., El Karoui, I., Rohaut, B., Faugeras, F., Gramfort, A., Cohen, L., Sigman, M., Dehaene, S., & Naccache, L. (2014). Large scale screening of neural signatures of consciousness in patients in a vegetative or minimally conscious state. *Brain*, *137*(Pt 8), 2258–2270. https://doi.org/10.1093/brain/awu141

Smith, B. C. (2019). *The promise of artificial intelligence: Reckoning and judgment.* MIT Press.

Song, C. (2021). Structural basis of consciousness. *Philosophy and the Mind Sciences*, *2*, 6. https://doi.org/10.33735/phimisci.2021.75

Sprevak, M. (2019). Triviality arguments about computational implementation. In M. Sprevak & M. Colombo (Eds.), *Routledge handbook of the computational mind* (pp. 175–191). Routledge.

Sprevak, M. (2020). Two kinds of information processing in cognition. *Review of Philosophy and Psychology*, *11*(3), 591–611. https://doi.org/10.1007/s13164-019-00438-9

Stevner, A., Vidaurre, D., Cabral, J., Rapuano, K., Nielsen, S. F. V., Tagliazucchi, E., Laufs, H., Vuust, P., Deco, G., Woolrich, M., & others. (2019). Discovery of key whole-brain transitions and dynamics during human wakefulness and non-REM sleep. *Nature Communications*, *10*(1), 1–14. https://doi.org/10.1038/s41467-019-08934-3

Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: From consciousness to its physical substrate. *Nature Reviews Neuroscience*, *17*(7), 450–461. https://doi.org/10.1038/nrn.2016.44

Tononi, G., & Koch, C. (2008). The neural correlates of consciousness: An update. *Annals of the New York Academy of Sciences*, *1124*, 239–261. https://doi.org/10.1196/annals.1440.004

Tsuchiya, N., Wilke, M., Frässle, S., & Lamme, V. A. F. (2015). No-report paradigms: Extracting the true neural correlates of consciousness. *Trends in Cognitive Sciences*, *19*(12), 757–770. https://doi.org/https://doi.org/10.1016/j.tics.2015.10.002

Van Es, T. (2019). Minimizing prediction errors in predictive processing: From inconsistency to non-representationalism. *Phenomenology and the Cognitive Sciences*, 1–21. https://doi.org/10.1007/s11097-019-09649-y

Vrselja, Z., Daniele, S. G., Silbereis, J., Talpo, F., Morozov, Y. M., Sousa, A. M., Tanaka, B. S., Skarica, M., Pletikos, M., Kaur, N., Zhuang, Z. W., Liu, Z., Alkawadri, R., Sinusas, A. J., Latham, S. R., Waxman, S. G., & Sestan, N. (2019). Restoration of brain circulation and cellular functions hours post-mortem. *Nature*, *568*(7752), 336. https://doi.org/10.1038/s41586-019-1099-1

Wallace, C. S., & Dowe, D. L. (1999). Minimum message length and Kolmogorov complexity. *The Computer Journal*, *42*(4), 270–283. https://doi.org/10.1093/comjnl/42.4.270

Wiese, W. (2020). The science of consciousness does not need another theory, it needs a minimal unifying model. *Neuroscience of Consciousness*, *2020*(1). https://doi.org/10.1093/nc/niaa013

Wiese, W. (2018). Toward a mature science of consciousness. *Frontiers in Psychology*, *9*, 693. https://doi.org/10.3389/fpsyg.2018.00693

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., & Friston, K. J. (2021). Examining the continuity between life and mind: Is there a continuity between autopoietic intentionality and representationality? *Philosophies*, *6*(11), 18. https://doi.org/10.3390/philosophies6010018