




Active Inference as a Computational Framework for Consciousness

Martina G. Vilas¹  · Ryszard Auksztulewicz¹ · Lucia Melloni^{1,2}

Accepted: 21 July 2021/Published online: 10 August 2021
© The Author(s) 2021

Abstract

Recently, the mechanistic framework of active inference has been put forward as a principled foundation to develop an overarching theory of consciousness which would help address conceptual disparities in the field (Wiese 2018; Hohwy and Seth 2020). For that promise to bear out, we argue that current proposals resting on the active inference scheme need refinement to become a process theory of consciousness. One way of improving a theory in mechanistic terms is to use formalisms such as computational models that implement, attune and validate the conceptual notions put forward. Here, we examine how computational modelling approaches have been used to refine the theoretical proposals linking active inference and consciousness, with a focus on the extent and success to which they have been developed to accommodate different facets of consciousness and experimental paradigms, as well as how simulations and empirical data have been used to test and improve these computational models. While current attempts using this approach have shown promising results, we argue they remain preliminary in nature. To refine their predictive and structural validity, testing those models against empirical data is needed i.e., new and unobserved neural data. A remaining challenge for active inference to become a theory of consciousness is to generalize the model to accommodate the broad range of consciousness explananda; and in particular to account for the phenomenological aspects of experience. Notwithstanding these gaps, this approach has proven to be a valuable avenue for theory advancement and holds great potential for future research.

Keywords Active inference · Consciousness · Computational model · Mechanistic model

✉ Martina G. Vilas
martina.vilas@ae.mpg.de

✉ Lucia Melloni
lucia.melloni@ae.mpg.de

¹ Department of Neuroscience, Max Planck Institute for Empirical Aesthetics, Frankfurt/M, Germany

² Department of Neurology, NYU Grossman School of Medicine, New York, USA

1 Theories of Consciousness and the Multiple Explanandum Problem

The nature of consciousness remains one of the most puzzling and least understood phenomena. In the early 90s, Crick and Koch proposed a research program focused on the neural correlates of consciousness (NCC), as inroads into understanding how the physical (brain) give rise to the phenomenal qualities (consciousness), arguing that ‘the problem of consciousness can, in the long run, be solved only by explanations at the neural level’ (Crick and Koch 1990). The NCC was defined as the minimum neuronal mechanisms jointly necessary for enabling any one specific conscious experience (Chalmers 2000). Since that seminal contribution, the NCC program has flourished. Many empirical investigations have been conducted aimed at unravelling the neural mechanism(s) that underpin consciousness, which have led to the development of a number of theories in recent years (for a recent review see Doerig et al. 2021). Yet, little agreement exists on their implications and the interpretation of their findings. This issue may partly be due to a lack of a unified explananda i.e., what is that the science of consciousness ought to explain, and of consensus on how consciousness should be operationalized and measured. We refer to this as the “explanandum problem”. Consciousness has been described as a “bundle of features” (Wiese 2018), each of which is tackled by different research programs, which themselves internally debate on the taxonomy and experimental paradigms to be used. Let us start by briefly reviewing some ways in which the study of consciousness has been approached.

Consciousness has been studied across two dimensions: Arousal or wakefulness (i.e., state of consciousness) and awareness (i.e., content of consciousness) (Laureys 2005).

Research on the *states of consciousness*, studies consciousness as a temporally extended state where having a subjective experience is possible, as opposed to other states where experience as a whole is absent (e.g. under anesthesia) (Bayne 2007). Whether those states are better thought of as resembling different “levels” in a continuum or “regions” in a multidimensional space remains an open question (Bayne et al. 2016). Researchers who investigate the neural correlates of conscious states usually compare how brain activity differs between states, with or without sensory stimulation (Bayne 2007). A few examples are studies contrasting patterns of brain activity, measured with functional Magnetic Resonance Imaging (fMRI), Magneto/Electroencephalography (M/EEG), Electrocorticography (ECoG), between wakefulness vs. sleep (e.g. Horowitz et al. 2009), between dreaming vs. dreamless sleep (e.g. Siclari et al. 2017), or between wakefulness vs. anaesthesia (Alkire et al. 2008) (for a recent review see Koch et al. 2016).

Research focusing on the *content of consciousness* investigates the neural patterns associated to a particular phenomenal content, such as a specific object, face, colour or sound. A useful distinction was proposed by Ned Block (1990, 1992, 1995) between phenomenal consciousness and access consciousness, who has further argued for the existence of at least two distinct NCC associated with those constructs, also differing in their experimental approaches. We review those in turn.

The phenomenal content of consciousness refers to the subjective qualities of conscious experience i.e., what differs between experiences of red versus green (Block 1990; Block 1992; Block 1995; Metzinger 2000). Studying phenomenology

involves dealing with the “hard problem” of consciousness, which refers to grappling with what it means to “feel” something. Why does seeing something *feel* different from hearing it? (Chalmers 1995). Understanding the subjective qualities of consciousness, as opposed to understanding the mechanisms that enable the report of such subjective experiences (conscious access), requires addressing at least two questions. One question refers to identifying the phenomenological properties that are essential and invariantly present in *any experience*. That is, what makes any experience an experience e.g., intrinsic, unified, unique, integrated and definite according to some theories (Tononi et al. 2016). Another question refers to what are the specific properties of a *particular experience* – what makes a given experience that experience e.g., why we see the color red when we look at an apple. To address these questions, phenomenology research collects first-person data using introspection, a form of qualitative research focusing on the study of an individual’s lived experiences within the world. Those methods, while very insightful at unravelling those subjective experiences, do not come without limitations. Perhaps the biggest challenge is that not all experiences are equally accessible to introspection (Nisbett and Wilson 1977). In addition, the mere exercise of introspecting can modify the content of those primal experiences. These limitations however do not disqualify the methodology per se. It merely reflects the fact that sources of evidence might be limited, a concept that is extensible to every scientific methodology. Yet, they remain essential tools for understanding how experiences present to us, and how we attribute meaning to them. The research tradition of neurophenomenology put forward by the late Francisco Varela (1996) and others, for instance in the framework of Integrated Information Theory (Haun and Tononi 2019), have aimed at addressing and better characterizing the structure of these phenomenal qualities and the neural mechanism behind them. A recent example for the latter approach concerns studies investigating the experience of space for which initial ideas have been advanced (Haun and Tononi 2019) and supported by empirical evidence (Song et al. 2017). In the tradition of neurophenomenology (Varela 1996), rigorous first-person data have been used in conjunction with neurophysiological data to shed light on the large-scale dynamics of consciousness (Lutz et al. 2002; for a recent review see Berkovich-Ohana et al. 2020).

Conscious access, on the other hand, refers to processes that enable specific information to be subjectively reported and made available for use by higher-level cognitive processes (e.g. reasoning, planning, decision-making, voluntary direction of attention, action control, etc.) (Block 2005). This information does not need to come from the senses, it can also be internal thoughts about our own perception (i.e. meta-cognition), or awareness about our own conscious experience (i.e. meta-awareness). Paradigms typically investigating access consciousness require participants to report about their internal experiences. For instance, in binocular rivalry experiments, participants are asked to report (e.g., via button press) their current percept, as perception alternates between competing stimuli that are concurrently presented to each eye. Neural representations between those trials in which a stimuli has been consciously perceived are contrasted against those in which the very same stimuli was not perceived. Another strategy consists of presenting stimulus that vary along a specific dimension known to affect perceptual detection (e.g., duration, contrast, spatial frequency) while collecting perceptual reports. Brain activity is then contrasted between those trials in which participants reported to

consciously perceive the stimuli with those in which they reported not perceiving the stimuli (for an overview of methods used to render stimuli invisible see Kim and Blake 2005).

Perhaps as a consequence of the heterogeneity in what the field of consciousness takes as its target of explanation (Signorelli et al. 2021), many theories of the neural correlates of consciousness have been put forward, each emphasizing one (or several) of the outlined consciousness explanandum (i.e. its states, contents, phenomenological nature, reportability, etc.) and selecting their preferred methodology accordingly. Reacting to this state of affairs, some argue that consciousness should not be partitioned, and instead should be studied as a unified phenomenon with a single underlying mechanism. For example, Bachmann and Hudetz (2014) argue that consciousness results from the interaction between the mechanisms that represent the contents of consciousness, and those that enable and modulate different states of conscious experience. Both mechanisms are necessary and co-dependent: contents cannot reach consciousness if there is no conscious state, and there is no conscious state without a content being represented. Thus, in their view, elucidating the integration of these two aspects is decisive for uncovering a sufficient NCC. In contrast, others have called for a strict separation between the concepts (Block 2005). Although concrete conceptual attempts have been made to bring different consciousness theories together, or at least to integrate their central claims (e.g. Wiese 2020), a common research agenda has not been agreed upon yet and progress in the field appears in many ways to have stagnated.

Against this background, there has been a push to integrate the study of the diverse explananda of consciousness and their respective experimental paradigms by interpreting the empirical findings using the explanatory constructs of a general-purpose modelling framework of brain function: active inference (Wiese 2018; Hohwy and Seth 2020). Some work has already been advanced to clarify which aspects of this framework might be particularly suitable for explaining common phenomena in consciousness research (Hohwy and Seth 2020).

Here we aim to gain some clarity on the issue of whether and how the computational modelling framework afforded by active inference can be exploited to its full potential to produce, refine and unify an explanatory theory of consciousness (when broadly defined). In reviewing the promise of active inference as a general framework for consciousness, we will for this piece set aside the important question of what the explananda of the research program on consciousness ought to be i.e., access consciousness or phenomenology. We will instead borrow inspiration from the systematic approach taken by Hohwy and Seth (2020) in which predictive processing is evaluated by its ability to provide a framework for explaining common challenges in consciousness research.

2 Active Inference as a Modelling Framework for Building a Process Theory

The Free Energy Principle states that all adaptive biological agents seek to minimize long-term surprise (i.e. entropy) (Friston 2010). Although this principle is normative (i.e. axiomatic, self-evident, non-falsifiable; Allen 2018) in nature and does not provide

mechanistic explanations of brain function, it can be used when building process models. A process model of brain function is one that specifies the neural processes that bring about a cognitive capacity, in terms of their structure, mechanisms and information flow (Andrews 2021). An example of such a process model developed under the free energy principle is active inference (Friston et al. 2016), which realizes a variant of a neural algorithmic scheme called predictive processing.

Predictive processing is at the core of hypotheses proposing that the brain seeks to reduce surprise by inferring the (hidden) states of the world giving rise to our sensorial experience using Bayesian inference mechanisms (Friston and Kiebel 2009). Such inference can be made by instantiating a generative model which specifies the likelihood of observations given hidden states in the world, and a prior probability of each state. The neural system uses this generative model to compute a posterior probability of the causes behind those observations. This generative model is said to be hierarchical, where each level of the hierarchy encodes states at nested timescales and each level takes as observations the hidden states of the levels below (Friston et al. 2017; Friston et al. 2018).

Under the predictive processing lens, the predictions of the generative model are compared against real observations, and the difference between the model and the observations i.e., “prediction error” (formally equivalent to free energy) is transmitted upwards in the hierarchy (Friston and Kiebel 2009). The goal of the neural system is to minimize these prediction errors across hierarchical levels. This can be achieved in two ways: by inferring the states of the world that maximize the probability of the observation, which corresponds to perception and is typically modelled by predictive coding schemes (Friston & Kiebel, 2009; Bastos et al. 2012); or by sampling or acting in the world to increase the chance of meeting those predictions (Friston et al. 2016). Models that assume this last property are what we will call active inference models.

Active inference models imply that the system not only minimizes surprise in the here and now: it also minimizes *expected* surprise through a policy selection process. This minimization can be achieved in two ways: either by performing a pragmatic action that maximizes the probability of obtaining the rewards encoded by the priors, or by carrying out an epistemic action that maximizes information gain through exploration. The generative model must therefore possess beliefs about future states, and counterfactual beliefs that encode the probability of some state and its outcome conditional on having selected a particular policy (Friston et al. 2018). The transitions between states at each level in the hierarchy are contextualized by the levels above (Friston et al. 2017).

The organism is also said to estimate the precision of its beliefs. Thus, the brain not only encodes beliefs about states of the world, how they produce observations, and how they transition over time; it also quantifies and modulates confidence in those beliefs. These so-called precision estimates can weight the impact of prediction errors (gain function), and when deployed in a descending (top-down) manner as expectations, they are thought of as attentional mechanisms (Feldman and Friston 2010). In this way, the neural system can amplify or down-regulate the ascending (bottom-up) prediction errors depending on the context and the goals of the organism (Aukstulewicz et al. 2017).

Taken together, these computational principles make up the architecture of active inference, which can be used to construct process theories of specific cognitive phenomena.

3 Active Inference Accounts of Consciousness

As an all-encompassing framework for building process theories of brain functions, active inference should in principle be able to accommodate consciousness. Accordingly, several attempts have been made to link the general-purpose computational mechanisms put forward by active inference with the diverse explananda of consciousness. In this section we will review some examples of these proposals. We will only discuss those that exclusively use conceptual tools from active inference, but see Marvan and Havlík (2021) for an overview of proposals that mix active inference with mechanisms from other theories. We will also omit theories of specific aspects of consciousness (e.g., the phenomenological property of conscious presence, or the self) that are based on predictive coding mechanisms alone rather than active inference (e.g. Seth et al. 2012; Woźniak 2018).

Proponents of active inference affirm that consciousness, as any other biological process, can be explained by the Free Energy Principle and Bayesian theories of brain processing where action plays a key role in reducing uncertainty. As defenders of active inference have argued: “conscious processing is about inferring the causes of sensory states, and thereby navigating the world to elude surprises” (Friston 2018).

The fact that consciousness is an inferential process does not mean that all organisms that perform inference are conscious. Instead, the active inference framework proposes that the difference between conscious and unconscious organisms or states is that the former are endowed with thick temporal and deep counterfactual generative models (Friston 2018). A conscious system is able to infer states of the world that have not happened yet (i.e. temporal thickness), relative to a selected course of action (i.e. counterfactual depth). Temporal thickness and counterfactual depth are graded features of generative models that allow the inference of states further away in time, and to compare a greater number of policies. Since these two properties are assumed to underlie conscious phenomena, proponents of active inference therefore argue that consciousness must therefore be a graded phenomenon. Indeed, non-human organisms differ in the extent to which they exhibit consciousness signatures, and humans are more or less conscious at different points in time. The claim is that during states of reduced consciousness (e.g., sedation) the brain’s generative model might lose temporal thickness and counterfactual depth.

Regarding conscious contents, they are said to be determined by the hypothesis with the highest posterior probability, parametrized by the inferences made at each level of the hierarchy. This is the belief that best explains away prediction errors and the one that will be used to sample the world or act on it (Hohwy 2012). The contents of our phenomenological experience are thus inferred states: “(...) seeing red and feeling pain (...) are themselves inferred causes, constructed to accommodate (i.e. best explain) the raw sensory flux – and the hierarchical machinations they induce” (Clark et al. 2019). Some have suggested that the contents of consciousness particularly represent the middle level beliefs of this posterior hierarchy (Clark et al. 2019; Whyte and Smith 2020). The relationship between these proposals and the temporal and counterfactual depth of generative models remains a matter of debate.

Some phenomenological properties can be explained by the hierarchical nature of the generative model as well. Lower levels are said to track sensorial information over short periods of time, while higher levels process more abstract and amodal

representations over longer time-windows. Correspondingly, top levels might encode information using discrete (and lower dimensional) representations, while lower levels may deal with information in a continuous manner (Friston et al. 2017). This might explain why phenomena such as memory (Barron et al. 2020) or imagination, associated with higher levels of the cortical hierarchy and involving sensory systems to a lesser degree, can feel different than canonical sensorial experiences like perceiving the redness of a ball (Clark et al. 2019).

Similarly, hierarchical precision deployment might explain why humans are capable of attending to and manipulating their own beliefs of the states of the world (i.e. metacognitive capacities). As mentioned before, top-down precision-weighting is often equated with attentional processes where higher level states modulate the confidence and importance ascribed to different perceptual states. Meta-awareness states, where one becomes aware and consciously manipulates these internal attentional processes, can be then thought as precision deployment over those levels in the hierarchy that deploy attention modulating mechanisms themselves (Sandved Smith et al. 2020).

4 Active Inference as a Computational Framework for Consciousness

Although many of the multiple explanandums of consciousness have been conceptually described through the active inference lens and these theoretical claims are a useful guide for research, they remain preliminary in nature and do not constitute a process theory in a strict sense. They typically do not provide mechanistic explanations, which identify the entities, operations and organizational features that have a causal effect in producing the phenomena of interest, and specify how the relevant components are implemented in the underlying system (Craver 2006; Kaplan 2011). The theory linking active inference and consciousness needs to be further constrained and specified to provide a mechanistic account, as has been already pointed out (Friston et al. 2020; Wiese and Friston 2020).

So far, the majority of researchers in the field have tried to make theoretical advancements through more precise conceptual specifications (for the latest examples see Clark et al. 2019; Friston et al. 2020; Limanowski and Friston 2020; Wiese and Friston 2020). However, as it will be discussed in the remaining sections, a promising alternative for better-specifying a theory is by building computational models (defined here as computer code formalizations of verbal theories) that implement its proposals (Guest and Martin 2021). Simulations and empirical data can be used to test this model, detect any incorrect mechanistic assumptions and thus refine the underlying theory (Smaldino 2017).

Proponents of active inference have long made use of computational modelling practices in their research (Friston et al. 2018), and thus such efforts might well be underway. In the following, we will discuss how computational models and experimental data can be and have been used to refine and validate the explanatory theories of consciousness under the active inference framework. As before, we will omit from this revision models that only partially implement the active inference scheme, such as empirically-based models whose parameters are interpreted post-hoc in terms of predictive coding mechanisms (e.g. Boly et al. 2011; Sanders et al. 2014) and do not refer to policy selection processes.

We focus on the mechanistic framework as a tool for evaluating the progress the field has made given the explicit intent to provide a neural process theory of consciousness (Miłkowski 2016a).

4.1 Towards an Explanatory Computational Model

4.1.1 Target Phenomena

When building a mechanistic model, the first step is to clearly identify and characterize the phenomena to be explained (Craver 2006). Models can then be evaluated by their ability to produce and modulate this explanandum, and the mechanisms proposed as explanatory for producing the phenomena of interest can be validated or falsified.

As stated before, one of the challenges in the field is that the target of consciousness research remains a much-debated issue and broadly differs across theories. Computational models of consciousness have consequently defined their explanandum differently depending on their embedding theoretical framework. This makes the formal and quantitative comparison of the mechanistic proposals put forward by different theories a difficult task.

Consciousness as the target of computational modelling also bears another important constraint: the phenomenon in itself is subjective, meaning that its presence is only accessible from a first-person perspective (Metzinger 2000) and thus cannot be objectively determined. Computational models cannot be probed with phenomenological interviews or other introspective methods commonly used in research with humans to characterize in detail internal states and subjective experience. To sidestep this limitation, computational models have been evaluated by their ability to simulate or fit the findings of experimental studies probing the explananda of consciousness. For example, computational models have been built to simulate electrophysiological markers traditionally believed to index conscious access mechanisms (e.g. Whyte and Smith 2020). Simulation approaches of this kind are also framed as building a replicatively valid model, meaning one which can generate known outputs from known inputs (Miłkowski 2016b). Building such a model is a promising approach to start refining a theory of consciousness. It will allow us to investigate which mechanistic processes (instantiated by the computational model) can give rise to observables related to conscious experience (the experimental markers). Yet, parallel work that formally relate those input-output models and simulations to subjective qualities (Carter et al. 2018) will be needed to fully validate this approach, at least when it comes to understanding phenomenal properties of consciousness, as opposed to cognitive access.

The fragmented definition of consciousness and the diversity of findings regarding neural and behavioral markers makes building a computational model of consciousness seem like an especially challenging endeavour. As stated before, it has been suggested that a potential solution to these issues might be to leverage the general-purpose mechanistic framework of brain processing of active inference. Such framework provides a single conceptual and formal scaffolding for situating the diverse explananda of consciousness (i.e. states, contents, phenomenology, reportability), traditionally studied in isolation with incommensurable theories. Situating the explananda under a common framework of brain function, whose components' interaction we understand, would inspire ideas on how the elements of the explananda

themselves might interact to give rise to consciousness. This unifying approach has the potential of developing a mechanistic model that is able to predict diverse behavioral and neural markers of consciousness, and is able to provide an explanation of how these integrate to produce a conscious experience.

How far are active inference models from providing a unified account of consciousness? In this section we will provide a synthesis of the existing models (see Table 1). We will restrict our discussion to work targeting phenomena traditionally associated with consciousness, like perceptual awareness and phenomenological experience. We will omit computational models of other high-level cognitive capacities (e.g. working memory), but it is important to mention these might be useful for better characterizing consciousness as they indirectly involve it (Reggia et al. 2017).

Predictively valid models of a variety of consciousness explananda have been built using the mechanistic tools of active inference. Such models have provided a partial account of conscious access markers. As anticipated (Vilas and Melloni 2020), they have also been able to accommodate the diversity of contents of conscious experience by simulating correlates of meta-awareness and meta-cognitive processes (Smith et al. 2019; Sandved Smith et al. 2020), or reports of hallucinations (Benrimoh et al. 2018; Benrimoh et al. 2019). Most notably, some of these efforts illustrate how two aspects of consciousness (e.g., meta-awareness and conscious access) might interact in a single model to modulate the reportability of internal emotional states (Smith et al. 2019). In contrast with these advancements, no computational model of active inference has so far reproduced findings on reduced states of consciousness, or was used to accommodate the structurally invariant properties of phenomenology, like the integrative nature of subjective experience.

What tasks have researchers used as an interface between computational models and experimental data? Active inference studies have simulated binocular rivalry (Parr et al. 2019), masking (Whyte and Smith 2020), inattention blindness (Whyte and Smith 2020), Troxler fading illusion (Parr et al. 2019) and oddball paradigms (Sandved Smith et al. 2020), as well as a working memory paradigm tapping into conscious access mechanisms (Smith et al. 2019), and a speech paradigm eliciting hallucinations (Benrimoh et al. 2018; Benrimoh et al. 2019). All studies used relatively simple stimuli with a low degree of ecological validity.

Active inference models simulate posterior beliefs of synthetic participants over the hidden states of a task, and these are then used to model behavioral and neural responses. The key contribution of this body of work is to use a similarly parameterized generative model to describe the evolution of hidden states with active inference mechanisms. At the same time, the models contain customized response models which map these hidden states to specific types of behavioral or neural markers, addressing the diversity of the explanandum. As behavioral markers, reports of perceptual experience (Whyte and Smith 2020), internal states (Smith et al. 2019) and visual saccades (Parr et al. 2019) were simulated. Neural markers included simulated firing rate responses (Smith et al. 2019; Whyte and Smith 2020), local field potentials (Smith et al. 2019) and event-related potentials (ERPs) (Whyte and Smith 2020). No work has so far simulated differences in the spatial distribution of brain activity.

Remarkably, active inference models have already helped to conceptually integrate the results of different experimental tasks probing the same aspect of consciousness. Whyte and Smith (2020) developed a model that could mimic the

Table 1 Active inference models of consciousness processes. This table summarizes the experimental paradigms and the behavioral and neural responses that were simulated in studies testing computational models of active inference. It also highlights which components of the computational model were modulated to better characterize and validate the relevant mechanisms for consciousness put forward by the active inference theory

	<i>Simulated task</i>	<i>Simulated behavioral markers</i>	<i>Simulated neural markers</i>	<i>Modulated model parameters</i>
Parr et al. (2019)	Troxler fading illusion	Saccades	None	Preferences over outcomes Transition precision
Whyte and Smith (2020)	Binocular rivalry	Forced choice report of perceptual content	Firing rates ERPs	Likelihood precision
	Masking Inattentional blindness			Likelihood precision modulated by parameters of signal strength and attention
Benrimoh et al. (2018)	Speech task	None ^a	None	Initial probabilities of hidden states Likelihood precision Precision over policies
Benrimoh et al. (2019)	Speech task	None ^a	None	Policy space Likelihood precision
Sandvev Smith et al. (2020)	Oddball paradigm (attentional variant)	None	None	Content transition matrix
				Likelihood precision
Smith et al. (2019)	Working memory task	Internal state categorization report	Firing rates Local field potentials	Top-down modulation of likelihood precision (of first order perceptual states and second order attentional states)
				Prior expectations of states Likelihood precision Prior bias over policies Transition precision
				Second order likelihood precision Second level transition precision

^a These studies simulated the content assumed to be perceived by the synthetic participant, but not the results of an experimental paradigm probing for such content

findings of two studies reporting the modulation of the P3 component (a marker once thought to signal access consciousness) in response to different experimental manipulations: one tapping into the attention of participants and the other into their expectation. The authors parameterized these input conditions into their model to independently modulate precision estimations, and simulated experiments that systematically varied their values to disentangle their effects on the P3. As another example, Parr et al. (2019) modelled how beliefs about the precision of states transitions might account for the stability of conscious perception in two different perceptual awareness tasks performed by the same synthetic subject. These two examples illustrate the potential this modelling framework has for integrating the results of different studies.

In sum, replicatively valid models using the active inference framework have been deployed to accommodate a number of consciousness explananda. This preliminary but quantitatively shows that this modelling framework is a serious candidate for providing a neural process theory of consciousness (broadly defined). Still, to further demonstrate this capacity studies need to extend these models to account for altered states of consciousness and the structure of phenomenological experience. Experiments should also simulate other neural markers associated with consciousness, since signatures such as the P3 are often contested in their relation to consciousness (Verleger 2020).

Having reviewed the ways in which active inference models deal with the diversity of the explanandum, we will now move on to the second step of building a mechanistic model: identifying the components that are critical for the theory and characterizing their effects on the phenomena of interest.

4.1.2 Explanatory Components

Not all of the mechanisms put forward by an all-encompassing model of brain function like active inference might be relevant for building an explanation of consciousness. A complete mechanistic model is one that identifies, specifically, all the *causally relevant* components for producing the phenomena. As a desiderata, the computational model should specify how the causally relevant components of the theory are instantiated, and separate them from auxiliary assumptions (Lee et al. 2019).

Equipped with a well spell-out set of critical components and their specification, then the assumptions and behavior of the model can be tested through simulations of empirical data (Smaldino 2017; Wilson and Collins 2019). A number of practices can be used to determine the components of the model truly relevant for producing the phenomena, and refine the mechanistic proposal of the theory.

For instance, critical insights can be gathered through simulating the whole parameter range of the model components to determine how this variation affects the simulated responses and how the elements of the model jointly influence the target. This approach can also test how robust the model is to changes of its auxiliary assumptions (Nassar and Frank 2016). In addition, model fit procedures can be used to estimate the parameter values that better explain the simulated data, or to contrast the explanatory power of a model against alternative ones (Wilson and Collins 2019). For instance, parameters of each model can be systematically pruned (e.g. using Bayesian model reduction; Friston et al. 2019), and alternative models can be explicitly compared (e.g. using Bayesian model selection; Friston and Penny 2011) and/or fused (e.g.,

using Bayesian model averaging; Trujillo-Barreto et al. 2004). Beyond this, the models fitted on particular datasets or trained for solving particular tasks could also be used to explain other datasets or tasks relevant to the phenomena of interest, to test the generalizability of the mechanisms proposed. These practices guard against confirmation bias (see Farrell and Lewandowsky 2010) and evaluate the level of commitment one should have to the theory.

We will now review how these practices have been implemented for refining theories linking active inference mechanisms and consciousness phenomena. Conceptual work has already suggested which specific components of the active inference framework might be those driving conscious experience. For instance, much work has been done in modelling the role of precision estimates and their hierarchical deployment in determining the content of a conscious experience. This component plays a central role on active inference as the content is defined by the hypothesis with the largest posterior probability which in turn is affected by the precision ascribed to those beliefs (Hohwy 2012).

To investigate the effect of the precision ascribed to likelihood estimates on the contents of experience, studies have manipulated their parameter values to determine how they affect conscious reports across masking and inattention blindness paradigms (Whyte and Smith 2020), as well as the number of perceptual switches in a binocular rivalry paradigm (Parr et al. 2019). In conjunction, both experiments demonstrate that increasing the uncertainty of the mapping between observations and the hidden states of the world influence if a content is consciously perceived or not. Going further, Benrimoh et al. (2018) analyzed how the precision estimation of different components of the model jointly interact to influence the target, showing that high prior precisions over policies can influence perception and induce hallucinations as long as they are accompanied by diminished likelihood precision, since in this scenario prior policies will have increased influence in establishing the perceptual hypothesis with the higher posterior probability.

Other work has characterized the role of priors in determining the contents of consciousness. Benrimoh et al. (2019) showed how hallucinations whose content is incongruous with the environment can only be simulated if an incorrect prior is added to the model. Similarly, Whyte and Smith (2020) reported that a prior consistent with new perceptual input increases the probability of the stimuli being consciously perceived. Since priors also affect the posterior probability of different perceptual hypothesis, taken together these studies further validate active inference's hypothesis that such model component is the one determining the content of conscious experience.

In contrast to concepts such as precision estimation and the content of the priors, other elements of the theoretical proposal linking active inference with consciousness remain underspecified. For example, the temporal thickness and counterfactual depth of the generative model, which are thought to index consciousness levels, remain ambiguously specified in the computational models. Some researchers have modelled the temporal depth as nested beliefs of hidden states (e.g. Whyte and Smith 2020), others as nested beliefs of model parameters (i.e. beliefs affecting the precision of components at lower levels of the hierarchy; see Sandved Smith et al. 2020). Those mechanisms could be complementary, yet whether and how those mechanism might differentially modulate consciousness levels has not been studied. The temporal depth property has also been translated to conscious access terms by defining it as the ability of the top levels of

the model's hierarchy to integrate information from lower levels, and transmit this integration in a top-down manner (Smith et al. 2019; Whyte and Smith 2020). The counterfactual property, which has been defined as the number of policies instantiated by the model, has not been manipulated to test its effects on consciousness levels, and the interaction with the temporal depth has not been investigated.

Not only the computational implementations of some of the relevant components of the modelling framework are unclear. These studies do not elucidate, and consequently manipulate, which are the auxiliary assumptions of the model. They also lack formal comparisons between the accuracy of their models from those of other consciousness theories. Simulation approaches are not the only way to fill these modelling gaps. Empirical data can also be used for such purposes, as we will explore in the next section.

4.2 Empirical Testing

4.2.1 Model Predictions

At some point during the theory development process, collecting empirical data becomes necessary to validate the adequacy of the computational model and further refine its mechanistic proposals. During the simulation process researchers parameterize and vary the explanatory components of the model, as well as the initial conditions or auxiliary assumptions, to characterize the mechanisms giving rise to the phenomena of interest. This means the model should be able to make a precise prediction of the behavior of the system under multiple parameter combinations, some of which possibly have not been tested empirically before. This is referred to as a “predictively valid model”, that is, a model that can correctly output predictions matching new observations (Milkowski 2016b).

Importantly, it is unnecessary and often impossible to test predictions stemming from the whole parameter space of the model. This is especially true for highly parameterized and complex models like those of active inference. The researcher needs to identify those predictions whose confirmation or falsification would be the most informative for understanding and assessing the adequacy of the explanatory mechanisms put forward by the theory (Wilson and Collins 2019). Simulations can be used to find those initial values where the contrast between predictions would be greater. For example, assuming some initial conditions, model X with component *x* would predict effect A, while model Y that is identical to model X but without component *x* would predict effect B. An experiment replicating those initial conditions can help disambiguate how necessary component *x* is for explaining the phenomena. Model fit metrics can then be used to compare how well these alternative mechanisms or models explain real and specific datasets (Lee et al. 2019). This approach can also be used to contrast the adequacy of models built from different theories. Adversarial collaboration projects are an example of this, where the most diverging predictions made by two competing theories are identified, and experiments are carried out by different research teams to verify or falsify those diagnostic predictions (Melloni et al. 2021).

How have active inference models been empirically tested to refine their mechanistic explanation of consciousness? Unfortunately, to the best of our knowledge, no study using this computational framework has motivated the collection of new empirical data

to evaluate and refine the models. More than 50 theoretical journal articles have been published making verbal and mathematical claims on how active inference might explain consciousness. The number of articles implementing these theoretical proposals into a computational model is considerably smaller – six studies according to this review. Of these, none has directly compared simulation results against new experimental data. Instead, they have been evaluated by their ability to reproduce findings from previous studies.

Yet, computational models of active inference that advance precise hypotheses to be tested in future empirical work have been put forward. For instance, through simulations Whyte and Smith (2020) identified attention and expectation as independent modulators of the common marker of consciousness P3, and proposed a new variant of a Posner cueing paradigm that would separately manipulate these components in a real experimental setting.

Empirical testing of the computational model is not only crucial for probing the effects of the explanatory components on the phenomena of interest. As the next section will illustrate, it is also the default approach for investigating how these components are instantiated in the system underlying the real capacity, i.e., their neural implementation.

4.2.2 Neural Implementation

It has been argued that a mechanistic explanation of a mental capacity such as consciousness is not complete until it shows how it is implemented in the human brain (Kaplan 2011). In other words, a mechanistic explanation requires building a structurally valid model, meaning one whose structure has shown a correspondence with that of the physical system implementing the capacity (Miłkowski 2016b). To create such a model in the neuroscience domain, the researcher needs to map the entities, properties and operations of the computational model to the anatomy and dynamics of the brain.

Detailed proposals have been put forward for how the components of the active inference model might be instantiated in the brain (Parr and Friston 2018). Researchers have measured how the behavior of these mechanisms can be explained by the activity of relevant neural structures or processes, or vice versa. For example, Schwartenbeck et al. (2015) showed that, in accordance with one of the anatomical proposals of active inference, participants' dopaminergic activity during a reward task can be predicted by the modelled precision estimates.

This modelling approach can also help refine the verbal and mathematical formulation of the theory and the computational model itself. For example, a biophysically realistic models of predictive coding and/or active inference would need to specify whether the term “precision” is implemented in the computational model as putative classical neuromodulation, other (e.g. NMDA-dependent) gain control mechanisms, synchronous activity, divisive normalization, lateral inhibition, or combinations thereof (Friston 2012; Shipp 2016; Auksztulewicz et al. 2018; Northoff and Mushiakhe 2020).

So far, methods that statistically relate brain recordings with the mechanistic components of the consciousness models of active inference are still missing. Some preliminary hypotheses have been nonetheless spelled out. For instance, Whyte and Smith (2020) described how the top-level of their model might mirror the activity of prefrontal and parietal areas while the lower-level would correspond to those of sensory regions. However, this proposal

remains conceptual in nature. Although they simulated neural firing rates independently encoding the posterior beliefs at each level of the hierarchy, they did not use these to simulate brain recordings based on high spatial resolution techniques (e.g. fMRI). This would have enabled a more direct comparison of their results to studies outlining the role of these brain regions in conscious access processes.

More generally, experiments such as the one of Whyte and Smith (2020) would also benefit from model-based fMRI approaches (O'Doherty et al. 2007), which use model fitting and regression techniques to identify the areas of the brain whose activity correlates with the neural activity predicted by the computational model.

A similar promising tool for linking neural recordings of a particular dataset with a computational model of brain function is dynamical causal modeling (DCM; Friston et al. 2003). DCMs are generative models of neural activity that can be used to infer brain responses to experimental manipulations by fitting the model to a specific dataset. One can use this approach to test how well the model explains real data, as well as to elucidate the optimal parameters values in particular experimental scenarios. DCMs have been used to evaluate predictive coding models of consciousness states and have shown, for example, that top-down processes projections from higher levels of the cortical hierarchy are impaired in the vegetative state (Boly et al. 2011), or that they are altered in visual illusions such as apparent motion (Sanders et al. 2014). Despite this progress, so far active inference models have not been combined with DCM techniques to directly fit empirical results of consciousness research, but rather used to reproduce more general patterns in data abstracted from several studies.

5 Moving Forward

We have reviewed some good examples of how computational models and simulation approaches have been used to specify and validate verbal theories linking mechanisms from active inference with conscious access processes and the contents of conscious experience. But this work is still very preliminary in nature, and the consciousness explananda is still partly, and disjointly, situated within the active inference framework.

To make significant progress, further studies need to deepen the knowledge on how the mechanistic components integrate and influence one another. Future work could also extend these efforts to cover more systematically the diversity of conscious contents (e.g. internal thoughts, representations about the self), for example by using a dimensional approach (Studerus et al. 2010; Lutz et al. 2015; Birch et al. 2020). But more importantly, the already developed models need to be tested against newly collected empirical data to test their predictive and structural validity.

Parallel to these, to fully account for the broad explanandum of consciousness, active inference models need to be developed to accommodate the spectrum of states of consciousness and, most critically, the structure and qualities of experience. Recently, Ramstead et al. (2021) have proposed an account termed computational phenomenology based on linking phenomenological data with computational generative models of experience in an attempt to formally describe the structure of consciousness. This proposal which remains to be explicitly formalized leaves the physical realization of consciousness aside, merely pursuing a methodological naturalization of phenomenology. Extending it to encompass a physical and neural implementation will go a long way to address the subjective qualities of experience.

Altogether, the development of computational models linking active inference mechanisms to consciousness phenomena holds the promise of advancing the theory by making more explicit the mechanism that within the framework relate to consciousness, while also opening avenues for future empirical research. Advancing the theory, however, suspends the question of what the explananda of the problem of consciousness are; and while the active inference framework is not expected to resolve this question, researchers of consciousness should ask themselves what a theory is about. The mystery of consciousness is and will always be its intrinsic, subjective nature, and how phenomenal properties are embedded in physical systems. For active inference (and/or any other theory) to be called a theory of consciousness, it needs to address these aspects.

Funding Open Access funding enabled and organized by Projekt DEAL. This work has been supported by the Max Planck Society and the European Commission's Marie Skłodowska-Curie Global Fellowship (750459 to R.A.).

Declarations

Conflict of Interest The authors have no conflicts of interest to declare that are relevant to the content of this article.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Alkire, M.T., A.G. Hudetz, and G. Tononi. 2008. Consciousness and anesthesia. *Science* 322: 876–880. <https://doi.org/10.1126/science.1149213>.
- Allen, Micah. 2018. The foundation: Mechanism, prediction, and falsification in Bayesian enactivism. *Physics of Life Reviews* 24: 17–20. <https://doi.org/10.1016/j.plrev.2018.01.007>.
- Andrews, Mel. 2021. The math is not the territory: navigating the free energy principle. *Biology & Philosophy* 36: 30. <https://doi.org/10.1007/s10539-021-09807-0>.
- Auksztulewicz, Ryszard, Karl J. Friston, and Anna C. Nobre. 2017. Task relevance modulates the behavioural and neural effects of sensory predictions. Edited by Ole Jensen. *PLOS Biology* 15: e2003143. <https://doi.org/10.1371/journal.pbio.2003143>.
- Auksztulewicz, Ryszard, Caspar M. Schwiedrzik, Thomas Thesen, Werner Doyle, Orrin Devinsky, Anna C. Nobre, Charles E. Schroeder, Karl J. Friston, and Lucia Melloni. 2018. Not all predictions are equal: “What” and “when” predictions modulate activity in auditory cortex through different mechanisms. *The Journal of Neuroscience* 38: 8680–8693. <https://doi.org/10.1523/JNEUROSCI.0369-18.2018>.
- Bachmann, Talis, and Anthony G. Hudetz. 2014. It is time to combine the two main traditions in the research on the neural correlates of consciousness: $C = L \times D$. *Frontiers in Psychology* 5. doi: <https://doi.org/10.3389/fpsyg.2014.00940>.
- Barron, Helen C., Ryszard Auksztulewicz, and Karl Friston. 2020. Prediction and memory: A predictive coding account. *Progress in Neurobiology* 192: 10182.

- Bastos, Andre M., W. Martin Usrey, Rick A. Adams, George R. Mangun, Pascal Fries, and Karl J. Friston. 2012. Canonical microcircuits for predictive coding. *Neuron* 76: 695–711. <https://doi.org/10.1016/j.neuron.2012.10.038>.
- Bayne, Tim. 2007. Conscious states and conscious creatures: Explanation in the scientific study of consciousness. *Philosophical Perspectives* 21: 1–22. <https://doi.org/10.1111/j.1520-8583.2007.00118.x>.
- Bayne, Tim, Jakob Hohwy, and Adrian M. Owen. 2016. Are there levels of consciousness? *Trends in Cognitive Sciences* 20: 405–413. <https://doi.org/10.1016/j.tics.2016.03.009>.
- Benrimoh, David, Thomas Parr, Peter Vincent, Rick A. Adams, and Karl Friston. 2018. Active inference and auditory hallucinations. *Computational Psychiatry* 2: 183. https://doi.org/10.1162/CPSY_a_00022.
- Benrimoh, David, Thomas Parr, Rick A. Adams, and Karl Friston. 2019. Hallucinations both in and out of context: An active inference account. Edited by Constantine Dovrolis. *PLOS ONE* 14: e0212379. <https://doi.org/10.1371/journal.pone.0212379>.
- Berkovich-Ohana, Aviva, Yair Dor-Ziderman, Fynn-Mathis Trautwein, Yoav Schweitzer, Ohad Nave, Stephen Fulder, and Yochai Ataria. 2020. The Hitchhiker's guide to neurophenomenology – the case of studying self boundaries with meditators. *Frontiers in Psychology* 11: 1680. <https://doi.org/10.3389/fpsyg.2020.01680>.
- Birch, Jonathan, Alexandra K. Schnell, and Nicola S. Clayton. 2020. Dimensions of animal consciousness. *Trends in Cognitive Sciences* 24: 789–801. <https://doi.org/10.1016/j.tics.2020.07.007>.
- Block, Ned. 1990. Consciousness and accessibility. *Behavioral and Brain Sciences* 13: 596–598. <https://doi.org/10.1017/S0140525X00080316>.
- Block, Ned. 1992. Begging the question against phenomenal consciousness. *Behavioral and Brain Sciences* 15: 205–206. <https://doi.org/10.1017/S0140525X00068266>.
- Block, Ned. 1995. On a confusion about a function of consciousness. *Behavioral and Brain Sciences* 18: 47.
- Block, Ned. 2005. Two neural correlates of consciousness. *Trends in Cognitive Sciences* 9: 46–52. <https://doi.org/10.1016/j.tics.2004.12.006>.
- Boly, M., M.I. Garrido, O. Gosseries, M.-A. Bruno, P. Boveroux, C. Schnakers, M. Massimini, V. Litvak, S. Laureys, and K. Friston. 2011. Preserved feedforward but impaired top-down processes in the vegetative state. *Science* 332: 858–862. <https://doi.org/10.1126/science.1202043>.
- Carter, Olivia, Jakob Hohwy, Jeroen van Boxtel, Victor Lamme, Ned Block, Christof Koch, and Naotsugu Tsuchiya. 2018. Conscious machines: Defining questions. Edited by Jennifer Sills. *Science* 359: 400. <https://doi.org/10.1126/science.aar4163>.
- Chalmers, David J. 1995. Facing up to the problem of consciousness. *Journal of Consciousness Studies* 2: 200–219.
- Chalmers, David J. 2000. What is a neural correlate of consciousness? In *Neural correlates of consciousness: Empirical and conceptual questions*. The MIT Press, 17–39.
- Clark, Andy, Karl Friston, and Sam Wilkinson. 2019. Bayesing qualia: Consciousness as inference, not raw datum. *Journal of Consciousness Studies* 26: 19–33.
- Craver, Carl F. 2006. When mechanistic models explain. *Synthese* 153: 355–376. <https://doi.org/10.1007/s11229-006-9097-x>.
- Crick, F., and C Koch. 1990. Towards a neurobiological theory of consciousness. *Seminars in the Neurosciences* 2.
- Doerig, Adrien, Aaron Schurger, and Michael H. Herzog. 2021. Hard criteria for empirical theories of consciousness. *Cognitive Neuroscience* 12: 41–62. <https://doi.org/10.1080/17588928.2020.1772214>.
- Farrell, Simon, and Stephan Lewandowsky. 2010. Computational models as aids to better reasoning in psychology. *Current Directions in Psychological Science* 19: 329–335. <https://doi.org/10.1177/0963721410386677>.
- Feldman, Harriet, and Karl J. Friston. 2010. Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience* 4. doi: <https://doi.org/10.3389/fnhum.2010.00215>.
- Friston, Karl. 2010. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience* 11: 127–138. <https://doi.org/10.1038/nrn2787>.
- Friston, Karl. 2012. Predictive coding, precision and synchrony. *Cognitive Neuroscience* 3: 238–239. <https://doi.org/10.1080/17588928.2012.691277>.
- Friston, Karl. 2018. Am I self-conscious? (or does self-organization entail self-consciousness?). *Frontiers in Psychology* 9: 579. <https://doi.org/10.3389/fpsyg.2018.00579>.
- Friston, Karl, and Stefan Kiebel. 2009. Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364: 1211–1221. <https://doi.org/10.1098/rstb.2008.0300>.
- Friston, Karl, and Will Penny. 2011. Post hoc Bayesian model selection. *NeuroImage* 56: 2089–2099. <https://doi.org/10.1016/j.neuroimage.2011.03.062>.

- Friston, K.J., L. Harrison, and W. Penny. 2003. Dynamic causal modelling. *NeuroImage* 19: 1273–1302. [https://doi.org/10.1016/S1053-8119\(03\)00202-7](https://doi.org/10.1016/S1053-8119(03)00202-7).
- Friston, Karl, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, and Giovanni Pezzulo. 2016. Active inference: A process theory. *Neural Computation* 29: 1–49. https://doi.org/10.1162/NECO_a_00912.
- Friston, Karl, Thomas Parr, and Bert de Vries. 2017. The graphical brain: Belief propagation and active inference. *Network Neuroscience* 1: 381–414. https://doi.org/10.1162/NETN_a_00018.
- Friston, Karl, Richard Rosch, Thomas Parr, Cathy Price, and Howard Bowman. 2018. Deep temporal models and active inference. *Neuroscience & Biobehavioral Reviews* 90: 486–501. <https://doi.org/10.1016/j.neubiorev.2018.04.004>.
- Friston, Karl, Thomas Parr, and Peter Zeidman. 2019. Bayesian model reduction. *arXiv:1805.07092 [stat]*.
- Friston, Karl, Wanja Wiese, and J. Allan Hobson. 2020. Sentience and the origins of consciousness: From Cartesian duality to Markovian monism. *Entropy (Basel, Switzerland)* 22. doi: <https://doi.org/10.3390/e22050516>.
- Guest, Olivia, and Andrea E. Martin. 2021. How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*: 174569162097058. doi: <https://doi.org/10.1177/1745691620970585>, 789, 802.
- Haun, Andrew, and Giulio Tononi. 2019. Why does space feel the way it does? Towards a principled account of spatial experience. *Entropy* 21: 1160. <https://doi.org/10.3390/e21121160>.
- Hohwy, Jakob. 2012. Attention and conscious perception in the hypothesis testing brain. *Frontiers in Psychology* 3. doi: <https://doi.org/10.3389/fpsyg.2012.00096>.
- Hohwy, Jakob, and Anil Seth. 2020. Predictive processing as a systematic basis for identifying the neural correlates of consciousness. *Philosophy and the Mind Sciences* 1. doi: <https://doi.org/10.33735/phimisci.2020.II.64>.
- Horowitz, S.G., A.R. Braun, W.S. Carr, D. Picchioni, T.J. Balkin, M. Fukunaga, and J.H. Duyn. 2009. Decoupling of the brain's default mode network during deep sleep. *Proceedings of the National Academy of Sciences* 106: 11376–11381. <https://doi.org/10.1073/pnas.0901435106>.
- Kaplan, David Michael. 2011. Explanation and description in computational neuroscience. *Synthese* 183: 339–373. <https://doi.org/10.1007/s11229-011-9970-0>.
- Kim, Chai-Youn, and Randolph Blake. 2005. Psychophysical magic: rendering the visible 'invisible'. *Trends in Cognitive Sciences* 9: 381–388. <https://doi.org/10.1016/j.tics.2005.06.012>.
- Koch, Christof, Marcello Massimini, Melanie Boly, and Giulio Tononi. 2016. Neural correlates of consciousness: Progress and problems. *Nature Reviews Neuroscience* 17: 307–321. <https://doi.org/10.1038/nrn.2016.22>.
- Laureys, Steven. 2005. The neural correlate of (un)awareness: Lessons from the vegetative state. *Trends in Cognitive Sciences* 9: 556–559. <https://doi.org/10.1016/j.tics.2005.10.010>.
- Lee, Michael D., Amy H. Criss, Berna Devezer, Christopher Donkin, Alexander Etz, Fábio P. Leite, Dora Matzke, Jeffrey N. Rouder, Jennifer S. Trueblood, Corey N. White, and Joachim Vandekerckhove. 2019. Robust modeling in cognitive science. *Computational Brain & Behavior* 2: 141–153. <https://doi.org/10.1007/s42113-019-00029-y>.
- Limanowski, Jakub, and Karl Friston. 2020. Attenuating oneself: An active inference perspective on “selfless” experiences. *Philosophy and the Mind Sciences* 1: 1–16. <https://doi.org/10.33735/phimisci.2020.I.35>.
- Lutz, A., J.-P. Lachaux, J. Martinerie, and F.J. Varela. 2002. Guiding the study of brain dynamics by using first-person data: Synchrony patterns correlate with ongoing conscious states during a simple visual task. *Proceedings of the National Academy of Sciences* 99: 1586–1591. <https://doi.org/10.1073/pnas.032658199>.
- Lutz, Antoine, Amishi P. Jha, John D. Dunne, and Clifford D. Saron. 2015. Investigating the phenomenological matrix of mindfulness-related practices from a neurocognitive perspective. *American Psychologist* 70: 632–658. <https://doi.org/10.1037/a0039585>.
- Marvan, Tomáš, and Marek Havlík. 2021. Is predictive processing a theory of perceptual consciousness? *New Ideas in Psychology* 61: 100837. <https://doi.org/10.1016/j.newideapsych.2020.100837>.
- Massimini, M. 2005. Breakdown of cortical effective connectivity during sleep. *Science* 309: 2228–2232. <https://doi.org/10.1126/science.1117256>.
- Melloni, Lucia, Liad Mudrik, Michael Pitts, and Christof Koch. 2021. Making the hard problem of consciousness easier. *Science* 372: 911–912. <https://doi.org/10.1126/science.abj3259>.
- Metzinger, Thomas. 2000. Introduction: Consciousness research at the end of the twentieth century. In *Neural correlates of consciousness: Empirical and conceptual questions*, 1–12. Place of publication not identified: Publisher not identified.

- Milkowski, Marcin. 2016a. A mechanistic account of computational explanation in cognitive science and computational neuroscience. In *Computing and philosophy*, ed. Vincent C. Müller, 191–205. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-23291-1_13.
- Milkowski, Marcin. 2016b. Explanatory completeness and idealization in large brain simulations: A mechanistic perspective. *Synthese* 193: 1457–1478. <https://doi.org/10.1007/s11229-015-0731-3>.
- Nassar, Matthew R., and Michael J. Frank. 2016. Taming the beast: Extracting generalizable knowledge from computational models of cognition. *Current Opinion in Behavioral Sciences* 11: 49–54. <https://doi.org/10.1016/j.cobeha.2016.04.003>.
- Nisbett, Richard E., and Timothy D. Wilson. 1977. Telling more than we can know: Verbal reports on mental processes. *Psychological Review* 84: 231–259.
- Northoff, Georg, and Hajime Mushiake. 2020. Why context matters? Divisive normalization and canonical microcircuits in psychiatric disorders. *Neuroscience Research* 156: 130–140. <https://doi.org/10.1016/j.neures.2019.10.002>.
- O'Doherty, J.P., A. Hampton, and H. Kim. 2007. Model-based fMRI and its application to reward learning and decision making. *Annals of the New York Academy of Sciences* 1104: 35–53. <https://doi.org/10.1196/annals.1390.022>.
- Parr, Thomas, and Karl J. Friston. 2018. The anatomy of inference: Generative models and brain structure. *Frontiers in Computational Neuroscience* 12: 90. <https://doi.org/10.3389/fncom.2018.00090>.
- Parr, Thomas, Andrew W. Corcoran, Karl Friston, and Jakob Hohwy. 2019, 2019. Perceptual awareness and active inference. *Neuroscience of Consciousness*: niz012. <https://doi.org/10.1093/nc/niz012>.
- Ramstead, Maxwell James, Casper Hesp, Lars Sandved-Smith, Jonas Mago, Michael Lifshitz, Giuseppe Pagnoni, Ryan Smith, et al. 2021. *From generative models to generative passages: A computational approach to (neuro)phenomenology*. Preprint. PsyArXiv. doi: <https://doi.org/10.31234/osf.io/k9pbn>.
- Reggia, James, Di-Wei Huang, and Garrett Katz. 2017. Exploring the computational explanatory gap. *Philosophies* 2: 5. <https://doi.org/10.3390/philosophies2010005>.
- Sanders, Lia Lira Olivier, Ryszard Auksztulewicz, Friederike U. Hohlefeld, Niko A. Busch, and Philipp Sterzer. 2014. The influence of spontaneous brain oscillations on apparent motion perception. *NeuroImage* 102: 241–248. <https://doi.org/10.1016/j.neuroimage.2014.07.065>.
- Sandved Smith, Lars, Casper Hesp, Antoine Lutz, Jérémie Mattout, Karl Friston, and Maxwell Ramstead. 2020. *Towards a formal neurophenomenology of metacognition: Modelling meta-awareness, mental action, and attentional control with deep active inference*. Preprint. PsyArXiv. doi: <https://doi.org/10.31234/osf.io/5jh3c>.
- Schwartenbeck, Philipp, Thomas H.B. FitzGerald, Christoph Mathys, Ray Dolan, and Karl Friston. 2015. The dopaminergic midbrain encodes the expected certainty about desired outcomes. *Cerebral Cortex* 25: 3434–3445. <https://doi.org/10.1093/cercor/bhu159>.
- Seth, Anil K., Keisuke Suzuki, and Hugo D. Critchley. 2012. An interoceptive predictive coding model of conscious presence. *Frontiers in Psychology* 2. doi: <https://doi.org/10.3389/fpsyg.2011.00395>.
- Shipp, Stewart. 2016. Neural elements for predictive coding. *Frontiers in Psychology* 7. doi: <https://doi.org/10.3389/fpsyg.2016.01792>.
- Siclari, Francesca, Benjamin Baird, Lampros Perogamvros, Giulio Bernardi, Joshua J. LaRocque, Brady Riedner, Melanie Boly, Bradley R. Postle, and Giulio Tononi. 2017. The neural correlates of dreaming. *Nature Neuroscience* 20: 872–878. <https://doi.org/10.1038/nn.4545>.
- Signorelli, Camilo Miguel, Joanna Szczotka, and Robert Prentner. 2021. Explanatory profiles of models of consciousness- towards a systematic classification. Preprint. PsyArXiv. doi: <https://doi.org/10.31234/osf.io/f5vdu>.
- Smaldino, Paul E. 2017. Models are stupid, and we need more of them. In *Computational social psychology*, 311–331.
- Smith, Ryan, Richard D. Lane, Thomas Parr, and Karl J. Friston. 2019. Neurocomputational mechanisms underlying emotional awareness: Insights afforded by deep active inference and their potential clinical relevance. *Neuroscience & Biobehavioral Reviews* 107: 473–491. <https://doi.org/10.1016/j.neubiorev.2019.09.002>.
- Smith, Ryan, Karl Friston, and Christopher Whyte. 2021. *A step-by-step tutorial on active inference and its application to empirical data*. Preprint. PsyArXiv. Doi: <https://doi.org/10.31234/osf.io/b4jm6>.
- Song, Chen, Andrew M. Haun, and Giulio Tononi. 2017. Plasticity in the structure of visual space. *ENEURO* 4: ENEURO.0080-17.2017. doi: <https://doi.org/10.1523/ENEURO.0080-17.2017>.
- Studerus, Erich, Alex Gamma, and Franz X. Vollenweider. 2010. Psychometric evaluation of the altered states of consciousness rating scale (OAV). Edited by Vaughan Bell. *PLoS ONE* 5: e12412. <https://doi.org/10.1371/journal.pone.0012412>.

- Tononi, Giulio, Melanie Boly, Marcello Massimini, and Christof Koch. 2016. Integrated information theory: From consciousness to its physical substrate. *Nature Reviews Neuroscience* 17: 450–461. <https://doi.org/10.1038/nrn.2016.44>.
- Trujillo-Barreto, Nelson J., Eduardo Aubert-Vázquez, and Pedro A. Valdés-Sosa. 2004. Bayesian model averaging in EEG/MEG imaging. *NeuroImage* 21: 1300–1319. <https://doi.org/10.1016/j.neuroimage.2003.11.008>.
- Varela, Francisco J. 1996. Neurophenomenology: A methodological remedy for the hard problem. *Journal of Consciousness Studies* 3: 330–349.
- Verleger, Rolf. 2020. Effects of relevance and response frequency on P3b amplitudes: Review of findings and comparison of hypotheses about the process reflected by P3b. *Psychophysiology* 57: e13542. <https://doi.org/10.1111/psyp.13542>.
- Vilas, Martina G., and Lucia Melloni. 2020. A challenge for predictive coding: Representational or experiential diversity? *Behavioral and Brain Sciences* 43: e150. <https://doi.org/10.1017/S0140525X19003157>.
- Whyte, Christopher J., and Ryan Smith. 2020. The predictive global neuronal workspace: A formal active inference model of visual consciousness. *Progress in Neurobiology* 101918: 101918. <https://doi.org/10.1016/j.pneurobio.2020.101918>.
- Wiese, Wanja. 2018. Toward a mature science of consciousness. *Frontiers in Psychology* 9: 693. <https://doi.org/10.3389/fpsyg.2018.00693>.
- Wiese, Wanja. 2020. The science of consciousness does not need another theory, it needs a minimal unifying model. *Neuroscience of Consciousness* 2020: niaa013. <https://doi.org/10.1093/nc/niaa013>.
- Wiese, Wanja, and Karl Friston. 2020. *The neural correlates of consciousness under the free energy principle: From computational correlates to computational explanation*. Preprint. PsyArXiv. Doi: <https://doi.org/10.31234/osf.io/7gefz>.
- Wilson, Robert C., and Anne G.E. Collins. 2019. Ten simple rules for the computational modeling of behavioral data. *eLife* 8: e49547. <https://doi.org/10.7554/eLife.49547>.
- Woźniak, Mateusz. 2018. “I” and “me”: The self in the context of consciousness. *Frontiers in Psychology* 9: 1656. <https://doi.org/10.3389/fpsyg.2018.01656>.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.