Mark Solms
and Karl Friston

How and Why Consciousness Arises

Some Considerations from Physics and Physiology

Abstract: *We offer a scientific approach to the philosophical 'hard problem' of consciousness, as formulated by David Chalmers in this journal. Our treatment is based upon two recent insights concerning (1) the endogenous nature of consciousness and (2) the minimal thermodynamic conditions for being alive. We suggest that a combination of these insights specifies sufficient conditions for attributing feeling to being.*

Keywords: consciousness; hard problem; affect; free energy; active inference; Markov blanket.

1. Introduction

The 'hard problem' of consciousness asks: 'How can we explain why there is something it is like to entertain a mental image or experience an emotion? It is widely agreed that experience arises from a physical basis, but we have no good explanation of how and why it so arises' (Chalmers, 1995). Here, we consider this question from a scientific perspective, but preface our answers with some disclaimers and definitions.

First, this is a *preliminary communication* in which we speak to the problem in broad outline. The topics we must discuss trench on many

Correspondence:
Email: mark.solms@uct.ac.za

Journal of Consciousness Studies, **25**, No. 5–6, 2018, pp. 202–38

specialist fields. Therefore, a comprehensive exegesis of our proposal — doing proper justice to the multiple literatures — requires a longer treatment than can be provided in a journal article. Second, the hard problem of consciousness, as formulated by Chalmers, is a *meta-physical* problem. To the extent that this fact implies that it cannot be 'solved' scientifically, we concede that we can only provide a scientific *response* to the problem as quoted above. Chalmers' hard problem rests upon Nagel's earlier claim that consciousness has an essential 'something-it-is-like-ness': 'an organism has conscious mental states if and only if there is something that it is like to *be* that organism — something it is like *for* the organism' (Nagel, 1974). Accordingly, we aim to sketch (in broad outline) a straightforward scientific answer to the question: *why is there something it is like to be an organism, for the organism, and how does this something-it-is-like-ness come about?* Nagel states that 'if we acknowledge that a physical theory of mind must account for the subjective character of experience, we must admit that no presently available conception gives us a clue about how this could be done' (*ibid.*). We hope to provide such a clue. But — and this is our last disclaimer — our physical theory is expressed in *functional* terms, which opens another philosophical can of worms that we would like to pre-empt, if we can, by defining what we mean by 'function'.[1]

After stating that 'a physical theory of mind must account for the subjective character of experience', Nagel adds: 'it seems unlikely that any physical theory of mind can be contemplated until more thought has been given to the general problem of subjective and objective' (*ibid.*). We agree. We use the terms 'subjective' and 'objective' to refer to *observational perspectives*. The subjective perspective upon the organism realizes the *being* of the organism. (In this paper, we will refer to this perspective as 'interoceptive', and we will explain why this perspective is *necessary* for the organism.) The objective perspective upon the very same organism realizes the *body* of the organism. (We will refer to this perspective as 'exteroceptive'.) We take the

[1] One of our referees notes that, to the extent that our argument rests on a type of functionalism (a functional role defined by variational free energy), we cannot solve the hard problem — because functionalism is a prime target of the hard problem. We take the view (see below) that affective functions need not suffer the same fate as cognitive ones in this respect. 'Something-it-is-like-ness' is less extrinsic to the function of feeling than it is to the function of visual perception — which has been the traditional model example of consciousness.

view (the metaphysical position) that neither of these observable realizations can be explained away by the other. In other words, data about an organism that are derived from both interoceptive and exteroceptive perspectives must be reducible to one and the same set of explanations. Therefore, biological explanations (as opposed to descriptions) are best formulated in neither interoceptive nor exteroceptive phenomenal terms, but rather as *abstractions*.

The explanations that we propose in this paper may be described as 'functional' for the following reason, using the example of memory: both the (subjective) re-experiencing of an event and the (objective) behaviour of the neuronal constellation encoding that event are explicable by laws governing an (abstracted) *function* called 'memory'. Memory itself is not subjective (mental) or objective (physical); it is both — and the abstracted laws governing memory are used to *explain* both of its realizations. (Ribot's and Miller's laws may be cited as examples.)

This seems perfectly straightforward. But here we are tasked with abstracting the laws governing a different function, not memory but 'consciousness', which introduces a special problem that does not apply to other biological functions. The nub of the problem is this: taking an interoceptive perspective upon the other functions does not always imply that there is 'something it is like' to be them. Thus, for example, the memory traces invoked above are not always experienced interoceptively, even when they exert exteroceptively observable effects. Procedural memory is the paradigmatic example. The something-it-is-like-ness of consciousness, by contrast, is intrinsic to what it *is*. This specific characteristic of subjectivity (which is sometimes present and sometimes not), and its particular effects upon the organism, are what we need to explain. The lack of any plausible scientific explanation of how and why *this* characteristic and *these* effects of subjectivity arise has led many philosophers to conclude that consciousness is a mere epiphenomenon of physical brain processes.

In our view, consciousness is not merely the subjective observational perspective upon the actual functions of organisms; consciousness realizes a biological force with definite causal powers of its own. To be clear, we want to explain both the mental and the physical manifestations of consciousness, by discerning the underlying functional laws that explain them both. It is in this (dual-aspect monist) sense that we aim to provide a physical theory of what Nagel calls the 'subjective character' of experience.

We will argue that the underlying function of consciousness is free energy minimization, and — in accordance with the above framework — we will argue that this function is realized in dual aspects: subjectively it is *felt* as affect (which enables feeling of perceptions and cognitions) and objectively it is *seen* as centrencephalic arousal (which enables selective modulation of postsynaptic gain).

2. Two Foundational Insights

Two recent scientific insights (Friston, 2013; Solms, 2013), when combined, yield a straightforward response to Chalmers' question cited above. The first of these insights (Solms, 2013; 2017a) is that the primary function of consciousness is not to register states of the external world but rather to register the internal states of the experiencing subject. This view is not based in philosophy but on the anatomical and physiological evidence regimented below (Section 6), which suggests that consciousness is quintessentially *interoceptive*. The basic argument goes as follows: conscious qualia arise primarily not from exteroceptive perception (i.e. vision, hearing, somatic sensation, taste, and smell), and still less from reflective awareness of such representations, but rather from the endogenous *arousal* processes that activate them.[2] Exteroceptive representations are intrinsically unconscious things (Kihlstrom, 1996): they do not inherently possess 'something-it-is-like-ness'. They only acquire conscious quality when they are, as Chalmers puts it, 'entertained' by the subject; i.e. when they are selectively activated by a more fundamental form of consciousness. In short, mental images can only be experienced *by* a conscious subject and they are in fact states *of* the conscious subject. The *arousal* processes that produce what is conventionally called 'wakefulness', in our view, therefore, *constitute the experiencing subject* — they are *consciousness itself*. To be clear: the term 'consciousness itself', neurophysiologically speaking, means the arousal functions of the centrencephalic structures that sustain wakefulness and behavioural responsivity *which in turn supply* the

[2] Please note, the arousal processes themselves are endogenous, not interoceptive. It might therefore be better to say that consciousness quintessentially arises *in response to* interoceptive events. Interoceptive events are visceral events. This issue is developed in the text below. In view of the centrality to our formulation of 'arousal', we operationalize the term (and show its relation to entropy and information) in an Appendix. At this point in our paper, we are referring to brain arousal in *vertebrates*. In Sections 3 and 4 we consider the elemental function performed by arousal in more formal terms.

conscious character of some higher cortical functions. The latter perceptual and cognitive functions (which are otherwise typically unconscious) derive their consciousness absolutely from the centrencephalic region.

Crucially, the evidence assembled in Section 6 shows that the arousal processes in question produce more than a merely quantitative 'level' of consciousness — they embody qualitative 'content' too: it *feels like something* to be awake. The processes of arousal, which we believe bring the experiencing subject into being, possess qualia of their own. These qualia are called *affects*; namely, *feelings* like hunger, lust, and surprise (Panksepp, 1998), which can exist independently of perceptual or cognitive representations — and arise even in the absence of the cortex (Merker, 2007).[3] Affect may be defined as the means by which organisms register their own states (Damasio, 2010); in other words, affect registers the state of the subject, not (primarily) of objects. Pfaff (2006) provides an analogy from physics: '[Affect] can be viewed as a vector. Arousal level determines the amplitude (length of the vector), while the exact feeling and object determine the angle of the vector' (p. 2).

This conception of consciousness has implications for the hard problem, since it refocuses the very nature of the thing we are trying to explain (i.e. the thing denoted by the word 'consciousness'). For example, if consciousness itself is feeling, then attended mental images only become conscious when they are *felt* — when they are, as it were, palpated by feeling — which is what usually happens when they are salient. The conventional focus on the cortical manifestation of conscious images (traditionally, visual ones) — as a model example of the neural correlates of consciousness (Crick, 1994) — may therefore have led us astray, because their consciousness is derivative. We would have done better to focus on the brain mechanisms of more basic forms of consciousness, like the feelings associated with thermoregulation, say, or hunger — which usually do not entail exteroceptive

[3] This suggests a three-level taxonomy of consciousness: (1) arousal, or consciousness itself, which we call 'affective consciousness'; (2) arousal of representations — that is attentional activation — which we call 'perceptual consciousness'; (3) reflective re-representation of affective and perceptual consciousness, which we call 'cognitive consciousness' (Solms and Panksepp, 2012). With reference to our introductory definitions: affective consciousness is *interoceptive*, perceptual consciousness is *exteroceptive*, and cognitive consciousness is *abstracted* from them both. (However, see previous footnote.)

images at all. Feelings *necessarily* entail something-it-is-like-ness. The function of affect is barely distinguishable from its feeling.

To be clear: we are not saying that perceptual and cognitive consciousness is really affective consciousness. The 'channel' functions of the cortex *stabilize* affective 'states' (Mesulam, 2000); by extending feeling onto perceptual images and thoughts (*cf*. Damasio, 2010), they transform raw feeling into a different kind of consciousness.[4] (We address this issue in more detail below, in Section 6, in relation to reconsolidation.)

If consciousness is a subjective state, then the hard problem raises the question: how and why do subjects become conscious? In other words, how and why do some subjects — but not others — feel like something (*cf.* Nagel, 1974)? Please note, in view of the affective nature of consciousness itself (the evidence for which is assembled below), it proves useful to recast 'something-it-is-like-ness' as 'feeling'. We said in our introductory remarks that the subjective perspective (being something) can be attributed to anything (like a computer or a zombie), but what are the minimal conditions for attributing feeling to being? An approach to this question might usefully start with the minimal conditions for being alive (or staying alive); since it is generally assumed that the subjectivity of non-living things is devoid of qualia.

The second of our two insights (Friston, 2013) concerns these minimal conditions. A fundamental property of living things (i.e. biological self-organizing systems) is *their tendency to resist the second law of thermodynamics*. This functional property emerges *naturally* within any ergodic[5] random dynamical system that possesses a *Markov blanket*.[6] On this view, their negentropic tendencies (i.e.

---

4 Freud (1895/1950; 1896/1950) introduced a useful concept here, called 'cathexis', to denote a 'free' endogenous energy that is 'bound' by cortical processing. (*Cf.* the 'free energy' principle discussed below; see also Carhart-Harris and Friston, 2010.)

5 'Ergodicity' is a statistical property, whereby the average of any measurable function of a random dynamical system *converges* over a sufficient period of time. In short, dynamical systems that possess measurable characteristics over periods of time must be (nearly) ergodic.

6 A 'Markov blanket' induces a statistical partitioning of internal and external states, and *hides* the latter from the former, so that the external states can only be registered vicariously (through the blanket) by the internal states of a system. The Markov blanket itself consists in two sets ('sensory' and 'active' states) which influence each other in a circular fashion: external states cause sensory states which influence — but are not influenced by — internal states, while internal states cause active states which influence

persistence of such systems in the face of exogenous perturbations) can always be cast as a Bayesian process of *active inference*. Active inference arises inevitably from the properties of a Markov blanket, because the internal states of such systems will necessarily model — and accordingly act upon — their external states to preserve the functional and structural integrity of the system. *This leads to homeostasis and a simple form of autopoiesis.* That is, biological systems must behave as if they had a generative model of the world that forges predictions about the sensory consequences of action (Conant and Ashby, 1970; Seth, 2015). This enables them to place an upper bound on the dispersion of their sensory states. In statistical terms, this dispersion or *entropy* is the expected self-information or *surprise* that can be quantified in terms of variational *free energy* (Friston, 2010).[7] In effect, self-organization entails the use of the sensory states (of a Markov blanket) to *infer* external states (that surround the Markov blanket).

Through acting upon the world — and sampling new sensory states — any biological self-organizing system automatically generates new predictions concerning the hidden causes of its sensory samples, through an iterative process that can be construed as *hypothesis testing* (Gregory, 1980; Seth, 2015). The link between the thermodynamic imperatives to minimize entropy or surprise and hypothesis testing rests upon the following fact: if we treat sensory samples as a data, then free energy becomes the negative logarithm of Bayesian model

— but are not influenced by — external states. The close similarity between the conditions just described and those of both bodies and minds is not accidental. See also footnote 13 regarding the definition of 'external' states.

7 On the relationship between (information theoretic) variational free energy and thermodynamic free energy — to clarify the physical realization of our 'functional' explanation — see Tozzi, Zare and Benasich (2016): 'Minimizing variational free-energy necessarily entails a metabolically efficient encoding that is consistent with the principles of minimum redundancy and maximum information transfer (Picard and Friston, 2014). Maximizing mutual information and minimizing metabolic costs are two sides of the same coin; by decomposing variational free energy into accuracy and complexity, one can derive the principle of maximum mutual information as a special case of maximizing accuracy, while minimizing complexity translates into minimizing metabolic costs (Friston et al., 2015). Thus, the basic form of Friston's free-energy principle supports the idea that the energetic levels of spontaneous brain activity, which are lower when compared with evoked activity, allow the CNS to obtain two apparent contradictory achievements: to minimize as much as possible the metabolic costs, and to the largest extent possible, maximize mutual information.' Please note: the functional principle *explains* the physical behaviour of neurons. On this important point, see also our Appendix.

evidence; see below and Friston (2013). In short, any self-organizing system *must* minimize its own free energy and therefore *must* engage in active inference; in virtue of maximizing model evidence. This is appealing because it identifies self-organization (Clark, 2017; Haken, 1983; Kauffman, 1993; Kelso, 1995; Maturana and Varela, 1980) with self-evidencing, i.e. garnering evidence for [my] models of — and beliefs about — the lived world and own body (Hohwy, 2016).

Thus, the existence of a Markov blanket, in our view, provides the minimal conditions not only for maintaining life but also for *selfhood* (for self-organizing existence within a world that can be separated from the self). This maintenance of selfhood generates a form of work (in the sense of statistical mechanics) that conforms to goal-directed notions of *intentionality*. This conclusion is based on formal mathematical arguments that we will now try to unpack.

3. Link to the Hard Problem

The above insights lead us to the following response to the hard problem. Since biological self-organizing systems are intrinsically intentional, because they must engage in active inference in order to avoid 'surprising' states (in order to maintain selfhood and stay alive), the internal state of such systems (their being) entails *existential value*. The same cannot be said in any simple way for the cognitive functions that have been the conventional focus of consciousness studies, such as visual perception. This, in our view, has substantial implications for the hard problem. Selfhood and intentionality, which are inherently linked to value (i.e. to the principle that survival is 'good'; see below), arise naturally within the parameters we have just described. These parameters underpin homeostasis, which is the most basic mechanism through which organisms stay alive. Later we will show that the brainstem nuclei that realize homeostasis, and thereby generate selfhood and intentionality (in vertebrate organisms) — properties which turn out to be deeply bound up with consciousness — are inextricable from the brainstem mechanisms for arousal. That is, we will show that *homeostatic regulation and the arousal of consciousness are effected by the same part of the brain*. But first we must formalize the relationship between homeostasis and affect.

Technically, on the Bayesian view, the 'value' mentioned above corresponds to prior 'beliefs'[8] or 'preferences' entailed by the generative model. In other words, a *preferred state* is a valuable state that a system — like you and me — expects to occupy (e.g. my core body temperature must remain between 36.5°C and 37.5°C). The presence of existential values or prior preferences underwrites the continued existence of the system (i.e. my survival) that is contingent upon *ongoing auto-assessment of free energy*, which, in turn, generates intentional acts (e.g. moving to cooler surroundings). The free energy of the system is identical with the demand for (anti-entropic) goal-directed work (via the resolution of surprise). That is, it is identical with what neurobiologists call 'drive' (Pfaff, 1999). This obligatory value relation (of selfhood to intentionality) may be described as proto-mental. All that is required, in order to render it so, is to view this relation from the subjective perspective of the self-organizing, self-evidencing system, which is justified precisely by its selfhood. In other words, it *must* be viewed interoceptively.

Viewing the relation from this perspective, we propose that the measurement — by a self-organizing system — of its own free energy, considered subjectively, gives rise to what we call affect. In other words, affect is this inherently evaluative aspect of biological being, predicated on the belief that survival is valuable (which belief, of course, is the value system that underpins all life forms). Please note, however: so far, we have only considered how the obligatory auto-measurement of free energy (which sustains selfhood and generates intentionality), which we are equating with affect, is necessary for *life*; we have not yet explained specifically how this function gives rise to *feeling*. That is why we call the mechanism we have described 'proto-mental'. In the next section, we will describe how feeling arises within the finer workings of this mechanism.

To preface our explanation, we propose that deviation away from each homeostatic settling point (away from the preferred state for that parameter) is registered as a *negative* affect, and returning towards a settling point is registered as the particular *positive* affect for that parameter. The settling point itself ('satiation') *resolves* the affect, which implies that affect is no longer generated at this (ideal) point.

8 Beliefs in this technical sense are taken to be probability distributions whose parameters or sufficient statistics again correspond to physical brain states. These will be identified later.

Affectivity in general, therefore, both negative and positive, registers continued demand for work (i.e. continued need). Affect, which becomes conscious in the manner we shall now describe, may be construed as an endogenous alarm mechanism that registers the existence and directionality of deviations from preferred (valued and expected) states.

Perhaps the easiest way to understand how this internal measurement of free energy entails feeling — from the viewpoint of the system — is to consider predictive coding as a (neurobiologically plausible) instance of self-evidencing.

4. Self-Evidencing, Precision, and Affect

Predictive coding formulates free energy or surprise in terms of precision weighted prediction errors. A prediction error (e) here is the difference between a sensation (φ) produced by some action (M) and the sensation predicted by a generative model $\psi(Q)$. Here, Q stands for internal expectations about — or representations of — hidden external states and $\psi(Q)$ is the prediction of sensory inputs that would have been encountered given those external states, under the generative model. Under some simplifying assumptions,[9] we can now associate free energy (F) with the amount of prediction error weighted by its precision (ω). Precision corresponds to the reliability, or inverse variance, of sensory fluctuations (in various modalities) and is an important aspect of inference; namely, the *representation of uncertainty*. One can think of precision as the confidence placed in the (predicted) consequences of an action or in a source of sensory evidence. Heuristically, one can regard prediction errors as newsworthy information and the precision of that information as its reliability. We will see below that prediction errors have a much greater effect on internal expectations or representations when they are afforded more precision.

[9] For clarity, we effectively reduce the free energy to the likelihood of a Gaussian distribution. In fuller treatments, one would consider hierarchical generative models (with precisions at each level — see Figure 2) and accommodate conditional uncertainty about external states. Furthermore, we have lumped all sensory prediction errors together — including exteroceptive, proprioceptive, and interoceptive modalities. (See footnote 13.) Please note: our usage of the term 'proprioceptive' denotes 'kinaesthetic' (we use 'proprioceptive' simply for alliterative harmony with 'exteroceptive' and 'interoceptive').

With these quantities in place, one can describe any self-organizing (i.e. self-evidencing) system with the following dynamics:

$$\frac{\partial}{\partial t}M = -\frac{\partial F}{\partial M} = -\frac{\partial F}{\partial e}\frac{\partial e}{\partial M} = \frac{\partial \varphi}{\partial M}\cdot\omega\cdot e \qquad \text{(1a: action)}$$

$$\frac{\partial}{\partial t}Q = -\frac{\partial F}{\partial Q} = -\frac{\partial F}{\partial e}\frac{\partial e}{\partial Q} = -\frac{\partial \psi}{\partial Q}\cdot\omega\cdot e \qquad \text{(1b: perception)}$$

$$\frac{\partial}{\partial t}\omega = -\frac{\partial F}{\partial \omega} = \frac{1}{2}\cdot(\omega^{-1} - e\cdot e) \qquad \text{(1c: affect)}$$

Where free energy and prediction error are:

$$F = \frac{1}{2}\cdot(e\cdot\omega\cdot e - \log(\omega)) \qquad (2)$$

$$e = \varphi(M) - \psi(Q)$$

We are mindful of the fact that we are following in the footsteps of the Helmholtz school of medicine, whose members swore an oath in 1842 to the effect that 'no forces other than the common physical chemical ones are at work in the organism' (Du Bois-Reymond, 1918). We therefore state our response to the hard problem using the quantities φ, ψ, ω, M, and Q, with an historical nod to a pupil of that school who first used them to 'represent psychical processes as quantitatively determinate states of specifiable material particles' (Freud, 1895/1950; 1896/1950).

In our formulation, the *ideal* adaptive state of the organism — where negentropic demand is met by optimal predictions — describes the 'Nirvana principle' (*cf.* Freud, 1920). This Nirvana corresponds to a curious world where there are no prediction errors and the expected free energy is absolutely minimized, which — by construction — corresponds to a self-state with no uncertainty or entropy. Under these conditions, the precision becomes infinitely high. This is easy to show by expressing the expected free energy or surprise in terms of ω (where $E[\cdot]$ denotes expectation or averaging).[10]

[10] Note that, in the current formulation, expected free energy decreases when precision increases. It is a simple matter to deal with inverse precision (i.e. the variance of — or uncertainty about — particular prediction errors), which might be more in line with Freud's original formulation of ω. Expected free energy and ω go hand-in-hand. However, we elected to deal with precision for consistency with current treatments of active inference in emotion and psychopathology (Clark, 2013).

$$F \approx -logP(\varphi(M)) \Rightarrow \quad (3)$$

$$E[F] \approx E[-logP(\varphi)] = H[P(\varphi)] = -\tfrac{1}{2} \cdot log(\omega)$$

The first expression says that free energy is (approximately) the logarithm of the probability of encountering some actively authored sensations. The second (approximate) equality says that the expected free energy decreases in proportion to log precision.

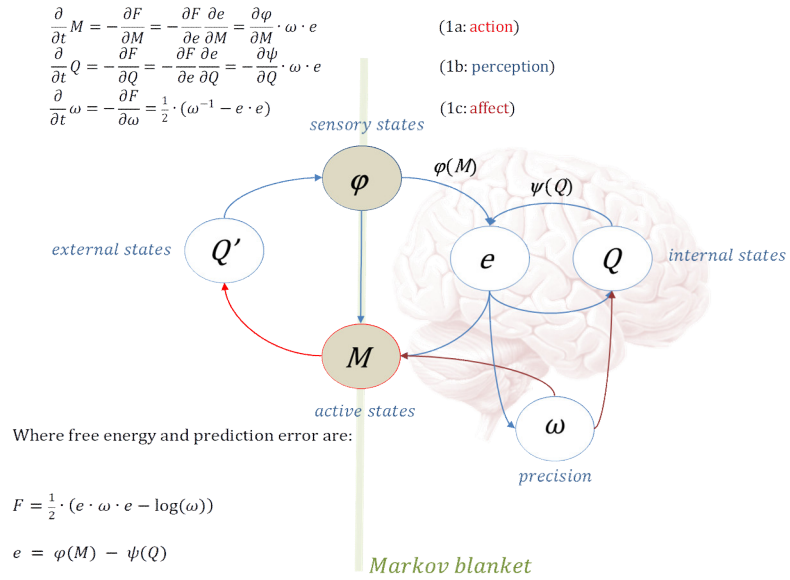

Figure 1. A diagrammatic representation of a self-organizing system's dynamics. From equation 1, it is evident that there are three ways to reduce free energy or prediction error. First, one can act to change sensations, so they match predictions (i.e. action). Second, one can change internal representations to produce a better prediction (i.e. perception). Finally, one can adjust the precision to optimally match the amplitude of prediction errors. It is this final optimization process — mandated by free energy minimization — that we associate with consciousness *per se* (see footnote 3) and the evaluation of free energy that underpins experience. In short, consciousness (as opposed to mere homeostasis) is constituted by inferring *changes in expected free energy* or, more simply, uncertainty about the experienced world and body. Inferred precision is *felt uncertainty*. Thus, precision increases when things promise to turn out as expected and it decreases when uncertainty prevails.

To appreciate the plausibility of this interpretation, it is worthwhile considering the computational and neurobiological architecture of predictive coding. In predictive coding schemes, prediction errors are formed by comparing ascending sensory evidence with descending predictions based upon (hierarchically disposed) expectations or representations (Mumford, 1992). The ensuing prediction errors are then passed up the hierarchy to adjust expectations (second equation above) in proportion to their precision or reliability. This sort of scheme — with recurrent exchanges of (ascending) prediction errors and (descending) predictions — closely resembles empirical message-passing in cortical and subcortical hierarchies (Adams, Shipp and Friston, 2013; Bastos *et al.*, 2012; Shipp, 2016). In this context, action reduces to proprioceptive (motor; see footnote 9) and interoceptive (autonomic; see footnote 13) reflexes that are driven by descending predictions from the brain's (hierarchical) generative model.

The crucial aspect of this formulation is the role of precision. From the above equations, it is clear that precision controls the influence of prediction errors on action and perception. Physiologically, precision is usually associated with the *postsynaptic gain* of cortical neuronal populations (e.g. superficial pyramidal cells or von Economo cells) reporting prediction errors (Brown *et al.*, 2013; Feldman and Friston, 2010; Friston, Kilner and Harrison, 2006). In this sense, precision can be associated — through free energy minimization — with *selective arousal* or *attentional selection* (Clark, 2013; Hohwy, 2013; Kanai *et al.*, 2015). This will be an important theme in what follows and ties our formulation of consciousness itself to the same sorts of mechanisms that mediate not only arousal but also attention (and sensory attenuation) and that are intimately involved in setting levels of consciousness during sleep and other mental states (Hobson, 2009). Furthermore, it is precisely this neuromodulatory synaptic mechanism that is targeted by psychotropic and psychedelic — i.e. consciousness altering — drugs (Nour and Carhart-Harris, 2017). This line of reasoning was initiated by Fotopoulou (2013), who has followed it empirically in relation to interoceptive sensitivity (Ainley *et al.*, 2016; Crucianelli *et al.*, 2017; Fotopoulou and Tsakiris, 2017a,b) and the social modulation of pain (Krahe *et al.*, 2013; Decety and Fotopoulou, 2015; Paloyelis *et al.*, 2016; von Mohr and Fotopoulou, 2017).

Conceptually, precision is a key determinant of free energy minimization and the enabling — or activation — of prediction errors. In other words, *precision determines which prediction errors are selected* and, ultimately, how we represent the world and our actions

upon it. In this sense, precision plays the role of *Maxwell's demon*[11] — selecting the passage of molecules (i.e. sensory signals) to confound the second law. On our view, *consciousness is nothing more or less than the activity of Maxwell's demon* (i.e. the optimization of precision with respect to free energy) — as opposed to the passage of molecules that are enabled (i.e. the perceptual sequelae of message-passing in cortical hierarchies). This optimization manifests in many guises. In the exteroceptive domain, it manifests as sensory attention and attenuation associated with the increase and decrease of sensory precision (Brown *et al.*, 2013; Feldman and Friston, 2010; Frith, Blakemore and Wolpert, 2000). In the proprioceptive domain, it corresponds to the selection and realization of motoric predictions of the sort associated with action and goal selection (Cisek and Kalaska, 2010; Frank, 2005; Friston *et al.*, 2014; 2012; Moustafa, Sherman and Frank, 2008). In the interoceptive domain, it literally determines 'gut feelings', i.e. the best explanation for interoceptive signals that have been enabled or selected (Hohwy, 2013; Seth, 2013). Note that this construction calls on the notion of *activating* expectations or representations in the sense that — in the absence of precision — prediction errors will fail to induce any perceptual synthesis, behaviour, or neuronal response. In other words, without precision, prediction errors would be sequestered at the point of their formation in the sensory epithelia.

Physiologically, these sorts of states are encountered every day; for example, during sleep (Hobson, 2009; Hobson and Friston, 2014). Furthermore, this activation implicates neuromodulatory systems (see Section 6 below) and their mediation through fast synchronized neuronal dynamics — that are affected in altered states of consciousness (Ferrarelli and Tononi, 2011; Lisman and Buzsaki, 2008; Uhlhaas and Singer, 2010). This formulation also has some construct validity in relation to neuronal versions of global workspace theories of consciousness (Dehaene and Changeux, 2011), that themselves can be traced back to the variational principles that underlie free energy minimization (Friston, Breakspear and Deco, 2012).

[11] Maxwell's demon is a thought experiment created by James Clerk Maxwell to suggest how the second law of thermodynamics might be violated: in brief, a demon controls a small door between two chambers of gas. As gas molecules approach, the demon opens and shuts the door, so that fast molecules pass to the other chamber, while slow molecules remain in the first, thus decreasing entropy.

The physiology of precision engineered hierarchical inference is as complex as the myriad neuromodulatory mechanisms mediating the postsynaptic gain of prediction error reporting pyramidal cells. It is useful to appreciate that every prediction error neuron (or neuronal population) is equipped with the postsynaptic gain and an implicit representation of precision. This affords explanatory latitude that may offer a useful framework to understand the distinction between exteroceptive and interoceptive attention (Kanai *et al.*, 2015) or, in the context of the present discussion, the relationship between attention and consciousness. For example, Koch and Tsuchiya (2012) and van Boxtel, Tsuchiya and Koch (2010) show that 'selective attention and visual consciousness have opposite effects: paying attention to [a] grating decreases the duration of its afterimage, whereas consciously seeing the grating increases the afterimage duration'. These sorts of effects are accommodated in hierarchical predictive coding by judicious tuning of the precision at various levels of the visual hierarchy; where (directed) attention is usually associated with increasing the precision of sensory evidence garnered in lower hierarchical levels. Conversely, 'seeing' normally entails precise or confident beliefs in perceptual posteriors at (higher) hierarchical levels — that best explain the sensorium. Interestingly, the precision at different levels usually operates in opposition in a highly context sensitive fashion (Dayan, 2012). On some accounts, getting the hierarchical balance of precision wrong — particularly in interoceptive inference — can have devastating effects on minimal selfhood and theory of mind in a neurodevelopmental setting (Fotopoulou and Tsakiris, 2017a; Quattrocki and Friston, 2014).

Our particular focus here is on the fundamental experience that is informed by ascending interoceptive signals (i.e. prediction errors) and how this process of *interoceptive inference* (Barrett and Simmons, 2015; Palmer, Seth and Hohwy, 2015; Seth, 2014; 2013) gives rise to consciousness through the optimization of ω. This self-state (i.e. the precision of interoceptive signals) is what we call *felt uncertainty* or affect. Affect, thus defined, is an *existential imperative*; it is the vehicle by which the system monitors and thereby maintains its functional and structural integrity. *The inherently qualitative-evaluative nature of this self-assessment explains 'how and why' it feels like something within the system, for the system.* Again, we must emphasize: this does not apply to the inherently unconscious perceptual/cognitive functions that gave rise to the hard problem. Specifically, expected increase in free energy just *is* 'bad' from the

(interoceptive) perspective of a self-organizing system — indeed it is an existential crisis — while expected decrease just *is* 'good'. These changes are recognized (i.e. interoceptively inferred [= felt]) in terms of changes in uncertainty and concomitant adjustments of ω.

Such self-valuation does not arise in non-living systems, which are spared the obligation of minimizing their free energy and therefore of engaging in active inference. In this regard, it is of capital importance to recognize that adjustment of ω is most imperative in relation to the homeostatic affects that broadcast *vital needs* (i.e. interoceptive determinants of Q). Imagine the consequences of adjusting the settling point — the preferred state — of core body temperature. (This applies only slightly less to 'emotional affects', described below.) This yields an important distinction between the exteroceptive and interoceptive sensory modalities of φ (another important theme in what follows). Homeostatic perturbations can only be managed to a limited extent by autonomic reflexes. (For example, as core body temperature rises, perspiration reaches the limit of its capacity to cool the organism.) At this limit, the biological imperatives encoded in preferred states absolutely demand free-energy-reducing, surprise-avoiding, volitional action. Life is difficult: most vital needs can only be managed via interaction with the *external* world.

The affective value implicit in ω must be an inherent property of any self-organizing system that proactively resists the second law. Precision optimization determines the extent to which this value will be *felt* (i.e. expressed via an enabling of belief updating). Precision entails selectivity and it thereby underwrites *choice*. To be clear: it is easy to envision an organism (or machine) in which precision values are set in such a way that the system's responses to prediction error are *automatized*. Indeed, large swathes of the human nervous system are organized in this way (see below). However, the capacity for *experiencing* changes in expected uncertainty (as described above) adds enormous adaptive value. This capacity is especially useful in the case of ambulant organisms — as opposed to plants, for example.

Below we will describe the relation of this capacity to neural *plasticity*. It is difficult to conceive of a self-organizing system adapting flexibly to inherently unpredictable environments in the absence of some such capacity. *This, in our view, is how and why consciousness arises*. We readily concede that conscious feeling is not the *only* way that changes in expected uncertainty (which encode the existential values described above, and thereby maintain selfhood and drive intentionality) can *conceivably* be registered subjectively and

proactively in the here-and-now by a self-organizing system, but it surely entails the *sufficient* conditions (*cf.* 'philosophical zombies'). It is therefore not surprising that feeling, once it evolved by natural selection, was conserved. In this respect, consciousness is no different from any other adaptive biological function. Ambulation, for example, does not *necessarily* require legs.

We do not mean to imply by our proposals that every (ergodic) random dynamical system that possesses a Markov blanket will experience feelings. We are claiming only that *these are the fundamental mechanisms whereby feeling arises*, and *consciousness is what such existential imperatives feel like in the vertebrate nervous system*. That is why we used the term 'proto-mental' rather than 'mental' for the basic mechanism we describe. What we describe is an elemental form of a self-maintaining mechanism that takes more complex forms in more complex biological systems (like vertebrates). Thus, while, on our view, the subjective states of protozoa must entail affect, as conceptualized here; we do not claim that they feel like we do. What we *are* claiming is that if a method could be devised by means of which any life form could 'declare' its existential and intentional states, as defined above, the output would be a functional of what we call affect. Such declarations in organisms equipped with the *precision optimization* mechanism just described, regarding the reliability of their sensory fluctuations (i.e. declarations that changes in expected uncertainty *feel* good, bad, or indifferent to and for the organism), would still be constrained by the epistemological problem of other minds; but this constraint applies equally to the self-report of *every* other being (including humans, whose declared feelings are discounted by behaviourists; see Panksepp *et al.*, 2016). Moreover, the properties of a Markov blanket *explain* this radical subjectivity of mental states (see footnote 6).

It is noteworthy that qualitative fluctuations in felt affect (i.e. ω) arise continuously from periodic comparisons between the sensory states that were predicted (based upon a generative model of the viscera and the world — $\psi(Q)$) and samples of the actual sensory states (φ). This recurrent assessment of sensory states only gives rise to changes in subjective quality (i.e. precision and feeling) when the amplitude of prediction errors *changes* — signalling a change in uncertainty about the state of affairs and, in particular, the consequences of action (M).

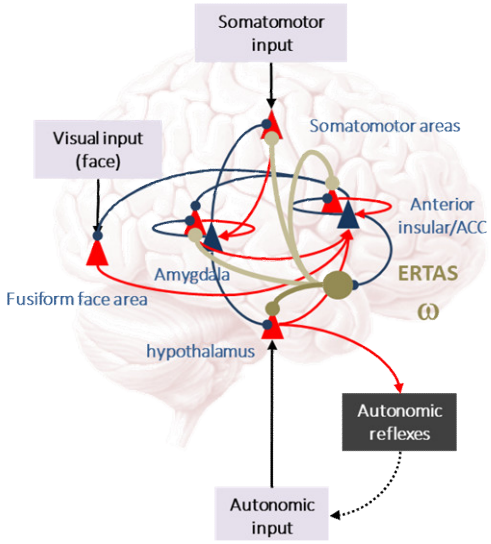

Figure 2. A simplified neural architecture underlying the (hierarchical) predictive coding of exteroceptive (visual), proprioceptive (somatomotor), and interoceptive (autonomic) signals. The anatomical designations, although plausible, are used to simply illustrate how predictive coding can be mapped onto neuronal systems. Red triangles correspond to neuronal populations (e.g. superficial pyramidal cells) encoding prediction error, while blue triangles represent populations (e.g. deep pyramidal cells) encoding expectations. These provide descending predictions to prediction error populations in lower hierarchical levels (blue connections). The prediction error populations then reciprocate ascending prediction errors to adjust the expectations (red connections). Arrows denote excitatory connections, while circles denote inhibitory effects (mediated by inhibitory interneurons). In this hypothetical example, which begins with exteroceptive events (as is currently conventional in consciousness studies), recurrent connections mediate innate (epigenetically specified) reflexes that elicit autonomic 'motor' (e.g. vasovagal) reflexes in response to appropriate somatosensory input. These reflexes depend upon high-level representations predicting both the somatosensory input and interoceptive consequences. The representations are activated by somatosensory prediction errors and send interoceptive predictions to the hypothalamic area — to elicit interoceptive prediction errors that are resolved in the visceral periphery by autonomic reflexes, to the limited extent that this is possible before somatomotor action is called for. Classical neuromodulators (in green) are shown to project to the hypothalamic area, to modulate the gain or precision of interoceptive prediction error units (this is affective processing). At the same time, the precision of exteroceptive and proprioceptive level prediction errors is adjusted to engage complementary attentional and motivational processing. (ERTAS: extended reticulothalamic activating system. ACC: anterior cingulate cortex).

In *exteroceptive* cases, the affect arises from (inferred) dispersive external states — registered as *not-self* states, although they have existential implications for the self — which persist in consciousness until the uncertainty is bound (i.e. precision is restored) by renewed predictive work. We surmise that two varieties of quality (*interoceptive* and *exteroceptive* precision) are registered differently by the system, as affective and perceptual consciousness respectively (see footnote 3). Whereas affect is an existential state, perception/cognition is an intentional one. Exteroceptive consciousness is *predictive work-in-progress.* This implies that it should wane in conditions of ambient monotony (as it does; see Riggs *et al.*, 1953). Please note: such work is always (ultimately) in the service of organismic needs.

The distinction between interoceptive and exteroceptive precision is thus central to our argument. Precision is not a single value; every sensation — and every hierarchical abstraction — must be equipped with a precision that has to be optimized. If brains are sympathetic organs of inference, assimilating exteroceptive, proprioceptive, and interoceptive data through prediction, then their respective precision is *about* something (*cf.* Brentano, 1874). Our position here is that interoceptive (and proprioceptive) precision are special in the sense that they oblige the organism to engage with the outside world and thereby determine active, embodied engagement with it. They are therefore inherently about selfhood and intentionality. It is in this sense that we associate interoceptive (and proprioceptive) precision with existential affect: see also Fotopoulou and Tsakiris (2017b), Seth (2013). Exteroceptive precision, although formally identical and closely related to affect (*cf.* 'sensory affects'; Panksepp, 1998), is normally conceptualized as *attention*, and proprioceptive precision as goal selection (or *motivation*).

5. A Comment on Dual-Aspect Monism

To clarify what we mean by sentient *being* and the associated intentionality of biological self-maintaining systems, an elaboration of our introductory remarks on the hard problem is called for. We believe that the mind/body problem, as typically formulated, is an artefact of observational perspective. Put simply, seeing oneself (exteroceptive perspective) and being oneself (interoceptive perspective) realize two aspects of the same thing: the 'physical' body (as seen) and 'mental' being (as felt) are dual aspects of a single entity. Experience accordingly does not 'arise from a physical basis'; rather, the physical (what

is seen — exteroception) and the mental (what is felt — interoception) are dual manifestations of unitary underlying processes. What is seen does not cause what is felt. Both have *hidden* causes. Consciousness (both exteroceptive and interoceptive) involves the quest for these unitary hidden causes, which must be inferred from the two sets (i.e. modalities) of data and *explain them both*. This is at the heart of inference to the best explanation that underlies free energy minimization and abductive inference we therefore engage in (Harman, 1965; Seth, 2015).

By analogy, casting the hard problem in perceptual-consciousness terms only, to clarify the difficulty (and thereby — for the sake of the analogy — equating vision with exteroception and hearing with interoception), we ask: how does lightning cause thunder? (*Cf.* Chalmers, 1996: 'How and why do neurophysiological activities produce the "experience of consciousness"?') The answer is: lightning does not cause or 'produce' thunder; both phenomena are caused by a not-directly-observable process that we infer from sensory data; namely, an *abstraction* called 'electrical discharge'. This unitary underlying explanation solves the 'hard problem' of the relation between lightning and thunder.

Neuropsychology requires an equivalent abstraction to explain the causal mechanism of consciousness, in both of its manifestations: exteroceptive and interoceptive. In our view, the abductive inference implied by minimizing (variational) free energy is this long-sought abstraction; namely, the analogous process of inferring the causes of lightning and thunder (i.e. the causes of our experienced sensations and feelings).

In terms of our Markov blankets: the sensory states *translate* external states, which can only be registered vicariously. Thus, both the exteroceptive (perceptual physical) and interoceptive (affective mental) states are registered subjectively *from the viewpoint of the system*. Moreover, the variational free energy itself (and its constituent precisions) is only *experienced* within the system when it is subjectively conceived; the experiences themselves cannot be observed from without, objectively. The qualitative value of variational free energy is therefore contingent upon *selfhood*. (This accounts for the philosophical problem of other minds, mentioned above.) This formulation is consistent with our argument that consciousness itself is affective, even when transformed (stabilized) in perceptual/cognitive consciousness.

6. Anatomical-Physiological Realization

In the scientific response to the hard problem on offer, formulated above in formal terms, we link the minimization of free energy with a demand for active inference. In this section, we locate these abstractions in vertebrate anatomy and physiology, and thereby demonstrate their explanatory power, using the mammalian brain as a model example.

The pivotal abstraction in the formulation is variational free energy, the information theoretic homologue of 'uncertainty' in statistical mechanics. The neurophysiological realization of this quantity is precision or 'selective arousal'. The relationship between free energy and precision (inverse uncertainty; see Equation 3) evokes the adaptive function of *salience*: precision = salience = arousal (Pfaff, 2006).[12] Our example therefore revolves around the relationship between brain arousal processes and hierarchical predictive coding, which defines the relation between affect and cognition (see Figure 2 and Appendix).

To rehearse the basic scheme: the organism infers hidden *external states* in terms of expectations thereof (Q) by minimizing variational free energy, based on *sensory states* (φ). Crucially, *the states that are external to the Markov blanket — when it is embodied — include the viscera*. In other words, the external (to the nervous system) states of both the lived world and our own bodies have to be inferred on the basis of (exteroceptive and interoceptive) sensory evidence.[13] The resulting predictions (ψ) recruit *active states* (M) — i.e. fire proprioceptive and autonomic reflexes — to realize predicted and preferred external states. This process is enabled by selecting precise sensory evidence through optimizing *precision* (ω).

In anatomical-physiological terms, this translates as follows. The organism infers its visceral states ($Q\acute{\eta}$) in order to minimize deviations from homeostatic settling points (i.e. prior beliefs) in relation to its

12 Strictly speaking, in active inference, salience is the opportunity to resolve uncertainty through minimizing expected free energy; thereby increasing the precision or confidence in beliefs about action (Friston *et al.*, 2015).

13 The embryonic neural tube is of course formed from the ectoderm, through invagination of the neural plate. To distinguish the two grades of 'external' state, we will (from here onward) use the terms Q and $Q\acute{\eta}$, respectively, for the external states that are inferred exteroceptively and interoceptively. Distinctions between exteroceptive and interoceptive φ states are made in the text above, as are those between kinaesthetic and autonomic M states. Freud (1985/1950) associated autonomic M with secretory 'key neurons'. He likewise distinguished between extrinsic and intrinsic stimuli: Q and $Q\acute{\eta}$.

vital needs (core body temperature, glucose metabolism, hydration in relation to salt, etc.) which tend to take precedence over all other preferred states. This is effected principally by 'need detectors' (i.e. interoceptive prediction errors) in the medial hypothalamus but also by other body-monitoring structures such as the circumventricular organs, parabrachial nucleus, area postrema, and solitary nucleus. Deviations from predicted values, being salient, trigger forebrain arousal (or precision), via the extended reticulothalamic activating system (ERTAS).

A core claim is that ERTAS arousal is *felt* as (precision-optimized) affect, and moreover that cortical perceptual qualia are contingent upon this prior affective arousal which is extended onto perception ('I feel like this *about* that'). These claims are based on several findings. All of consciousness (both affective and perceptual/cognitive) is obliterated by relatively small ERTAS lesions (Parvizi and Damasio, 2001), whereas even relatively large cortical lesions obliterate only certain aspects of perceptual/cognitive consciousness (Penfield and Jasper, 1954). This implies a hierarchical dependency relation. The dependency applies even to those cortical regions that have been most closely linked with consciousness of affect, namely the insula (Craig, 2009) and prefrontal convexity (LeDoux and Brown, 2017). The fact that total ablation of insular cortex does not obliterate feeling (Damasio, Damasio and Tranel, 2012) and that — like prefrontal damage (Harlow, 1868) — it is actually associated with increased affectivity, shows that affective consciousness cannot be *generated* there. Indeed, even hydranencephalic children — born without a cortex — show a full range of basic emotions in response to adequate stimuli (Shewmon, Holmse and Byrne, 1999; Merker, 2007). Likewise, completely decorticate animals show increased, not decreased, affectivity (Huston and Borbely, 1974). The fallacy initiated by Moruzzi and Magoun (1949) to the effect that ERTAS arousal generates a quantitative 'level' of consciousness — while its qualitative 'contents' are generated in the cortex — is exposed also by the simple fact that psychotropic drugs like antidepressants and antipsychotics act upon the single-neurotransmitter systems sourced in the ERTAS, *viz.* the (serotonergic) raphe, (noradrenergic) locus coeruleus, and (dopaminergic) ventral tegmental area (VTA). Moreover, damage to these structures collectively does not affect mood only; it obliterates consciousness as a whole. In fact, the smallest possible coma-inducing lesion is located in the brainstem periaqueductal grey (PAG); electric stimulation of which produces *the most intense affective experiences*

known to man (Panksepp, 1998; Merker, 2007). These findings support the view that core brainstem arousal generates not only the global 'level' but also the affective 'state' of consciousness (Mesulam, 2000).

Returning, then, to our model example: prediction errors, which monitor the vital functions of the internal milieu ($Q\acute{\eta}$), activate ERTAS arousal when deviations from homeostatic setpoints exceed plausible thresholds (in relation to prior preferences). The qualitative vectors of arousal thus generated (and selected by ω modulation) in relation to the different need-parameters are felt as homeostatic affects — hunger, sleepiness, coldness, etc. — also known as 'drives' (Pfaff, 1999). These make (often urgent) demands upon the forebrain for action (M), which are channelled from the VTA through mesocortical-mesolimbic dopaminergic neurons projecting via the lateral hypothalamus into the limbic striatum (nucleus accumbens and amygdala) and other (mainly frontal) forebrain sites. This is an all-purpose *foraging* system that generates spontaneous behaviour and epistemophilic emotions like curiosity, optimism, and enthusiasm, and brings the animal into contact with need-satisfying opportunities (Wright and Panksepp, 2012).[14] The existence of this system (and the vital needs it serves) explains why animals cannot minimize their free energy by simply finding a dark, unchanging chamber and staying there (Friston, Thornton and Clark, 2012). Furthermore, it speaks to the role of dopamine in setting the precision and selecting appropriate courses of action (Friston *et al.*, 2014; 2012).

Two general rules emerge here. First: the crossing of innate (prior) thresholds concerning vital need parameters arouses *negative affects* which in turn trigger *instinctual predictions* as to which stereotyped behaviour is adequate to meet the relevant need (to resolve the affect and resolve uncertainty about the current state of affairs; Pezzulo, Rigoli and Friston, 2015). These actions are automatic. Second: for the animal to survive in unpredicted environments, the prior instinctual action plans — embodied mainly in the basal ganglia — must be supplemented by *learning from experience*, via the generation of experienced *positive affects* that point toward satiation (i.e.

[14] Once the circular action–perception causality described above has been established, new iterations can be initiated at any point in the cycle (e.g. by fortuitous sensory states which possess 'incentive salience'). Action is of course initiated both by needs and opportunities.

homeostatic settling points), like successful instinctual actions do. These actions, which underpin *plasticity*, are voluntary. Since they entail uncertainty and choice, they must be guided in the here-and-now by changes in expected values (predicted 'good' *vs*. 'bad' outcomes). According to our formulation of precision (above), such here-and-now evaluation is achieved through *feeling*. Thus, learning from experience is made possible by affective *qualia* (conscious assessment of the existential 'good' *vs.* 'bad', encoded as changes in expected uncertainty), which guide (i.e. select) actions and sensations in accordance with prior preferences (i.e. existential values). The organism's voluntary acts would otherwise be subverted (see Equation 1a), leading to a failure of active inference and adynamia/abulia (*cf.* Parkinsonian diseases).

This is the *causal contribution* of qualia. As stated above, the same contribution could conceivably be made by non-conscious 'feelings' — i.e. precision-weighted prediction errors — if evolution had found another way for organisms to pre-emptively register and prioritize (to themselves and for themselves) such inherently qualitative existential risks. But the fact that something can conceivably be done differently doesn't mean that it is not done in the way that it actually is in the mammalian nervous system. (As Jean-Martin Charcot is reputed to have said: 'Theory is good, but it doesn't prevent things from existing.')

Panksepp (1998) calls this dopaminergic foraging activity 'SEEKING' while Berridge (1996) calls it 'wanting'. The distinction between their terms reflects the impact of learning upon the primary instinctual mechanism. Primary appetitive SEEKING brings the animal into contact with external states (encoded by Q) that happen to satisfy its needs, thereby producing consummatory *experiences* (what Berridge calls 'liking', leading to satiation), which link particular sensory states (φ) with positive affects (ω), causing subsequent wanting of those states. Such experiences of satisfaction give rise to the cause-and-effect predictions that constitute the very fabric of the brain's generative model (ψ), i.e. long-term memory (LTM). Subsequent iterative testing and refining of hypotheses implicit in LTM is mediated by the same dopaminergic system, through coding of what Schultz (2016) calls 'reward prediction error'. On the proposed view, this 'reward prediction error' is the precision of beliefs about (proprioceptive) action. In other words, when sensory cues resolve any uncertainty about 'what to do next', precision increases in a way that is plausibly mediated by phasic dopaminergic responses. These

discharges enable action selection and facilitate learning (habituation) through an enabling of synaptic plasticity (Frank, 2005; Friston *et al.*, 2014; Hazy, Frank and O'Reilly, 2010).

It is important to recognize that the range of organismic needs and their associated feelings and instinctual behaviours exceed those associated with the *homeostatic* affects (like hunger, thirst, and sleepiness; Peters, McEwan and Friston, 2017; Seth and Friston, 2016). Similar mechanisms to those just described for SEEKING — involving mainly limbic circuitry arising from the upper brainstem — apply to all the *emotional* affects (LUST, FEAR, RAGE, PANIC/GRIEF, CARE, PLAY) and to the *sensory* affects too (surprise, disgust, pain, etc.; see Panksepp, 1998).[15] Each of these many vectors (or 'flavours') of affective qualia (which attribute biological meaning [prior preferences] to survival situations that the animal is bound to encounter through SEEKING) selects its own innate prediction as to what the organism must *do* — when in a need state of, say, thirst versus separation distress versus disgust — and these prior predictions must all be supplemented by learning from experience. This is the main task of mental life (i.e. learning how to meet organismic needs in the world; to stay alive and reproduce). This task is greatly assisted by qualia.

We have foregrounded the role of consciousness in learning processes. However, the ideal of learning is to *automatize* reliable predictions, through consolidation, ultimately down to subcortical non-declarative memory systems (which are 'hard to learn and hard to forget' and in some respects 'indelible'; LeDoux, 1996); so that these acquired predictions may come to resemble the innate ones. The

15 For example: the FEAR circuit (mainly glutamatergic but modulated by peptides DBI, CRF, CCK, alpha MSH, and NPY, projecting from lateral and central amygdala via anterior and medial hypothalamus to PAG), which mediates the *need* for the animal to avoid danger, triggers feelings of trepidation and behaviours of freezing/fleeing; and the PANIC/GRIEF circuit (mainly opioidergic but modulated by oxytocin, prolactin, CRF, and ACh, projecting from ACC via various diencephalic sites to PAG), which mediates the need for close proximity to caregivers, triggers feelings of separation distress and behaviours of protest vocalization/searching. However, the animal has to learn *what* to fear and attach to. These survival tools (emotional affects) are intrinsic *brain* states, embodying preferred self/other relations of universal biological significance; they are not read-outs of current *visceral* states. *Cf.* the James-Lange theory of emotion. The term 'emotional affects' (as opposed to 'homeostatic affects') does not imply that emotional affects are regulated by non-homeostatic mechanisms. It implies only that they are not driven by current bodily needs. The same applies to 'sensory affects' (see Panksepp, 1998).

cortical declarative memory systems, by contrast, are always ready, on the basis of prediction error, to consciously 'entertain a mental image'. In other words, declarative systems readily return long-term memories (LTM) to the short-term (STM) state of *conscious* working memory — in order to update them. This necessarily entails re-activation (i.e. selection) of salient cortical representations of relations between active and sensory states, by way of the relevant vectors of (i.e. affective 'flavours' of) upper brainstem/limbic arousal. The salient cortical traces are thus palpated with feeling (attended to), rendering them conscious once more. This reversal of the consolidation process (*reconsolidation*; Nader, Schafe and LeDoux, 2000) renders memory-traces labile, through literal dissolution of the proteins that initially 'wired' them (Hebb, 1949). This iterative feeling and re-feeling one's way through declarable problems is the mechanism of cortical (exteroceptive and proprioceptive) qualia, which have so dominated contemporary consciousness studies. In short, *predictive-work-in-progress* (see above) *is reconsolidation*. One is reminded of Freud's (1920) obscure dictum: 'consciousness arises instead of a memory-trace' (i.e. a labile trace is not a trace; see Solms, 2017b). Perceptual/cognitive consciousness (activated via attention), no less than affect, is a product of *uncertainty*. Non-declarative (subcortical) memory-traces are far less uncertain — more precise but also less complex — than declarative (cortical) ones. The relative degree of precision typically attaching to cortical versus subcortical versus autonomic prediction errors, therefore, coincides with the relative plasticity (resistance to change) of their associated beliefs (i.e. more precision = less local plasticity).

Lastly, minimal selfhood need not involve what we call cognitive consciousness (cortical re-representation of the subject as an object; see footnote 3). Only reflective consciousness requires both a sentient self *and* a self-representation. Affect *just is* a self-state (and through feeling — i.e. precision optimization — it necessarily generates consciousness itself), which selects salient perceptual representations, which eventually include cognitive re-representations of the self. As these (reflective re-representations, which are greatly facilitated by language) supervene, *thinking* becomes possible. Thinking is a virtual form of acting, a virtual form of hypothesis testing (Attias, 2003; Baker, Saxe and Tenenbaum, 2009; Hobson and Friston, 2014; Metzinger, 2003), which consists in prefrontal activation of cortical representations only (including representations of the self) without necessarily triggering action in the world. Testing predictions in this

(virtual) way saves lives, which is presumably why reflective cognition — which hides so much else from view — evolved. Minimal selfhood in our model example, by contrast, requires nothing located above the level of the superior colliculi and PAG — the 'synencephalic bottleneck' of Merker (2007) or 'SELF' of Panksepp (1998) — the final common pathway for all target and action selection guided by motivational states.

7. Conclusion

Descartes famously claimed that each of us knows only one thing for certain: 'I think, therefore I am.' In this article, we have made two related claims. (1) Descartes' reflective declaration can be reduced to a simpler truth: for us vertebrates, at least, being is feeling, i.e. *what it is like to be* is to feel. (2) Being (and therefore feeling) is ultimately further reducible to *resisting entropy* — resisting dissipation — a process that arises naturally from the fact that any ergodic random dynamical system must differentiate itself from its environment (literally come into being) through the formation of a Markov blanket, whereafter it can respond only to its own states, which (through precision optimization) are *felt*.

In one sense, this inverts the Cartesian position to imply that 'I am, therefore I think'. In other words, I am ergodic, therefore I must infer states of my body and the world from my (interoceptive and exteroceptive) sensorium. The thesis on offer here goes further: it suggests that experience rests upon selecting those aspects of the sensorium that underwrite 'thinking', i.e. abductive inference, or inference to the best explanation (Seth, 2015). This private feeling of one's own abduction (i.e. beliefs about beliefs) appears to be mediated by primitive neurobiological systems that are deeply implicated in (affective) consciousness by neuropsychological, neurophysiological, and neuropharmacological evidence; namely, the ascending neuromodulatory systems that broadcast signals to the entire brain. On this view, feeling and awareness become formally isomorphic with attentional selection (cast here in terms of precision control), in the sense that we cannot be aware of that which is not attended. Crucially, from a technical perspective, consciousness therefore arises from best guesses about beliefs (i.e. the inferred precision of a probability distribution over the causes of sensations). This leads to the notion: 'I am, because I feel, therefore I think.'

Rather than ending with a philosophical aphorism, however, we would like to conclude with the hope that we have persuaded some readers that the imperatives of our physiology and the physics of self-organization provide a plausible scientific response to the psychological question: *why is there something it is like to be an organism, for the organism, and how does this something-it-is-like-ness come about?*

As nicely summarized by one of our reviewers: 'The free energy framework provides an advance over previous suggestions for ['correlates' of sentience] because it comes with some properties that make it a good fit for central aspects of consciousness: clear articulations of affect, attention, and exteroception, and their common ground in precision optimization. In particular, the idea that active inference is associated with a sense of a self being there, through expected free energy, is coming close to capturing an intrinsic aspect of consciousness that other accounts tend to ignore. Together, these properties of the free energy framework make it an attractive candidate for further study in the science of consciousness.'

Acknowledgments

MS is funded by a National Research Foundation Incentive Grant (Ref: 95781). KF is funded by a Wellcome Trust Principal Research Fellowship (Ref: 088130/Z/09/Z) and would also like to express his gratitude to J. Allan Hobson for foundational discussions and scholarly direction.

References

Adams, R.A., Shipp, S. & Friston, K.J. (2013) Predictions not commands: Active inference in the motor system, *Brain Structure and Function*, **218**, pp. 611–643.

Ainley, V., Apps, M.A.J., Fotopoulou, A. & Tsakiris. M. (2016) 'Bodily precision': A predictive coding account of individual differences in interoceptive accuracy, *Philosophical Transactions of the Royal Society B*, **371**, 20160003, [Online], http://dx.doi.org/10.1098/rstb.2016.0003.

Attias, H. (2003) Planning by probabilistic inference, *Proceedings of the 9th International Workshop on Artificial Intelligence and Statistics*.

Baker, C.L., Saxe, R. & Tenenbaum, J.B. (2009) Action understanding as inverse planning, *Cognition*, **113**, pp. 329–349.

Barrett, L.F. & Simmons, W.K. (2015) Interoceptive predictions in the brain, *Nature Reviews Neuroscience*, **16**, pp. 419–429.

Bastos, A.M., Usrey, W.M., Adams, R.A., Mangun, G.R., Fries, P. & Friston, K.J. (2012) Canonical microcircuits for predictive coding, *Neuron*, **76**, pp. 695–711.

Berridge, K. (1996) Food reward: Brain substrates of wanting and liking, *Neuroscience & Biobehavioral Reviews*, **20**, pp. 1–25.

Brentano, F. (1874) *Psychology from an Empirical Standpoint*, London: Routledge.
Brown, H., Adams, R.A., Parees, I., Edwards, M. & Friston, K. (2013) Active inference, sensory attenuation and illusions, *Cognitive Processing*, **14**, pp. 411–427.
Carhart-Harris, R. & Friston, K. (2010) The default mode, ego functions and free energy: A neurobiological account of Freudian ideas, *Brain*, **133**, pp. 1265–1283.
Chalmers, D.J. (1995) Facing up to the problem of consciousness, *Journal of Consciousness Studies*, **2** (3), pp. 200–219.
Chalmers, D.J. (1996) *The Conscious Mind: In Search of a Fundamental Theory*, Oxford: Oxford University Press.
Cisek, P. & Kalaska, J.F. (2010) Neural mechanisms for interacting with a world full of action choices, *Annual Review of Neuroscience*, **33**, pp. 269–298.
Clark, A. (2013) The many faces of precision, *Frontiers of Psychology*, **4**, art. 270.
Clark, A. (2017) How to knit your own Markov blanket, in Metzinger, T.K. & Wiese, W. (eds.) *Philosophy and Predictive Processing*, Frankfurt am Main: MIND Group.
Conant, R.C. & Ashby, W.R. (1970) Every good regulator of a system must be a model of that system, *International Journal of Systems Science*, **1**, pp. 89–97.
Craig, A.D. (2009) How do you feel — now? The anterior insula and human awareness, *Nature Reviews Neuroscience*, **10**, pp. 59–70.
Crick, F. (1994) *The Astonishing Hypothesis*, New York: Scribner.
Crucianelli, L., Krahé, C., Jenkinson, P. & Fotopoulou, A. (2017) Interoceptive ingredients of body ownership: Affective touch and cardiac awareness in the rubber hand illusion, *Cortex*, [Online], http://dx.doi.org/10.1016/j.cortex.2017.04.018.
Damasio, A. (2010) *Self Comes to Mind*, New York: Pantheon.
Damasio, A., Damasio, H. & Tranel, D. (2012) Persistence of feeling and sentience after bilateral damage of the insula, *Cerebral Cortex*, **23** (4), pp. 833–846.
Dayan, P. (2012) Twenty-five lessons from computational neuromodulation, *Neuron*, **76**, pp. 240–256.
Decety, J. & Fotopoulou, A. (2015) Why empathy has a beneficial impact on others in medicine: Unifying theories, *Frontiers in Behavioral Neuroscience*, **8**, art. 457.
Dehaene, S. & Changeux, J.-P. (2011) Experimental and theoretical approaches to conscious processing, *Neuron*, **70**, pp. 200–227.
Du Bois-Reymond, E. (1918) *Jugenbriefe, von Émil du Bois-Reymond an Eduard Hallmann*, du Bois-Reymond, E. (ed.), Berlin: D. Reimer.
Feldman, H. & Friston, K.J. (2010) Attention, uncertainty, and free-energy, *Frontiers in Human Neuroscience*, **4**, art. 215.
Ferrarelli, F. & Tononi, G. (2011) The thalamic reticular nucleus and schizophrenia, *Schizophrenia Bulletin*, **37**, pp. 306–315.
Fotopoulou, A. (2013) Beyond the reward principle: Consciousness as precision seeking, *Neuropsychoanalysis*, **15**, pp. 33–38.
Fotopoulou, A. & Tsakiris, M. (2017a) Mentalizing homeostasis: The social origins of interoceptive inference, *Neuropsychoanalysis*, **19**, pp. 3–28.

Fotopoulou, A. & Tsakiris, M. (2017b) Mentalizing homeostasis: The social origins of interoceptive inference — replies to commentaries, *Neuropsychoanalysis*, **19**, pp. 71–76.
Frank, M.J. (2005) Dynamic dopamine modulation in the basal ganglia: A neurocomputational account of cognitive deficits in medicated and nonmedicated Parkinsonism, *Journal of Cognitive Neuroscience*, **1**, pp. 51–72.
Freud, S. (1895/1950) Project for a scientific psychology, *Standard Edition of the Complete Psychological Works of Sigmund Freud*, **1**, pp. 281–397, London: Hogarth.
Freud, S. (1896/1950) Letter of January 1 to Wilhelm Fliess, *Standard Edition of the Complete Psychological Works of Sigmund Freud*, **1**, pp. 388–391, London: Hogarth.
Freud, S. (1920) Beyond the pleasure principle, *Standard Edition of the Complete Psychological Works of Sigmund Freud*, **18**, pp. 7–64, London: Hogarth.
Friston, K. (2010) The free-energy principle: A unified brain theory?, *Nature Reviews Neuroscience*, **11**, pp. 127–138.
Friston, K. (2013) Life as we know it, *Journal of the Royal Society Interface*, **10**, 20130475.
Friston, K., Kilner, J. & Harrison, L. (2006) A free energy principle for the brain, *Journal of Physiology, Paris*, **100**, pp. 70–87.
Friston, K., Breakspear, M. & Deco, G. (2012) Perception and self-organized instability, *Frontiers in Computational Neuroscience*, **6**, art. 44.
Friston, K., Thornton, C. & Clark, A. (2012) Free-energy minimization and the dark-room problem, *Frontiers in Psychology*, **3**, art. 130.
Friston, K., Shiner, T., Fitzgerald, T., Galea, J.M., Adams, R., Brown, H., Dolan, R.J., Moran, R., Stephan, K.E. & Bestmann, S. (2012) Dopamine, affordance and active inference, *PLoS Computational Biology*, **8**, e1002327.
Friston, K., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T. & Dolan, R.J. (2014) The anatomy of choice: Dopamine and decision-making, *Philosophical Transactions of the Royal Society of London B*, **369**, 20130481.
Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T. & Pezzulo, G. (2015) Active inference and epistemic value, *Cognitive Neuroscience*, **6**, pp. 187–214.
Frith, C.D., Blakemore, S.J. & Wolpert, D.M. (2000) Abnormalities in the awareness and control of action, *Philosophical Transactions of the Royal Society of London B*, **355**, pp. 1771–1788.
Gregory, R.L. (1980) Perceptions as hypotheses, *Philosophical Transactions of the Royal Society of London B*, **290**, pp. 181–197.
Haken, H. (1983) *Synergetics: An Introduction. Non-Equilibrium Phase Transition and Self-Organisation in Physics, Chemistry and Biology*, Berlin: Springer Verlag.
Harman, G.H. (1965) The inference to the best explanation, *Philosophical Review*, **74**, pp. 88–95.
Harlow, J.M. (1868) Recovery from the passage of an iron bar through the head, *Publications of the Massachusetts Medical Society*, **2** (3), pp. 327–347.
Hazy, T.E., Frank, M.J. & O'Reilly, R.C. (2010) Neural mechanisms of acquired phasic dopamine responses in learning, *Neuroscience & Biobehavioral Review*, **34**, pp. 701–720.
Hebb, D. (1949) *The Organization of Behavior: A Neuropsychological Theory*, New York: Wiley and Sons.

Hobson, J.A. (2009) REM sleep and dreaming: Towards a theory of protoconsciousness, *Nature Reviews Neuroscience*, **10**, pp. 803–813.
Hobson, J.A. & Friston, K.J. (2014) Consciousness, dreams, and inference: The Cartesian theatre revisited, *Journal of Consciousness Studies*, **21** (1–2), pp. 6–32.
Hohwy, J. (2013) *The Predictive Mind*, Oxford: Oxford University Press.
Hohwy, J. (2016) The self-evidencing brain, *Noûs*, **50**, pp. 259–285.
Huston, J. & Borbely, A. (1974) The thalamic rat: General behaviour, operant learning with rewarding hypothalamic stimulation, and effects of amphetamine, *Physiology & Behavior*, **12**, pp. 433–448.
Kanai, R., Komura, Y., Shipp, S. & Friston, K. (2015) Cerebral hierarchies: Predictive processing, precision and the pulvinar, *Philosophical Transactions of the Royal Society of London B*, **370**, 20140169.
Kauffman, S. (1993) *The Origins of Order: Self-Organization and Selection in Evolution*, Oxford: Oxford University Press.
Kelso, J.A.S. (1995) *Dynamic Patterns: The Self-Organization of Brain and Behavior*, Cambridge, MA: MIT Press.
Kihlstrom, J. (1996) Perception without awareness of what is perceived, learning without awareness of what is learned, in Velmans, M. (ed.) *The Science of Consciousness: Psychological, Neuropsychological and Clinical Reviews*, pp. 23–46, London: Routledge.
Koch, C. & Tsuchiya, N. (2012) Attention and consciousness: Related yet different, *Trends in Cognitive Sciences*, **16**, pp. 103–105.
Krahé, C., Springer, A., Weinman, J. & Fotopoulou, A. (2013) The social modulation of pain: Others as predictive signals of salience — a systematic review, *Frontiers in Human Neuroscience*, **7**, art. 386.
LeDoux, J. (1996) *The Emotional Brain*, New York: Simon and Shuster.
LeDoux, J. & Brown, R. (2017) A higher-order theory of emotional consciousness, *Proceedings of the National Academy of Sciences*, **114**, pp. E2016–E2025.
Lisman, J. & Buzsaki, G. (2008) A neural coding scheme formed by the combined function of gamma and theta oscillations, *Schizophrenia Bulletin*, **34**, pp. 974–980.
Maturana, H.R. & Varela, F. (1980) Autopoiesis: The organization of the living, in Maturana, H.R. & Varela, F. (eds.) *Autopoiesis and Cognition*, Dordrecht: Reidel.
Merker, B. (2007) Consciousness without a cerebral cortex: A challenge for neuroscience and medicine, *Behavioral and Brain Sciences*, **30**, pp. 63–134.
Mesulam, M.M. (2000) Behavioral neuroanatomy: Large-scale networks, association cortex, frontal syndromes, the limbic system and hemispheric lateralization, in *Principles of Behavioral and Cognitive Neurology*, 2nd ed., pp. 1–120, New York: Oxford University Press.
Metzinger, T. (2003) *Being No One: The Self-Model Theory of Subjectivity*, Cambridge, MA: MIT Press.
Moruzzi, G. & Magoun, H. (1949) Brain stem reticular formation and activation of the EEG, *Electroencephalography & Clinical Neurology*, **1**, pp. 455–473.
Moustafa, A.A., Sherman, S.J. & Frank, M.J. (2008) A dopaminergic basis for working memory, learning and attentional shifting in Parkinsonism, *Neuropsychologia*, **46**, pp. 3144–3156.
Mumford, D. (1992) On the computational architecture of the neocortex. II, *Biological Cybernetics*, **66**, pp. 241–251.

Nader, K., Schafe, G.E. & LeDoux, J. (2000) Fear memories require protein synthesis in the amygdala for reconsolidation after retrieval, *Nature*, **406**, pp. 722–726.
Nagel, T. (1974) What is it like to be a bat?, *Philosophical Review*, **83**, pp. 435–450.
Nour, M.M. & Carhart-Harris, R.L. (2017) Psychedelics and the science of self-experience, *The British Journal of Psychiatry*, **210**, pp. 177–179.
Palmer, C.J., Seth, A.K. & Hohwy, J. (2015) The felt presence of other minds: Predictive processing, counterfactual predictions, and mentalising in autism, *Consciousness & Cognition*, **36**, pp. 376–389.
Paloyelis, Y., Krahé, C., Maltezos, S., Williams, S.C., Howard, M.A. & Fotopoulou, A. (2016) The analgesic effect of oxytocin in humans: A double-blind, placebo-controlled cross-over study using laser-evoked potentials, *Journal of Neuroendocrinology*, **28**, 10.1111/jne.12347.
Panksepp, J. (1998) *Affective Neuroscience: The Foundations of Animal and Human Emotions*, New York: Oxford University Press.
Panksepp, J., Lane, R.D., Solms, M. & Smith, R. (2016) Reconciling cognitive and affective neuroscience perspectives on the brain basis of emotional experience, *Neuroscience & Biobehavioral Reviews*, **76**, pp. 187–215.
Parvizi, J. & Damasio, A. (2001) Consciousness and the brainstem, *Cognition*, **79**, pp. 135–159.
Penfield, W. & Jasper, H. (1954) *Epilepsy and the Functional Anatomy of the Human Brain*, Oxford: Little & Brown.
Peters, A., McEwen, B.S. & Friston, K. (2017) Uncertainty and stress: Why it causes diseases and how it is mastered by the brain, *Progress in Neurobiology*, **156**, pp. 164–188.
Pezzulo, G., Rigoli, F. & Friston, K. (2015) Active inference, homeostatic regulation and adaptive behavioural control, *Progress in Neurobiology*, **134**, pp. 17–35.
Pfaff, D. (1999) *Drive: Neurobiological and Molecular Mechanisms of Sexual Motivation*, Cambridge, MA: MIT Press.
Pfaff, D. (2006) *Brain Arousal and Information Theory*, Cambridge, MA: Harvard University Press.
Picard, F. & Friston, K. (2014) Predictions, perception and a sense of self, *Neurology*, **83**, pp. 1112–1118.
Quattrocki, E. & Friston, K. (2014) Autism, oxytocin and interoception, *Neuroscience and Biobehavioral Reviews*, **47c**, pp. 410–430.
Riggs, L.A., Ratliff, F., Cornsweet, J.C. & Cornsweet, T.N. (1953) The disappearance of steadily fixated visual test objects, *Journal of the Optical Society of America*, **43**, pp. 495–501.
Schultz, W. (2016) Dopamine reward prediction error coding, *Dialogues in Clinical Neuroscience*, **18**, pp. 23–32.
Seth, A.K. (2013) Interoceptive inference, emotion, and the embodied self, *Trends in Cognitive Sciences*, **17**, pp. 565–573.
Seth, A.K. (2014) The cybernetic brain: From interoceptive inference to sensorimotor contingencies, in Metzinger, T.K. & Windt, J.M. (eds.) *Open MIND*, Frankfurt am Main: MIND Group.
Seth, A.K. (2015) Inference to the best prediction, in Metzinger, T.K. & Windt, J.M. (eds.) *Open MIND*, Frankfurt am Main: MIND Group.

Seth, A.K. & Friston, K.J. (2016) Active interoceptive inference and the emotional brain, *Philosophical Transactions of the Royal Society of London B*, **371**, 20160007.

Shannon, C. (1948) A mathematical theory of communication, *Bell System Technical Journal*, **27**, pp. 379–423.

Shewmon, D., Holmse, D. & Byrne, P. (1999) Consciousness in congenitally decorticate children: Developmental vegetative state as a self-fulfilling prophecy, *Developmental Medicine & Child Neurology*, **41**, pp. 364–374.

Shipp, S. (2016) Neural elements for predictive coding, *Frontiers in Psychology*, **7**, art. 1792.

Solms, M. (2013) The conscious id, *Neuropsychoanalysis*, **15**, pp. 5–19.

Solms, M. (2017a) Consciousness by surprise: A neuropsychoanalytic approach to the hard problem, in Poznanski, R., Tuszynski, J. & Feinberg, T. (eds.) *Biophysics of Consciousness: A Foundational Approach*, pp. 129–148, New York: World Scientific.

Solms, M. (2017b) What is 'the unconscious' and where is it located in the brain? A neuropsychoanalytic perspective, *Annals of the New York Academy of Sciences*, **1406**, pp. 90–97.

Solms, M. & Panksepp, J. (2012) The 'id' knows more than the 'ego' admits, *Brain Sciences*, **2**, pp. 147–175.

Tozzi, A., Zare, M. & Benasich, A. (2016) New perspectives on spontaneous brain activity: Dynamic networks and energy matter, *Frontiers in Human Neuroscience*, **10**, art. 247.

Uhlhaas, P.J. & Singer, W. (2010) Abnormal neural oscillations and synchrony in schizophrenia, *Nature Reviews Neuroscience*, **11**, pp. 100–113.

van Boxtel, J.J.A., Tsuchiya, N. & Koch, C. (2010) Opposing effects of attention and consciousness on afterimages, *Proceedings of the National Academy of Sciences*, **107**, pp. 8883–8888.

von Mohr, M. & Fotopoulou, A. (2017) The cutaneous borders of interoception: Active and social inference of pain and pleasure on the skin, in Tsakiris, M. & de Preester, H. (eds.) *The Interoceptive Basis of the Mind*, Oxford: Oxford University Press.

Wright, J. & Panksepp, J. (2012) An evolutionary framework to understand foraging, wanting, and desire: The neuropsychology of the SEEKING system, *Neuropsychoanalysis*, **14**, pp. 5–39.

Paper received August 2017; revised January 2018.

Appendix

In his exhaustive treatment of the topic of brain arousal, Pfaff[16] (2006, pp. 2–6) comments as follows:

> Satisfying the need for an 'energy source' for behavior, arousal explains the initiation and persistence of motivated behavior in a wide variety of species… Arousal, fuelling drive mechanisms, potentiates behavior,

16 Quoted with permission.

while specific motives and incentives explain why an animal does one thing and not another… The *Dictionary of Ethology* not only emphasizes arousal in the context of the sleep–wake cycle but also refers to the overall state of responsiveness of the animal, as indicated by the intensity of stimulation necessary to trigger a behavioral reaction. Arousal 'moves the animal towards readiness for action from a state of inactivity.' In the case of directed action, a founder of ethology, Nikko Tinbergen, would say arousal provides the motoric energy for a 'fixed action pattern' in response to a 'sign stimulus'. The dictionary does not eschew neurophysiology, as it also covers arousal levels indicated by the cortical electroencephalogram (EEG)… Generations of behavioral scientists have both theorized and experimentally confirmed that a concept like arousal is necessary to explain the initiation, strength, and persistence of behavioral responses. Arousal provides the fundamental force that makes animals and humans active and responsive so they will perform instinctive behaviors or learned behaviors directed toward goal objects. The strength of a learned response depends on arousal and drive. Hebb saw a state of generalized activation as fundamental to optimal cognitive performance. Duffy goes even further by invoking the concept of 'activation' to account for a significant part of an animal's behavior.[17] She anticipated that quantitative physiologic or physical measures would allow a mathematical approach to this aspect of behavioral science… Cannon brought in the autonomic nervous system as a necessary mechanism by which arousal prepares the animal for muscular action. Entire theories of emotion were based on the activation of behavior… Malmo brought all of this material together by citing EEG evidence and physiologic data, which go along with behavioral results in establishing activation and arousal as primary components driving all behavioral mechanisms…

This is the classic arousal problem: How do internal and external influences wake up brain and behavior, whether in humans or in other animals, whether in the laboratory or in natural, ethological settings? It is important to reformulate and solve this problem because we are dealing with responsivity to the environment, one of the elementary requirements for animal life. It is also timely to reformulate and solve this problem now because new neurobiologic, genetic, and computational tools have opened up approaches to 'behavioral states' that were never possible before… Explaining arousal will permit us to understand the states of behavior that lie beneath large numbers of specific response mechanisms. Not only is it strategic to accomplish the analysis of many behaviors all at once but also elucidating mechanisms of behavioral states leads to an understanding of mood and temperament.

---

17 Pfaff's own principal component analyses suggest that the proportion of behaviour across a wide range of data that can be accounted for by 'generalized arousal' is between 30% and 45%.

> To put it another way, much of twentieth-century neuroscience was directed at explaining the particularity of specific stimulus/response connections. Now we are in a position to reveal mechanisms of entire classes of responses under the name of 'state control'. Most important are the mechanisms determining the level of arousal…
>
> Any truly universal definition of arousal must be elementary and fundamental, primitive and undifferentiated, and not derived from higher CNS functions. It cannot be limited by particular, temporary conditions or measures. For example, it cannot be confined to explaining responses to only one stimulus modality. Voluntary motor activity and emotional responses should also be included. Therefore, I propose the following as an operational definition that is intuitively satisfying and that will lead to precise quantitative measurements: *'Generalized arousal' is higher in an animal or human being who is: (S) more alert to sensory stimuli of all sorts, and (M) more motorically active, and (E) more reactive emotionally.* This is a concrete definition of the most fundamental force in the nervous system… All three components can be measured with precision… Clearly there is a neuroanatomy of generalized arousal, there are neurons whose firing patterns lead to it, and genes whose loss disrupts it. Therefore… generalized arousal is the behavioral state produced by arousal pathways, their electrophysiological mechanisms, and genetic influences. The fact that these mechanisms produce the same sensory alertness (S), motor reactivity (M) and emotional reactivity (E) as our definition affirms the existence of a generalized arousal function and the accuracy of its operational definition.

Pfaff continues: 'Because CNS arousal depends on surprise and unpredictability, its appropriate quantification depends on the mathematics of *information*' (p. 13, emphasis added). Shannon's (1948) equation makes information measurable:

> If any event is perfectly regular, say the ticking of a metronome, the next event (the next tick) does not tell us anything new. It has an extremely high probability (p) of occurrence in exactly that time bin… We have no uncertainty about whether, in any given time bin, the tick will occur. In Shannon's equation, the information in any event is in inverse proportion to its probability. Put another way, the more uncertain we are about the occurrence of that event, the more information is transmitted, inherently, when it does happen… When all events in an array of events are equally probable, information is at its top value. Disorder maximises information flow. Coming from thermodynamics, the technical term for disorder in Shannon's equation is entropy. His symbol for entropy is H… The information content inherent in some event x is:
>
> $$H[p(x)] = -\sum_x p(x) \log_e p(x)$$

Where $p(x)$ is the probability of event x.

Pfaff sums up (pp. 19–20):

> For a lower animal or human to be aroused, there must be some change in the [interoceptive or exteroceptive] environment. If there is change, there must be some uncertainty about the state of the environment. Quantitatively, to the degree that there is uncertainty, predictability is decreased. Given these considerations, we can use [Shannon's equation] to state that the less predictable the environment and the greater the entropy, the more information is available. Arousal of brain and behavior, and information calculations, are inseparably united.

In short, unknown, unexpected, disordered, and unusual (high-information) stimuli produce and sustain arousal responses.

> Information theory has been lurking behind behavioral investigations and neurophysiologic data all along. First, in clear and simple logic, consider what is required for an animal or human being to rouse itself to action. Second, consider what is required to recognize a familiar stimulus (habituation) and to give special attention to a novel stimulus. Third, from the experimenter's point of view, information theory provides methods for calculating the meaningful content of spike trains and quantifying the cognitive load of certain environmental situations. New questions can be asked: How much distortion of a sensory stimulus field is required for novelty? What kinds of generalization from a specific type of stimulus are allowed for a given type of response? The information theoretic approach will help us to turn the combination of genetics, neurophysiology, and behavior into a quantitative science. We can use the 'mathematics of arousal' to help analyse neurobiologic mechanisms. (*ibid.*, p. 23)

Pfaff ultimately concludes (pp. 138–45):

> CNS arousal systems battle heroically against the Second Law of Thermodynamics in a very special way. They respond selectively to environmental situations that have an inherently high entropy — a high degree of uncertainty and therefore information content. But in responding, CNS arousal systems effectively reduce entropy by compressing all of that information into a single, lawful response… Arousal neurobiology is the neuroscience of change, uncertainty, unpredictability, and surprise — that is, of information science. Throughout all of the analyses of arousal mechanisms in the CNS so far — neuroanatomic, physiologic, genetic, and behavioural — the concepts of information theory have proven useful. The mathematics of information provides ways of classifying responses to natural stimuli. Nerve cells actually encode probabilities and uncertainties, with the result that they can guide behavior in unpredictable circumstances. CNS arousal itself absolutely depends on change, uncertainty, unpredictability, and surprise. The huge phenomenon called habituation, a decline in response ampli-

> tude on repetition of the same stimulus, pervades neurophysiology, behavioral science, and autonomic physiology; and it shows us how declining information content leads to declining CNS arousal. Thus, arousal theory and information theory were made for each other.

It is important to recognize that the 'mathematics of information' explains the behaviour of neurons in both arousal processes and learning/memory processes, which, combined, determine what the brain *does*. Therefore, although 'information' is not a physical construct, it explains (i.e. lawfully organizes) the physical activity of the brain. It is the *function* that is selected by evolution; the phenotypes follow.