entropy

MDPI

Article

Sentience and the Origins of Consciousness: From Cartesian Duality to Markovian Monism

Karl J. Friston [1,*], Wanja Wiese [2,*] and J. Allan Hobson [3,*]

[1] The Wellcome Centre for Human Neuroimaging, Institute of Neurology, Queen Square, London WC1N 3AR, UK

[2] Department of Philosophy, Johannes Gutenberg University Mainz, Jakob-Welder-Weg 18, 55128 Mainz, Germany

[3] Division of Sleep Medicine, Harvard Medical School, 74 Fenwood Road, Boston, MA 02115, USA

* Correspondence: k.friston@ucl.ac.uk (K.J.F.); wawiese@uni-mainz.de (W.W.); allan_hobson@hms.harvard.edu (J.A.H.)

Received: 4 February 2020; Accepted: 16 April 2020; Published: 30 April 2020

Abstract: This essay addresses Cartesian duality and how its implicit dialectic might be repaired using physics and information theory. Our agenda is to describe a key distinction in the physical sciences that may provide a foundation for the distinction between mind and matter, and between sentient and intentional systems. From this perspective, it becomes tenable to talk about the physics of sentience and 'forces' that underwrite our beliefs (in the sense of probability distributions represented by our internal states), which may ground our mental states and consciousness. We will refer to this view as Markovian monism, which entails two claims: (1) fundamentally, there is only one type of thing and only one type of irreducible property (hence *monism*). (2) All systems possessing a Markov blanket have properties that are relevant for understanding the mind and consciousness: if such systems have mental properties, then they have them partly by virtue of possessing a Markov blanket (hence *Markovian*). Markovian monism rests upon the information geometry of random dynamic systems. In brief, the information geometry induced in any system—whose internal states can be distinguished from external states—must acquire a dual aspect. This dual aspect concerns the (intrinsic) information geometry of the probabilistic evolution of internal states and a separate (extrinsic) information geometry of probabilistic beliefs about external states that are parameterised by internal states. We call these intrinsic (i.e., mechanical, or state-based) and extrinsic (i.e., Markovian, or belief-based) information geometries, respectively. Although these mathematical notions may sound complicated, they are fairly straightforward to handle, and may offer a means through which to frame the origins of consciousness.

Keywords: consciousness; information geometry; Markovian monism

1. Introduction

The aim of this essay is to emphasise a couple of key technical distinctions that seem especially prescient for an understanding of the beliefs and intentions that underpin pre-theoretical notions of consciousness. What follows is an attempt to describe constructs from information theory and physics that place certain constraints on the dynamics of self-organising creatures, such as ourselves. These constraints lend themselves to an easy interpretation in terms of beliefs and intentions; provided one defines their meaning carefully in relation to the mathematical objects at hand. The benefit of articulating a calculus of beliefs (and intentions) from first principles has yet to be demonstrated; however, just having a calculus of this sort may provide useful perspectives on current philosophical debates. Furthermore, trying to articulate pre-theoretical notions in terms of maths should, in principle, expand the scope of dialogue in this area. To illustrate this, we will try to license talk about physical forces causing beliefs in a non-mysterious way—a way that clearly identifies systems or artefacts that are and are not equipped with processes that can ground mental capacities and consciousness.

To make a coherent argument along these lines, it will be necessary to introduce a few technical concepts. The formal basis of the arguments in this—more philosophical—treatment of sentience and physics can be found in [1]. The current paper starts were Friston (ibid.) stops; namely, to examine the philosophical implications of Markov blankets and the ensuing Bayesian mechanics. For readers who are more technically minded, the derivations and explanations of the equations in this paper can be found in [1] (using the same notation). We have attempted to unpack the derivations for non-mathematical readers but will retain key technical terms, so that the lineage of what follows can be read clearly. To avoid cluttering the narrative with definitions, a glossary of terms and expressions is provided at the end of the paper. In brief, we first establish the basic setup used to describe physical systems that evince the phenomenology necessary to accommodate pre-theoretical notions of consciousness. This will involve the introduction of Markov blankets and the distinction between the internal and external states of a system or creature.

Having established the distinction between external and internal states, we introduce the notion of information length and information geometry. This is the first key move in the theoretical analysis on offer. Crucially, information geometry allows us to establish a calculus of beliefs in terms of probability distributions. This calculus enables a distinction to be made between the probability distribution *about things* and the probability distribution *of things*. This distinction is then treated as one way of describing an account that (literally) maps belief states onto physical states; here, beliefs about external states that are parameterised, represented, encoded or coherent with internal states. We shall call the ensuing view *Markovian monism* because it is predicated on the existence of a Markov blanket.

This brings us to a modest representationalism[1], which allows one to talk about flows, energy gradients and forces that shape the dynamics of internal states and, necessarily, the beliefs they parameterise. The next section considers the nature of these beliefs and, in particular, beliefs about how internal states couple to external states; namely, beliefs about action upon the world 'out there'. To do this formally, we have to look at two distinct ways of describing the dynamics and introduce the notion of trajectories via the path integral formulation. Having done this, we can then associate intentions with beliefs about action—that, in turn, depend upon beliefs about the consequences of action. At this point, we can make a distinction between systems that have a rudimentary information geometry of a reflexive, instantaneous sort—and systems that hold beliefs about the future. It is this quantitative distinction that may provide a spectrum of intentional or agential systems, ranging from protozoa to people. We conclude with a brief discussion of related formulations—and how the central role of sentience, observation, measurement, or inference opens the door for further developments of a sentient physics. In particular, we will discuss how Markovian monism can be interpreted in terms of existing theories regarding the relationship between mind and matter, such as neutral monism and panprotopsychism.

[1] The important point is that such systems can be described 'as if' they represent probability distributions. More substantial representationalist accounts can be built on this foundation, see Section 13.

The primary target of this paper is sentience. Our use of the word "sentience" here is in the sense of "responsive to sensory impressions". It is not used in the philosophy of mind sense; namely, the capacity to perceive or experience subjectively, i.e., phenomenal consciousness, or having 'qualia'. Sentience here, simply implies the existence of a non-empty subset of systemic states; namely, sensory states. In virtue of the conditional dependencies that define this subset (i.e., the Markov blanket partition), the internal states are necessarily 'responsive to' sensory states and thus the dictionary definition is fulfilled. The deeper philosophical issue of sentience speaks to the hard problem of tying down quantitative experience or subjective experience within the information geometry afforded by the Markov blanket construction. We will return to this below.

While most of this paper deals with sentience in the sense just specified, it may shed light on the origins of consciousness. First, applying the concept of subjective, phenomenal consciousness to a system trivially presupposes that this system can be described from two perspectives (i.e., from a third- and from a first-person perspective). Second, the minimal form of goal-directedness and 'as if' intentionality—that one can ascribe to sentient systems—provide conceptual building blocks that ground more high-level concepts, such as physical computation, intentionality, and representation, which may be useful to understand the evolutionary transition from non-conscious to conscious organisms, and thereby illuminate the origins of consciousness.

2. Markov Blankets and Self-Organisation

Before we can talk about anything, we have to consider what distinguishes a 'thing' from everything else. Mathematically, this requires the existence of a particular partition of all states a system could be in into external, (Markov) blanket and internal states. A Markov blanket comprises a set of states that renders states internal to the blanket conditionally independent of external states. The term was originally coined by Pearl in the context of Bayesian networks [2]. For a Bayesian network (i.e., a directed acyclic graphical model) the Markov blanket comprises the parents, children, and parents of the children of a state or node. For a Markov random field (i.e., an undirected graphical model), the Markov blanket comprises the parents and children, i.e., its neighbours. For a dependency network (i.e., a directed cyclic graphical model) the Markov blanket comprises just the parents. For treatments of Markov blankets in the life sciences, please see [3–8]. The three-way partition induced by the Markov blanket enables one to distinguish internal and external states via their conditional independence, given blanket states. The blanket states themselves can be further partitioned into sensory and active states, where sensory states are not influenced by internal states and active states are not influenced by external states [9]. Note that all we have done here is to stipulatively define a 'thing' in terms of its internal states (and Markov blanket) in terms of what does *not* influence what. The requisite absence of specific influences are precisely those described above; namely, internal states and external states only influence each other via the Markov blanket, while sensory states are not influenced by internal states, a similar relationship is true for active and external states. A key insight here is that structure emerges from influences that *are not there*, much like a sculpture emerges from the material removed. There are lots of interesting implications of defining things in terms of Markov blankets (please see Figure 1 for a couple of intuitive examples); however, we will place the notion of a Markov blanket to one side for the moment and consider how systemic states behave in general. After this, we will then consider the implications of this generic behaviour, when there is a Markov blanket play.

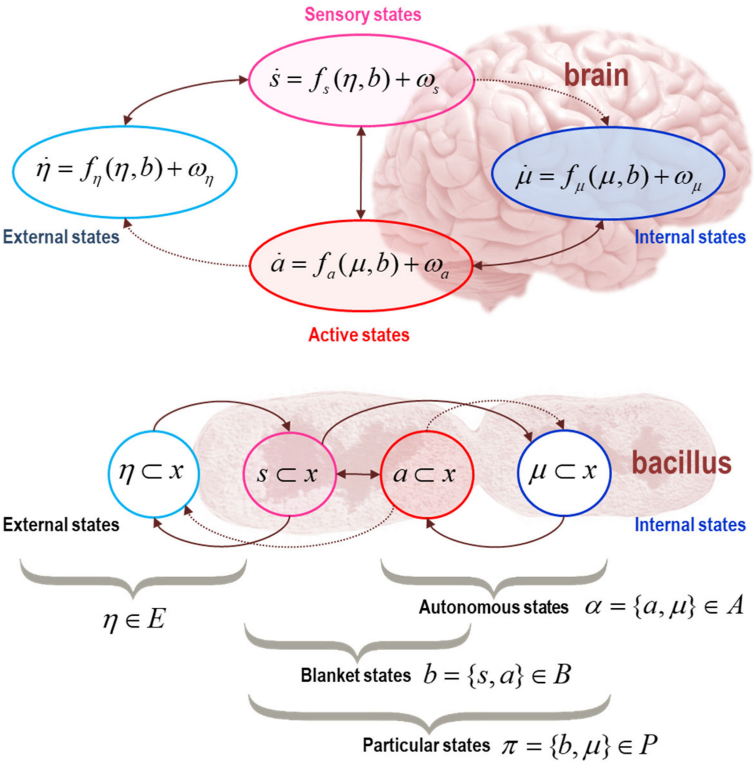
Sensory states
$\dot{s} = f_s(\eta, b) + \omega_s$
brain
$\dot{\eta} = f_\eta(\eta, b) + \omega_\eta$
$\dot{\mu} = f_\mu(\mu, b) + \omega_\mu$
External states
Internal states
$\dot{a} = f_a(\mu, b) + \omega_a$
Active states
$\eta \subset x$
$s \subset x$
$a \subset x$
$\mu \subset x$
bacillus
External states
Internal states
$\eta \in E$
Autonomous states $\alpha = \{a, \mu\} \in A$
Blanket states $b = \{s, a\} \in B$
Particular states $\pi = \{b, \mu\} \in P$

Figure 1. (*Markov blankets*): This schematic illustrates the partition of systemic states into internal states (blue) and hidden or external states (cyan) that are separated by a Markov blanket—comprising sensory (magenta) and active states (red). The upper panel shows this partition as it would be applied to action and perception in the brain. The ensuing self-organisation of internal states then corresponds to perception, while action couples brain states back to external states. The lower panel shows the same dependencies but rearranged so that the internal states are associated with the intracellular states of a Bacillus, while the sensory states become the surface states or cell membrane overlying active states (e.g., the actin filaments of the cytoskeleton).

3. The Langevin Formalism and Density Dynamics

Starting from first principles, if we assume that a system exists, in the sense that it has measurable characteristics over some nontrivial period of time,[2] then we can express its evolution in terms of a random dynamical system. This just means that the system can be described in terms of changes in states over time that are subject to some random fluctuations:

$$x(\tau) = f(x, \tau) + \omega. \tag{1}$$

This is a completely general specification of (Langevin) dynamics that underwrites nearly all of physics [10–12]. In brief, the dynamics in (1) can be described in terms of two equivalent formulations—the dynamics of the accompanying probability density over the states and the path integral formulation.[3]

[2] In the sense that anything just is a Markov blanket, the relevant timescale is the duration over which the thing exists. Generally, smaller things last for short periods of time and bigger things last longer. This is a necessary consequence of composing Markov blankets of Markov blankets (i.e., things of things). In terms of sentient systems, the relevant time scale is the time over which a sentient system persists (e.g., the duration of being a sentient person).

[3] In turn, this leads to quantum, statistical and classical mechanics, which can be regarded as special cases of density dynamics under certain assumptions. For example, when the system attains nonequilibrium steady-state, the solution to the density dynamics (i.e., Fokker Planck equation) becomes the solution to the Schrödinger equation that underwrites quantum electrodynamics. When random fluctuations become negligible (in large systems), we move from the dissipative

We will be interested in systems that have measurable characteristics, which means that they must converge to some attracting set or manifold, known as a random or pullback attractor [13].[4] After a sufficient period of time, as the system evolves, it will trace out a trajectory—in state space—that circulates, usually in a highly itinerant fashion, on the attracting manifold. This means that if we observe the system at random, there is a certain probability of finding it in a particular state. This is known as the nonequilibrium steady-state density [12].

It is natural to ask whether a single attracting manifold is an appropriate construct to describe a system or creature over its lifetime; especially when certain 'life-cycles' have distinct developmental stages or indeed feature metamorphosis. From the perspective of the current argument, it helps to appreciate that the attracting manifold is itself a random set.[5] In other words, a particle or person is never 'off' their manifold—they just occupy states that are more or less likely, given the kind of thing they are (i.e., something's *characteristic* states are an attracting set of states that it is likely to occupy). Technically, this peripatetic itinerancy corresponds to stochastic chaos, where excursions from the attracting set—driven by random fluctuations—are an integral aspect of the dynamics. These excursions are repaired through the flow that counters the effects of random fluctuations and underwrites the information geometry of self-organisation. This formulation can, in principle, accommodate slow changes to the attracting set—and implicit Markov blanket—that may require the notion of wandering sets [14].

The reason that this is interesting is that one can use standard descriptions of density dynamics to express the flow of states as a gradient flow on something called *self-information* or *surprisal* [15–18]. Without going into details, this is the steady-state solution to the Fokker Planck equation [19–23]. This equation says that, on average, the states of any system with an attracting set must conform to a gradient flow on surprisal; namely, the negative logarithm of the probability density at nonequilibrium steady state [24,25].

$$\begin{aligned} f(x) &= (Q - \Gamma) \cdot \nabla \Im(x) \\ \Im(x) &= -\ln p(x) \end{aligned} \tag{2}$$

This is the solution to the Fokker-Planck equation when the system has attained nonequilibrium steady-state. It says that the average flow of systemic states has two parts. The first (gradient) component involves surprisal gradients, while the second circulates on iso-probability contours. The gradient flow effectively counters the dispersion due to random fluctuations, such that the probability density does not change over time. See Figure 2 for an intuitive illustration of this solution.

The key move now is to put the Markov blanket back in play. The above equation holds (nontrivially) for the internal, blanket, and external states, where we can drop the appropriate states from the gradient flows, according to the specification of the Markov blanket in Figure 1. In particular, if we just focus on internal and active states—which we will refer to as *autonomous* states—we have the following flows[6] (see p. 17 and pp. 20,21 in [1]).

$$\begin{aligned} f_\alpha(\pi) &= (Q_{\alpha\alpha} - \Gamma_{\alpha\alpha}) \nabla_\alpha \Im(\pi) \\ \alpha &= \{a, \mu\} \\ \pi &= \{s, \alpha\} \end{aligned} \tag{3}$$

thermodynamics to conservative classical mechanics. A technical treatment along these lines can be found in [1] with worked (numerical) examples.

[4] Technically, Equation (1) only holds on the attracting set. However, this does not mean the dynamics collapse to a single point. The attracting manifold would usually support stochastic chaos and dynamical itinerancy—that may look like a succession of transients.

[5] Note that the attracting set is in play throughout the 'lifetime' of any 'thing' because, by definition, a 'thing' has to be at nonequilibrium steady-state. This follows because the Markov blanket is a partition of states at nonequilibrium steady-state.

[6] Note that as in (2) $Q_{\alpha\alpha}$ and $\Gamma_{\alpha\alpha}$ denote antisymmetric and leading diagonal matrices, respectively.

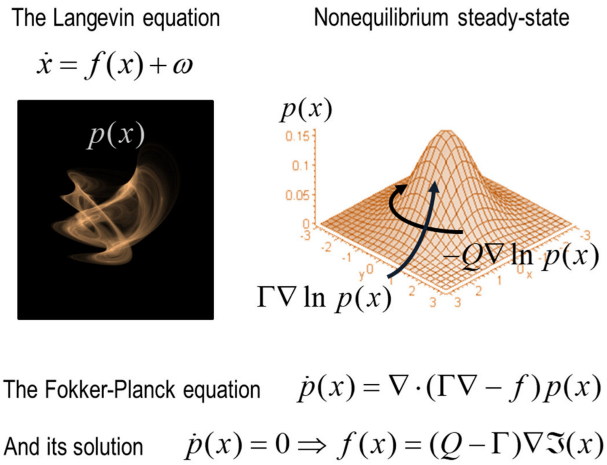
The Langevin equation
$\dot{x} = f(x) + \omega$
$p(x)$
Nonequilibrium steady-state
$p(x)$
0.15
0.1
0.05
0
$\Gamma\nabla \ln p(x)$
$-Q\nabla \ln p(x)$
y
x
The Fokker-Planck equation $\dot{p}(x) = \nabla \cdot (\Gamma\nabla - f)p(x)$
And its solution $\dot{p}(x) = 0 \Rightarrow f(x) = (Q - \Gamma)\nabla\Im(x)$

Figure 2. (*density dynamics and pullback attractors*): This figure illustrates the fundaments of density or ensemble dynamics in random dynamical systems—of the sort described by the Langevin equation. The left panel pictures some arbitrary random attractor (a.k.a., a pullback attractor) that can be thought of in two ways: first, it can be considered as the trajectory of (two) systemic states as they evolve over time. For example, these two states could be the depolarisation and current of a nerve cell, over several minutes. At a larger timescale, this trajectory could reflect your daily routine, getting up in the morning, having a cup of coffee, going to work and so on. It could also represent the slow fluctuations in two meteorological states over the period of a year. The key aspect of this trajectory is that it will—after itinerant wandering and a sufficient period of time—revisit particular regimes of state space. These states constitute the attracting set or pullback attractor. The second interpretation is of a probability density over the states that the system will be found in, when sampled at random. The evolution of the probability density is described by the Fokker-Planck equation. Crucially, when any system has attained nonequilibrium steady state, we know that this density does not change with time. This affords the solution to the Fokker-Planck equation—a solution that means that there is a lawful relationship between the flow of states at any point in state space and the probability density. This solution expresses the flow in terms of gradients of log density or surprisal and the amplitude of random fluctuations. In turn, the nonequilibrium steady-state solution can always be expressed, via the Helmholtz decomposition, in terms of two orthogonal components. One component is a gradient flow that rebuilds probability gradients in a way that is exactly countered by the dispersion of states due to random fluctuations. The other component is a solenoidal or divergence-free flow that circulates on isoprobability contours. These two components are shown in the schematic on the right, in terms of a curl-free gradient flow—that depends only on the amplitude of random fluctuations Γ– and a divergence-free solenoidal flow—that depends upon an antisymmetric matrix *Q*. This example shows the flow around the peak of a probability density, with a Gaussian form. Please see [1,25] for details.

This means anything that can be measured (i.e., a system with a Markov blanket and attracting set) must possess the above gradient flows. In turn, this means that internal and active states will look as if they are trying to minimise exactly the same quantity; namely, the surprisal of states that constitute the thing, particle, or creature. These are the internal states and their Markov blanket; i.e., *particular states*.[7] This means that anything that exists must, in some sense, be self-evidencing [37].

[7] In itself, this is remarkable, in the sense that it captures the essence of many descriptions of adaptive behaviour, ranging from expected utility theory in economics [26–28] through to synergetics and self-organisation [21,29]. See Figure 3. To see how these descriptions follow from the gradient flows in (3), we only have to note that the mechanics of internal and active states can be regarded as *perception* and *action*, where both are in the service of minimising a particular surprisal. This surprisal can be regarded as a cost function from the point of view of engineering and behavioural psychology [30–32]. From the perspective of information theory, surprisal corresponds to self-information, leading to notions such as the principle of minimum redundancy or maximum efficiency [33]. The average value of surprisal is entropy [17]. This means that anything that exists will—appear to—minimise the entropy of its particular states over time [29,34]. In other words, it will appear to

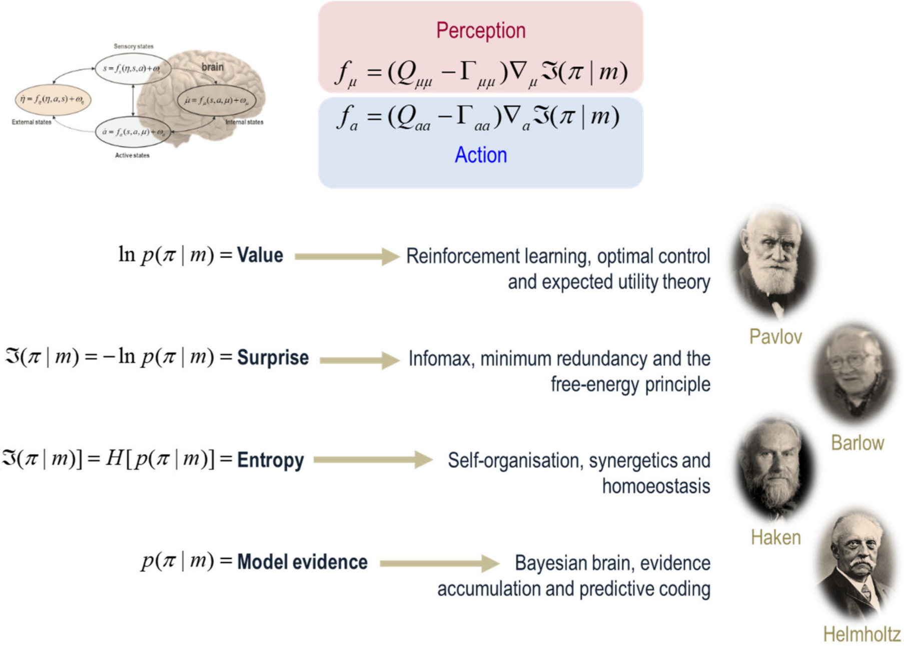
Sensory states
$s = f_s(\eta, s, a) + \omega_s$
brain
$\dot{\eta} = f_\eta(\eta, a, s) + \omega_\eta$
External states
$\dot{\mu} = f_\mu(s, a, \mu) + \omega_\mu$
Internal states
$\dot{a} = f_a(s, a, \mu) + \omega_a$
Active states
Perception
$f_\mu = (Q_{\mu\mu} - \Gamma_{\mu\mu})\nabla_\mu \Im(\pi \mid m)$
$f_a = (Q_{aa} - \Gamma_{aa})\nabla_a \Im(\pi \mid m)$
Action
$\ln p(\pi \mid m) =$ Value
Reinforcement learning, optimal control and expected utility theory
$\Im(\pi \mid m) = -\ln p(\pi \mid m) =$ Surprise
Infomax, minimum redundancy and the free-energy principle
$E[\Im(\pi \mid m)] = H[p(\pi \mid m)] =$ Entropy
Self-organisation, synergetics and homoeostasis
$p(\pi \mid m) =$ Model evidence
Bayesian brain, evidence accumulation and predictive coding
Pavlov
Barlow
Haken
Helmholtz

Figure 3. (*Markov blankets and other formulations*): This schematic illustrates the various interpretations of a gradient flow on surprisal. Recall that the existence of a Markov blanket implies a certain lack of influences among internal, blanket, and external states. At nonequilibrium steady-state, these independencies have an important consequence; internal and active states are the only states that are not influenced by external states, which means their dynamics (i.e., perception and action) are a function of, and only of, particular states; i.e., a particular surprisal.[8] This surprisal has a number of interesting interpretations. Given it is the negative log probability of finding a particle or creature in a particular state, minimising particular surprisal corresponds to maximising the *value* of a particle's state. This interpretation is licensed by the fact that the states with a high probability are, by definition, attracting states. On this view, one can then spin-off an interpretation in terms of reinforcement learning [30], optimal control theory [31] and, in economics, expected utility theory [39]. Indeed, any scheme predicated on the optimisation of some objective function can now be cast in terms of minimising a particular surprisal—in terms of perception and action (i.e., the flow of internal and active states). The minimisation of particular surprisal leads to a series of influential accounts of neuronal dynamics; including the principle of maximum mutual information [40,41], the principles of minimum redundancy and maximum efficiency [33] and—as we will see later—the free energy principle [42]. Crucially, the average or expected surprisal (over time or particular states of being) corresponds to entropy. This means that action and perception look as if they are minimising a particular entropy. The implicit resistance to the second law of thermodynamics leads us to theories of self-organisation, such as synergetics in physics [29,43,44] or homoeostasis in physiology [35,45,46]. Finally, the probability of any particular states given a Markov blanket (*m*) is, on a statistical view, model evidence [18,47]. This means that all the above formulations are internally consistent with things like the Bayesian brain hypothesis, evidence accumulation and predictive coding; most of which inherit from Helmholtz's motion of unconscious inference [48], later unpacked in terms of perception as hypothesis testing in 20th century psychology [49] and machine learning [50]. In short, the very existence of something leads in the natural way to a whole series of optimisation frameworks in the physical and life sciences that lends each a construct validity in relation to the others.

resist the second law of thermodynamics (which is again remarkable, because we are dealing with open systems that are far from equilibrium). From the point of view of a physiologist, this is nothing more than a generalised homoeostasis [35]. Finally, from the point of view of a statistician, the negative surprisal would look exactly the same as Bayesian model evidence [36].

4. Bayesian Mechanics

Thus, if we can describe anything as self-evidencing—in the sense of possessing a dynamics that tries to minimise a particular surprisal—or maximise a particular model evidence, what is the model? It is at this point we get into the realm of inference and Bayesian mechanics, which follows naturally from the density dynamics of the preceding section. The key move here rests upon another fundamental but simple consequence of possessing a Markov blanket.

Technically, the stipulative existence of a Markov blanket means that internal and external states are conditionally independent of each other, when conditioned on blanket states. This has an important consequence. In brief, for every given blanket state there must exist a density over internal states and a density over external states. The former must possess an expectation (i.e., average) or mode (i.e., maximum). This means for every conditional expectation of internal states there must be a conditional density over external states. In short, the mapping between the expected (i.e., average) internal state (for any given blanket state) and a conditional density over external states (i.e., a Bayesian belief about external states) inherits from the conditional independencies that define a Markov blanket. In turn, anything that exists is defined by its Markov blanket. A more formal treatment of this can be found on p. 84 of [1]. See also [3,38] for further discussion.

Therefore, if internal and external states are conditionally independent, then for every given blanket state there is an expected internal state and a conditional probability density over external states. In other words, there must be a one-to-one relationship between the average internal state of a particle (or creature) and a probability density over external states, for every given blanket state.[9] This means that we can express the posterior or conditional density over external states as a probabilistic belief that is parameterized by internal states:

$$\begin{aligned} q_{\boldsymbol{\mu}}(\eta) &= p(\eta|b) = p(\eta|\pi) \\ \boldsymbol{\mu}(b) &\triangleq \operatorname{argmax}_{\mu} p(\mu|b) \end{aligned} \tag{4}$$

On the assumption that the number or dimensionality of internal states is greater than the number of blanket states, the dimensionality of the internal (statistical) manifold—defined by the second equality in (4)—corresponds to the dimensionality of blanket states (which ensures an injective and surjective mapping). This is important because it means there is a subspace (i.e., statistical manifold) of internal states whose dimensionality corresponds to dimensionality of the blanket state (e.g., cardinality of sensory receptors). Heuristically, this means that many external states of affairs can only be represented probabilistically; in a way that depends upon the number of blanket states. Furthermore, the states parameterising this conditional density are conditional expectations; namely, the average internal state, for each blanket state—please see Figure 18 in [1] for a worked (numerical) example.

This is important from a number of perspectives. First, it allows us to interpret the flow of (expected) autonomous states (i.e., action and perception) as a gradient flow on something called variational free energy.[10]

[8] Note that in going from Equation (3) to the equations in Figure 3, we have assumed that the solenoidal coupling (Q) has a block diagonal form. In other words, we are ignoring the solenoidal coupling between internal and active states [9]. The interesting relationship between conditional independence and solenoidal coupling is pursued in a forthcoming submission to Entropy [38].

[9] $\boldsymbol{\mu}(b)$ could also be defined as the expected value of $p(\mu|b)$ which will we approximated by ensemble averages of internal states.

[10] This functional can be expressed in several forms; namely, an expected energy minus the entropy of the variational density, which is equivalent to the self-information associated with particular states (i.e., *surprisal*) plus the KL divergence between the variational and posterior density (i.e., *bound*). In turn, this can be decomposed into the negative log likelihood of particular states (i.e., *accuracy*) and the KL divergence between posterior and prior densities (i.e., *complexity*). In short, variational free energy constitutes a *Lyapunov function* for the expected flow of autonomous states.Variational free energy, like particular surprisal, depends on, and only on, particular states. Without going into technical details, it is sufficient to note that working with the variational free energy resolves many analytic and computational problems of working with surprisal *per se*; especially, if we want to interpret perception in terms of approximate Bayesian inference. It is perhaps sufficient to note that this variational free energy underlies nearly every statistical procedure in the physical and data

$$\begin{aligned} f_\alpha(\pi) &\approx (Q_{\alpha\alpha} - \Gamma_{\alpha\alpha})\nabla_\alpha F(\pi) \\ \\ F(\pi) &\geq \Im(\pi) \\ F(\pi) &\triangleq \underbrace{E_q[\Im(\eta,\pi)]}_{energy} - \underbrace{H[q_\mu(\eta)]}_{entropy} \\ &= \underbrace{\Im(\pi)}_{surprisal} + \underbrace{D[q_\mu(\eta)\|p(\eta|\pi)]}_{bound} \\ &= \underbrace{E_q[\Im(\pi|\eta)]}_{inaccuracy} + \underbrace{D[q_\mu(\eta)\|p(\eta)]}_{complexity} \end{aligned} \tag{5}$$

The second thing that (4) brings to the table is an *information geometry* and attending calculus of beliefs. From now on, we will associate beliefs with the probability density above that is parameterised by (expected) internal states. Note that these beliefs are non-propositional, where 'belief' is used in the sense of 'belief propagation' and 'Bayesian belief updating' that can always be formulated as minimising variational free energy [51,52,58]. To license a description of this conditional density in terms of beliefs, we can now appeal to information geometry [23,59–61].

5. Information Geometry and Beliefs

Information geometry is a formalism that considers the metric or geometrical properties of statistical manifolds. Generally speaking, a collection of points in some arbitrary state space does not, in and of itself, have any geometry or associated notion of distance, e.g., one cannot say whether one point is near another. To equip a space with a geometry, one has to supply something called a metric tensor–such that small displacements in state space can be associated with a metric of distance. For familiar Euclidean spaces, this metric tensor is the identity matrix. In other words, moving one centimetre in this direction means that I have moved a distance of 1 cm. However, generally speaking, metric spaces do not have such a simple tensor form[11]. Provided the metric tensor is symmetrical and positive (for all dimensions of the states in question), the geometry is said to be Riemannian. So, what is special about the Riemannian geometry of statistical manifolds?

A statistical manifold is a special state space, in which the states represent the parameters of a probability distribution. For example, a two-dimensional manifold, whose coordinates are mean and precision, would constitute a statistical manifold for Gaussian distributions. In other words, for every point on the statistical manifold there would be a corresponding Gaussian (bell shaped) probability density. The important thing here is that any statistical manifold is necessarily equipped with a unique metric tensor, known as the Fisher information metric [23,59,62].[12]

$$\begin{aligned} &d\ell^2 = g_{ij} d\mu^i d\mu^j \\ &g(\mu) = \nabla_{\mu'\mu'} D[q_{\mu'}(\eta)\|q_\mu(\eta)]|_{\mu'=\mu} = E_q[\nabla_\mu \ln q_\mu(\eta) \times \nabla_\mu \ln q_\mu(\eta)] \end{aligned} \tag{6}$$

sciences [51–56]. For example, it is the (negative) evidence lower bound used in state of the art (variational autoencoder) deep learning [53,55]. In summary, the variational free energy is always implicitly or explicitly under the hood of any inference process, ranging from simple analyses of variance through to the Bayesian brain [57].

[11] For example, if I set off in a straight line and travelled 40,075 km, I will have moved exactly no distance, because I would have circumnavigated the globe.

[12] The notion of a metric is very general; in the sense that any metric space is defined by the way that it is measured. In the special case of a statistical manifold, the metric is supplied by the way in which probability densities change as we move over the manifold. In this instance, the metric is the Fisher information. Technically, the Fisher information metric can be thought of as an infinitesimal form of the relative entropy (i.e., the Kullback-Leibler divergence between the densities encoded by two infinitesimally close points on the manifold). Specifically, it is the Hessian of the divergence. Heuristically, this means the Fisher information metric scores the number of distinguishable probability densities encountered when moving from one point on the manifold to another.

Here, $d\ell$ is the information length associated with small displacements on the statistical manifold $d\mu = \mu' - \mu$ induced by a probability density $q_\mu(\eta)$. It is not important to understand the details of this metric; other than to note that it must exist. In brief, the distance between two points on the statistical manifold obtains by accumulating the Kullback-Leibler divergence between the probability distributions encoded as we move along a path from one point to another. In other words, the information length scores the number of different probabilistic or belief states encountered in moving from one part of a statistical manifold to another. The path with the smallest length is known as a geodesic. So why is this interesting?

If we return to the independencies induced by the Markov blanket, Equation (4) tells us something fundamental. The (expected) internal states have acquired an information geometry, because they parameterise probabilistic beliefs about external states. This geometry is uniquely supplied by the Fisher information metric specified by the associated beliefs. In short, we now know that there is a unique geometry in some belief space that can be associated with the internal (physical) state of any particle or creature. Furthermore, we also know that the gradient flows describing the dynamics of internal states can be expressed as a gradient flow on a variational free energy functional (i.e., function of the function) *of beliefs*: see (5). All this follows from first principles and yet we have something quite remarkable in hand: if anything exists, its autonomous states will (appear to) be driven by gradient forces established by an information geometry or, more simply, probabilistic beliefs.[13] From (5):

$$\begin{aligned} &f_\alpha(\pi) \approx (Q_{\alpha\alpha} - \Gamma_{\alpha\alpha})\nabla_\alpha F \\ &F(\pi) \equiv F[s, q_\mu(\eta)] \end{aligned} \tag{7}$$

We will call the information geometry that follows from this an *extrinsic* information geometry because it rests upon probabilistic (Bayesian) beliefs about external states. Bayesian beliefs are just conditional probability distributions *that are manifest* in the sense of being encoded by the (internal) states of a physical system. This means it would be perfectly sensible to say that a bacterium has certain Bayesian beliefs about the extracellular milieu—that are encoded by intracellular states. Similarly, in a brain, neuronal activity in the visual cortex parameterizes a Bayesian belief about some visible attribute of the sensorium. Clearly, these kinds of beliefs are not propositional in nature.

Things get even more interesting when we step back and think about the density dynamics of the internal states. Recall from above, that an information geometry is a necessary property of any statistical manifold constituted by parametric states. So, are there any parameters of the probability density over the internal states themselves? The answer here is yes. In fact, these parameters are thermodynamic variables (e.g., pressure) that underwrite thermodynamics or statistical mechanics [62,64]. An important parameter of this kind is time itself. This follows because if we start the internal states from any initial probability density, it will evolve over time to its non-equilibrium steady-state solution. Crucially, this means that we can parameterise the density over internal states with time—and time becomes our statistical manifold. This leads to the challenging intuition, that distance travelled in time can change as we move into the future. In virtue of the existence of the attracting set, the increase in this information length will eventually slow down and stop (as the probability density in the distant future approaches its nonequilibrium steady state)[14]. In turn, the information length furnishes a useful measure of distance from any initial conditions to nonequilibrium steady-state–that has been

[13] In turn, this flow will, in a well-defined metric sense, cause movement in a belief space. This is just a statement of the way things must be—if things exist. Having said this, one is perfectly entitled to describe this sort of sentient dynamics (i.e., the Bayesian mechanics) as being caused by the same forces or gradients that constitute the (Fisher information) metric in (6). This is nothing more than a formal restatement of Johann Friedrich Herbart's "mechanics of the mind"; according to which conscious representations behave like counteracting forces [63].

[14] An intuition here, can be built by considering what you will be doing in a few minutes, as opposed to next year. The difference between the probability over different 'states of being' between the present and in 2 min time is much greater than the corresponding differences between this time next year and this time next year, plus two min.

exploited in characterising self-organisation in random, chaotic dynamical systems [23,62]. We will refer to the accompanying information geometry as an *intrinsic* geometry, because it is intrinsic to the density dynamics of the states *per se*.[15] From our point of view, this means there are two information geometries in play with the following metrics:

$$\begin{aligned} g(\tau) &= \nabla_{\tau'\tau'} D[p_{\tau'}(\mu) \| p_\tau(\mu)]|_{\tau'=\tau} \quad && \text{intrinsic} \\ g(\mu) &= \nabla_{\mu'\mu'} D[q_{\mu'}(\eta) \| q_\mu(\eta)]|_{\mu'=\mu} \quad && \text{extrinsic} \end{aligned} \tag{8}$$

First, there is an intrinsic information geometry inherent in the information length based upon time-dependent probability densities over internal states. This information length characterises the system or creature in terms of itinerant, self-organising density dynamics that forms the basis of statistical mechanics in physics, i.e., a physical, material, or *mechanical* information geometry that is *intrinsic* to the system. At the same time, there is an information geometry in the space of internal states that refers to belief distributions over external states. This is the *extrinsic* information geometry that inherits from the *Markovian* conditions that define, stipulatively, autonomous states (via their Markov blanket). The extrinsic geometry is conjugate to the intrinsic geometry but measures distances between beliefs. Both are measurable, and both supervene on the same Langevin dynamics.

Again, this is not mysterious it is just a mathematical statement of the way things are. What is interesting here is that internal states have a dual aspect information geometry that seems to be related to the dual aspect monism—usually advanced to counter Cartesian (matter and mind) duality. On a simple interpretation, one might associate the information length of internal states with the material behaviour of particles or creatures, while the mindful aspects are naturally associated with the probabilistic beliefs that underwrite the extrinsic information geometry of internal states. However, the existence of a dual aspect information geometry does not, in and of itself, give a system mental states and consciousness, but only computational properties (including probabilistic beliefs). Furthermore, the extrinsic information geometry is ultimately reducible to the intrinsic information geometry (and the other way around), in the sense that there is a necessary link between them cf. [65], pp. 11–13. Still, physical, and computational properties are not identical.[16]

An interesting special case arises if we assume that the conditional beliefs are Gaussian in form (denoted by $\boldsymbol{N}$ in equation (9) below). In this instance, the Fisher information metric becomes the curvature or 'deepness' of free energy minima, which is the same as the precision (i.e., inverse covariance) of the beliefs *per se*.

$$\begin{aligned} q(\mu) &= \textstyle\sum(\mu)^{-1} = \nabla_{\mu\mu} F = -\nabla_{\mu\mu} \ln q(\eta) \\ q(\eta) &= \boldsymbol{N}(\sigma(\mu), \textstyle\sum(\mu)) \end{aligned} \tag{9}$$

In other words, distances in belief space depend upon conditional precision or the confidence ascribed to beliefs about external states of affairs 'out there'. We will return to this interesting case in the conclusion. At the moment, notice that we have a formal way of talking about the 'force of evidence' in moving beliefs and how the degree of movement depends upon conditional precision, confidence, or certainty [67–69].

[15] Another way of thinking about the distinction between the intrinsic and extrinsic information geometries is that the implicit probability distributions are over internal and external states, respectively. This means the intrinsic geometry describes the probabilistic behaviour of internal states, while the extrinsic geometry describes the Bayesian beliefs encoded by internal states *about external states*.

[16] This is also how the following statement could be interpreted: "We are dualists only in asserting that, while the brain is material, the mind is immaterial" [66]. Technically, the link between the intrinsic and extrinsic information geometries follows because any change in internal states implies a conjugate movement on both statistical manifolds. However, these manifolds are formally different: one is a manifold containing parameters of beliefs about external states, while the other is a manifold containing parameters of the probability density over (future) internal states; namely, time (or appropriate statistical parameter apt for describing thermodynamics).

6. A Force to Be Reckoned with

To make all this concrete, it is perfectly permissible to express the gradient flows in terms of forces supplied by the extrinsic, belief-based information geometry. This just requires a specification of the units of the random fluctuations in terms of Boltzmann's constant. This means that we can rewrite (7) in terms of a thermodynamic potential $U(\pi)$ and associated forces $f_m(\pi)$, where, at nonequilibrium steady-state (see pp. 65–67 in [1]):

$$\begin{aligned}
f_\alpha(\pi) &= (\mu_m - Q_m) f_m(\pi) \\
&= (Q_m - \mu_m)\nabla U(\pi) \\
&\approx (Q_{\alpha\alpha} - \Gamma_{\alpha\alpha})\nabla_\alpha F(\pi) \\
\\
f_m(\pi) &\triangleq -\nabla U(\pi) \\
Q_{\alpha\alpha} &\triangleq k_B T \cdot Q_m \\
\Gamma_{\alpha\alpha} &\triangleq k_B T \cdot \mu_m \\
U(\pi) &\triangleq k_B T \cdot \Im(\pi) \\
&\approx k_B T \cdot F(\pi)
\end{aligned} \tag{10}$$

The last equality is known as the Einstein–Smoluchowski relation, where μ_m is a mobility coefficient. This means, we have factorised the amplitude of random fluctuations $\Gamma_{\alpha\alpha} = \mu_m k_B T$ into mobility and temperature [12]. Nothing has changed here. All we have done is assign units of measurement to the amplitude of random fluctuations, so that we can interpret the ensuing flow as responding to a force, which can be interpreted as a gradient established by a thermodynamic potential. This thermodynamic potential is just (scaled) surprisal or our free energy functional of beliefs.

These equalities cast the appearance of Cartesian duality in pleasingly transparent terms. The forces that engender our physical dynamics can either be expressed as thermodynamic forces or as self-evidencing; in virtue of the extrinsic information geometry supplied by variational free energy. Mathematically, this duality arises from the fact that the surprisal and variational free energy are conjugate: one rests upon the probability of particular states, while the other is a functional of blanket states and beliefs that are parameterised by internal states. They are conjugate in that they refer to probability densities over conditionally independent (i.e., orthogonal) states; namely, internal and external states.

The point here is that there is no difficulty in moving between descriptions afforded by statistical thermodynamics and self-evidencing (i.e., minimising variational free energy). On this reading, variational free energy is a feature of an extrinsic information geometry induced by beliefs encoded by internal states that have an intrinsic information geometry. This free energy has gradients that exert forces on internal states so that they come to parameterise new beliefs. These new beliefs depend upon blanket (e.g., sensory) states; thereby furnishing a mathematical image of perception. Furthermore, the same Bayesian mechanics applies to active states that change external states—and thereby mediate action upon the world. So, is there anything more to the story?

7. Active Inference and the Future

Active inference will become a key aspect of the arguments below, when thinking about different kinds of generative models; specifically, generative models of the consequences of action. On the above arguments, anything (that exists in virtue of possessing a Markov blanket) can be cast as performing some elemental form of inference—and possessing an implicit generative model. However, not all generative models are equal; in the sense that no two things are the same. Later, we will look at special kinds of generative models that underwrite active inference.

Above, we introduced variational free energy as an expression of particular surprisal. This variational form is a functional of sensory states and a conditional density or belief distribution encoded by internal states. However, the variational free energy also depends upon the surprisal of joint particular and external states, $\Im(\eta, \pi) \equiv -\ln p(\eta, \pi)$, see (5). On a statistical view, the corresponding nonequilibrium steady-state density $p(\eta, \pi)$ is known as a *generative model*. In other words, it constitutes a probabilistic specification of how external and particular states manifest. It is this generative model that licenses an interpretation of particular surprisal in terms of Bayesian mechanics and self-evidencing [37]. So, what does this mean for our formulation of beliefs and intention?

Note that we can always describe the dynamics of internal states in terms of a gradient flow on variational free energy. This means that the dynamical architecture of any particle or creature can also be expressed as a functional of some generative model that, in some sense, must be isomorphic with the nonequilibrium steady-state density. This has some interesting implications: from the point of view of self-organisation, it tells us immediately that if we interpret the action of a particle or creature in terms of self-evidencing, it says that the implicit generative model—which supplies the forces that change internal and active (i.e., autonomous) states—must be a sufficiently good model of systemic states. This is exactly the good regulator theory that emerged in the formulations of self-organisation at the inception of cybernetics [45,70].[17]

8. Active Inference and the Path Integral Formulation

We will first preview, heuristically, the final argument in this essay. Because active states depend upon internal states (and the beliefs that they parameterise)—but active states do not depend upon external states—it will look as if particles or creatures are acting on the basis of their beliefs about external states. Furthermore, if a particle or creature acts in a dextrous, precise and adaptive way to fluctuations in its blanket states, it will look as if it is acting to minimise its particular surprisal (or variational free energy). In other words, it will look as if it is trying to minimise the surprisal, expected following an action. This means, it would look as if it is behaving to minimise expected surprisal or self-information, which is uncertainty or its particular entropy.

Anthropomorphically, a creature will therefore (appear to) have beliefs about the consequences of its action, which means it must have beliefs about the future. So how far into the future? One can formalise a response to this question by turning to the path integral formulation of random dynamical systems [12,13,77,78]. In this formulation, we are not concerned with the probability density over states but rather over trajectories or sequences of states. Specifically, we are interested in the probability of trajectories of autonomous states, often referred to as 'policies' in the optimal control literature [32]. So, what can one say about the probability of different courses of action in the future?

We can now turn to the information length associated with the evolution of systemic states to answer this question (for a more detailed treatment, see pp. 86–88 in [1]). Recall from above, that the information length reflects the accumulated changes in probability densities as time progresses. If a system attains nonequilibrium steady state after a period of time, then the information length asymptotes to the distance between the initial (particular) state and the final (steady) state. This means that we can characterise a certain kind of particle (or creature) that returns to steady state in terms of the (critical) time $\boldsymbol{\tau}$ it takes for the information length to stop increasing:

$$d\ell(\boldsymbol{\tau}) \approx 0 \Leftrightarrow D[q_{\boldsymbol{\tau}}(\eta_{\boldsymbol{\tau}}, \pi_{\boldsymbol{\tau}}) \| p(\eta_{\boldsymbol{\tau}}, \pi_{\boldsymbol{\tau}})] \approx 0 \tag{11}$$

[17] There are many interesting issues here. For example, it means that the intrinsic anatomy and dynamics (i.e., physiology) of internal states must, in some way, recapitulate the dynamical or causal structure of the outside world [71–73]. There are many examples of this. One celebrated example is the segregation of the brain into ventral ('what') and dorsal ('where') streams [74] that may reflect the statistical independence between 'what' and 'where'. For example, knowing what something is does not, on average, tell me where it is. Another interesting example is that it should be possible to discern the physical structure of systemic states by looking at the brain of any creature. For example, if I looked at my brain, I would immediately guess that my embodied world had a bilateral symmetry, while if I looked at the brain of an octopus, I might guess that it's embodied world had a rotational symmetry [75]. These examples emphasise the 'body as world' in a non-radical enactive or embodied sense [76]. This begs the question of how the generative model—said to be entailed by internal states—shapes perception and, crucially, action.

The probability density $q_\tau(\eta_\tau, \pi_\tau)$ is the *predictive density* over hidden and sensory states, conditioned upon the initial state of the particle and subsequent trajectory of autonomous states. In brief, particles with a short critical time[18] will, effectively, converge to nonequilibrium steady-state quickly and show a simple self-organisation (e.g., the Aplysia gill and siphon withdrawal reflex) mathematically, these sorts of particles quickly 'forget' their initial conditions. Conversely, particles with a long critical time will exhibit itinerant density dynamics (e.g., you and me). Particles like you and me 'remember' our initial conditions and look as if we are pursuing long-term plans.

Convergence to nonequilibrium steady state in the future allows us to relate the surprisal of a trajectory of autonomous states (i.e., a policy) to the variational free energy expected under the predictive density above:

$$\begin{aligned}
G(\boldsymbol{\alpha}[\boldsymbol{\tau}]) &\approx \mathcal{A}(\boldsymbol{\alpha}[\boldsymbol{\tau}]|\pi_0) \\
G(\alpha[\tau]) &\triangleq \underbrace{E_{q_\tau}[\Im(\eta_\tau, \pi_\tau)]}_{energy} - \underbrace{H[q_\tau(\eta_\tau|\pi_\tau)]}_{entropy} \\
&= \underbrace{E_{q_\tau}[\Im(\pi_\tau|\eta_\tau)]}_{ambiguity} + \underbrace{D[q_\tau(\eta_\tau|\pi_\tau)||p(\eta_\tau)]}_{risk}
\end{aligned} \tag{12}$$

The expected free energy in (12) has been formulated to emphasise the formal correspondence with variational free energy in (5): where the complexity and accuracy terms become *risk* (i.e., expected complexity) and *ambiguity* (i.e., expected inaccuracy). This path integral formulation says that if the probability density over systemic states has converged to nonequilibrium steady state after some critical time, then there can be no further increase in information length. At this point, the probability of an autonomous path into the future becomes the variational free energy the agent expects to encounter.

The equality in (11) is a little abstract but has some clear homologues in stochastic thermodynamics (in the form of integral fluctuation theorems) [12,79]. Here, it tells us something rather interesting. It means that creatures that have an adaptive response to changes in their external milieu will look as if they are selecting their long-term actions on the basis of an expected free energy. Crucially, this free energy is based upon a generative model that must extend at least to a (critical) time in the future when nonequilibrium steady state is restored. Conversely, if certain kinds of creatures select their actions on the basis of minimising expected free energy, they will respond adaptively to changes in external states.

This formulation offers a description of different kinds of particles or creatures quantified by their critical time or temporal depth in (11). For example, if a certain kind of particle (e.g., a trial or protozoan) has a short temporal horizon or information length, it will respond quickly and reflexively to any perturbations—for as long as it exists. Conversely, creatures like us (e.g., politicians and pontiffs) may be characterised by deep generative models that see far into the future; enabling a move from homoeostasis to allostasis and, effectively, the capacity to select courses of action that consider long term consequences [80–82]. Given that the imperative for this action selection is to minimise expected free energy (i.e., expected surprisal or uncertainty), we now have a plausible description of intentional behaviour that will, to all intents and purposes, look like uncertainty resolving, information seeking, epistemic foraging [26,81,83–90]. Alternatively, on a more (millennial) Gibsonian view, action selection responds to long-term epistemic affordances [91–93].

This temporal depth may distinguish between different kinds of sentient particles. Again, all of this follows in a relatively straightforward way from information theory and statistical physics. Furthermore, the equations above can be used to simulate perception and intentional behaviour. To illustrate the difference between short term (shallow) inference based upon Equation (5) and

[18] Note that time here does not refer to clock or universal time, it is the time since an initial (i.e., known) state at any point in a systems history. This enables the itinerancy of nonequilibrium steady-state dynamics to be associated with the number of probabilistic configurations a system will pass through, over time, when prepared in some initial state.

long-term (deep) active inference based upon Equation (12), we provide two examples in Figures 4 and 5. The first uses simulated handwriting that is elicited purely on the basis of reflexive responses prescribed by a dynamic generative model (i.e., a pattern generator), while the second calls on the notion of epistemic affordance by simulating saccadic searches and active vision. In the present thesis, simulations of the second sort of active inference may offer a better account of intentional behaviour; namely, beliefs about the consequences of action and subsequent action selection.

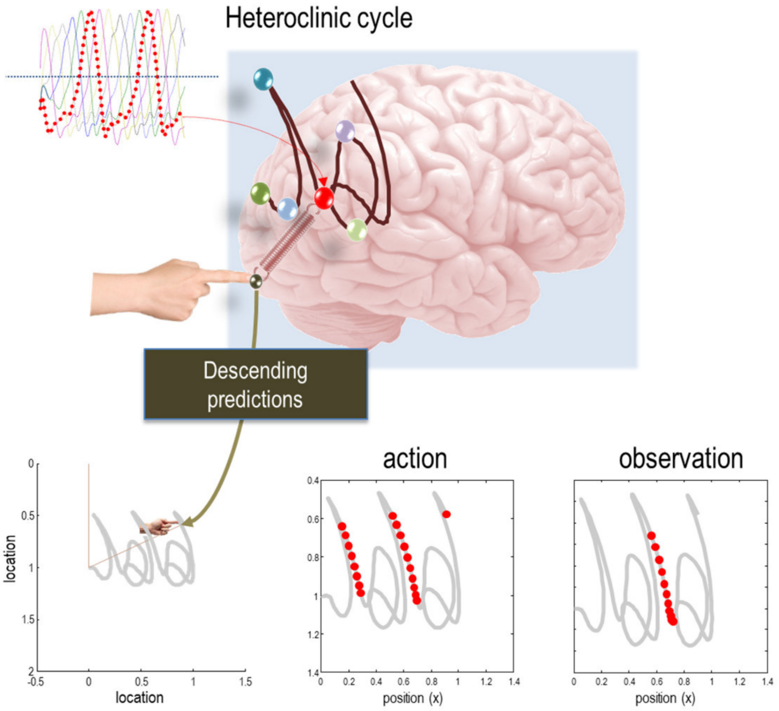

Figure 4. (*autonomous movement*). This figure shows the results of simulating active inference (i.e., writing), in terms of conditional expectations about hidden states of the world, consequent predictions about sensory input and the ensuing behaviour. The autonomous dynamics that underlie this behaviour rest upon prior expectations about states with Lotka-Volterra dynamics (c.f., a central pattern generator): these are the six (arbitrarily) coloured lines in the upper left panel. In this generative model, each state is associated with a location in Euclidean space that attracts the agent's finger. In effect, the internal states then supply predictions of what sensory states should register, if these prior beliefs were true. Active states try to suppress the ensuing prediction error (i.e., sensory surprisal) by reflexively fulfilling expected changes in angular velocity, through exerting forces on the agent's joints (not shown). The subsequent movement of the arm is traced out in the lower left panel. This trajectory has been plotted in a moving frame of reference, so that it looks like synthetic handwriting (e.g., a succession of 'j' and 'a' letters). The lower left panels show the activity of one (the fourth attractor) conditional expectation under 'action', and 'action-observation'. During action, sensory states register both the visual and proprioceptive consequences of movement, while under action observation, only visual sensations are available—as if the agent was watching another agent. The red dots correspond to the time bins during which this state exceeded an amplitude threshold of two arbitrary units. They key thing to note here is that this unit responds preferentially when, and only when, the motor trajectory produces a down-stroke, but not an up-stroke. Please see [94] for further details. Furthermore, with a slight delay, this internal state responds during action and action observation. From a biological perspective, this is interesting because it speaks to an empirical phenomena known as mirror neuron activity [95–97].

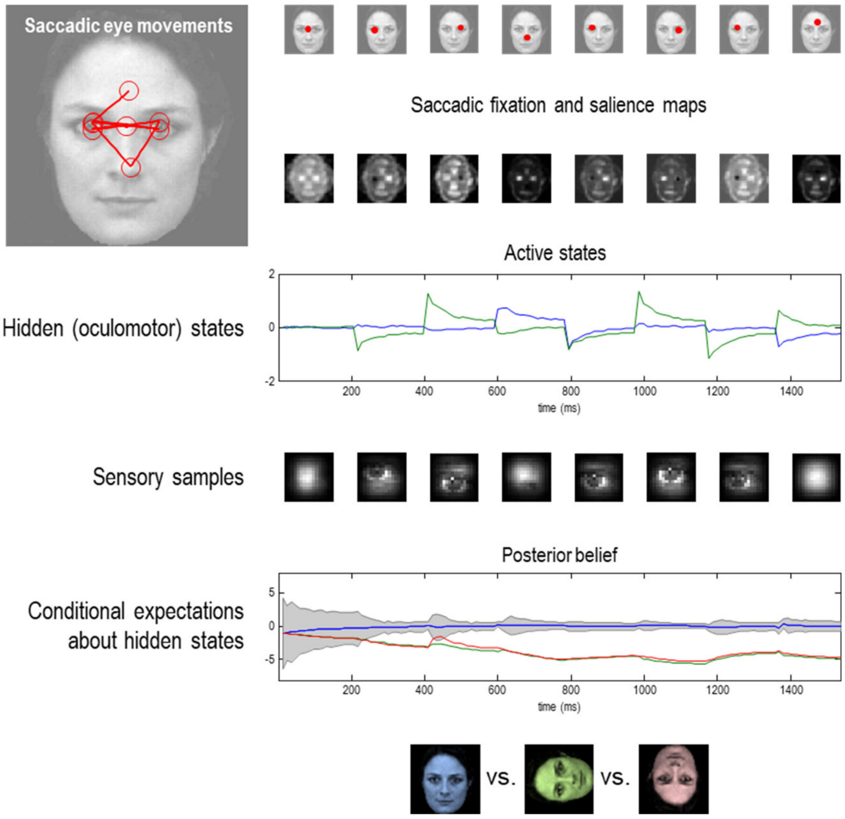

Figure 5. (*epistemic foraging*). This figure shows the results of a simulation in which a face was presented to an agent, whose responses were simulated by selecting active states that minimise expected free energy following an eye movement. The agent had three internal images or hypotheses about the stimuli she might sample (an upright face, and inverted face and a rotated face). The agent was presented with an upright face and her conditional expectations were evaluated over 16 (12 ms.) time bins, until the next saccade was emitted. This was repeated for eight saccades. The ensuing eye movements are shown as red dots at the end of each saccade in the upper row. The corresponding sequence of eye movements is shown in the insert on the upper left, where the red circles correspond roughly to the proportion of the visual image sampled. These saccades are driven by prior beliefs about the direction of gaze based upon the saliency maps in the second row. These saliency maps are the expected free energy as a function of policies; namely, where to look next. Note that these maps change with successive saccades as posterior beliefs about external states, including the stimulus, become progressively more precise or confident. Note also that salience is depleted in locations that were foveated in the previous saccade because these locations no longer have an epistemic affordance (i.e., the ability to reduce uncertainty or expected free energy). This is a nice illustration of a ubiquitous phenomenon, known as inhibition of return. Oculomotor responses are shown in the third row in terms of the two hidden oculomotor states, corresponding to vertical and horizontal eye movements. The associated portions of the image sampled (at the end of each saccade) are shown in the fourth row. The final two rows show the posterior beliefs in terms of their sufficient statistics and the stimulus categories, respectively. The posterior beliefs are plotted here in terms of conditional expectations and the 90% confidence interval about the true stimulus. The key thing to note here is that the expectation about the true stimulus supervenes over its competing expectations and, as a result, conditional confidence about the stimulus category increases (the confidence intervals shrink to the expectation). This illustrates the nature of evidence accumulation when selecting a hypothesis or percept the best explains sensory data. Within saccade accumulation is evident even during the initial fixation with further stepwise decreases in uncertainty as salient information is sampled at successive saccades. Please see [98] for further details.

9. Markovian Monism

Above, we have shown that a duality—between two ways in which states of a system can be conceived of—already arises at a very fundamental level; namely, for all systems that possess a Markov blanket. Their internal states can both be associated with an intrinsic and with an extrinsic information geometry. What metaphysical implication does this have? Does it follow that all systems with a Markov blanket have a mind (because they have probabilistic beliefs about external states)? Are such systems conscious? The formalism itself does not answer these questions: different metaphysical interpretations of the existence of a dual information geometry are possible. In fact, one might ask whether it has any metaphysical significance whatsoever. For the existence of an extrinsic information geometry only means that one *can* map internal states to conditional probability distributions (over external states, given blanket states). It does not mean that the resulting descriptions refer to entities that actually exist (just as we can ascribe to a lectern the propositional belief that the best way to persist is to do nothing—which does not mean that the lectern actually has a propositional belief; see [99]).

Hence, any metaphysical conclusions must be drawn with care. In what follows, we will first argue that the formalism speaks in favour of monistic views—if we assume that the existence of an extrinsic information geometry has any relevance for understanding the mind and consciousness in the first place. After that, we will discuss different interpretations of the dual perspective afforded by the two information geometries: panprotopsychism, neutral monism, dual-aspect theories, and physicalism. We will argue that physicalism provides the most plausible interpretation. However, we acknowledge that competing interpretations cannot conclusively be ruled out. Hence, we dub the resulting view 'Markovian monism'. Markovian monism consists of two claims: (1) Fundamentally, there is only one type of thing and only one type of irreducible property (this is why it is a Markovian *monism*). (2) All systems possessing a Markov blanket have properties that are relevant for understanding the mind and consciousness: if such systems have mental properties, then they have them partly by virtue of possessing a Markov blanket (this is why it is a *Markovian* monism).

Why do we rule out dualistic interpretations of the dual information geometry? First, note that dualism is still consistent with the existence of an extrinsic information geometry. However, consider any properties that a system has by virtue of the fact that its internal states encode probability distributions over external states. Since the dynamics that can be described with reference to these properties can equivalently be described without regarding internal states as representations of probability distributions, there is a sense in which both perspectives are reducible to one another. Hence, the dual information geometry itself does not entail property dualism. Therefore, if one believes that there are irreducible mental properties, one has to posit them in addition to, and largely independently of, properties entailed by the existence of an extrinsic information geometry. But this means that mental properties will not be instantiated (partly) by virtue of the existence of a Markov blanket (contradicting claim (2) above). In other words, dualism is more or less orthogonal to the formal treatment.

However, we do believe that the existence of an extrinsic information geometry tells us something interesting about the origin of minds and consciousness. Under the assumption that properties entailed by the existence of a Markov blanket are relevant to understanding mental properties, we therefore have to reject dualism. This still leaves different metaphysical options open.

10. Markovian Monism as Panprotopsychism?

According to panpsychism, mental properties are fundamental non-physical properties and are instantiated by all micro-level entities. Hence, this amounts to a form of property dualism, which we ruled out above. Note, again, that dualism is compatible with the formal treatment presented here, but it would not be an interpretation in which properties entailed by the existence of a Markov blanket have any explanatory relevance to the existence of minds and consciousness—because panpsychism already presupposes mentality as a fundamental part of reality.

However, there is a variant of panpsychism, viz. panprotopsychism, that could, in principle, be described as a Markovian monism. In short, panprotopsychism is "the view that fundamental entities are *proto-conscious*, that is, that they have certain special properties that are precursors to consciousness and that can collectively constitute consciousness in larger systems." [100], p. 259. These special, non-structural properties are *protophenomenal* properties that are not identical to (micro-)physical properties (otherwise, even physicalism could be considered as a form of panprotopsychism, [100], p. 260). There is nothing it is like to be a system that has just a single protophenomenal property. However, if a system displays a sufficiently large number of protophenomenal properties, or if they are arranged in the right way, then the system will also have phenomenal properties (which are constituted by collections of protophenomenal properties).

From the point of view of Markovian monism, one could identify properties entailed by the existence of a Markov blanket with protophenomenal properties. An example is the property of encoding a conditional probability distribution over external states. However, it is unclear to us to what extent this could be regarded as a non-structural property. Furthermore, a robust version of panprotopsychism would have to presuppose that all systems with a Markov blanket actually represent probability distributions—as opposed to just being systems that *can be described as if* they represented such distributions. Below, we will suggest that a realist interpretation of descriptions afforded by the extrinsic information geometry should be contingent on further conditions. This is why we would not interpret Markovian monism as a version of panprotopsychism.

11. Markovian Monism as Neutral Monism?

Neutral monism is normally read as a family of views; according to which the fundamental layer of reality consists of ontologically neutral entities. Different versions of the theory make different claims about the sense in which basic entities are neutral (see [101], who lists five different options). The most popular options seem to be views according to which the basic entities are (a) intrinsically *neither* mental nor physical or (b) intrinsically *both* mental and physical.

A great advantage of neutral monism is that it solves the mind-body problem without postulating two basic types of entity (mental and physical)—the significance of this is that worries about psycho-physical interaction (that plagued Cartesian dualism) disappear. The only causal interaction in question involves neutral entities (however, the problem of mental causation may reappear, in the sense that macro-level mental properties may still be causally irrelevant, see [102], pp. 33–34).

Markovian monism could be specified as a version of neutral monism in which basic entities are intrinsically neither mental nor physical. There are two conjugate ways in which things that exist can be described: either from the perspective of the intrinsic information geometry or from the perspective of the extrinsic information geometry. Under the assumption that neither perspective is privileged, one would have to conclude that reality is, fundamentally, ontologically neutral.

However, this would also presuppose a realist interpretation of descriptions in terms of the extrinsic information geometry (i.e., one would have to assume that all systems with a Markov blanket actually represent probability distributions and perform computations). Furthermore, it would have the consequence that even relatively simply systems, such as single-cell organisms, would have a mind (as suggested by [103]). For these reasons, we would not interpret Markovian monism as a version of neutral monism.

12. Markovian Monism as a Dual-Aspect Theory?

Dual-aspect monism is the position that reality has two aspects: a mental and a physical aspect. Dual-aspect monism is very similar to neutral monism. Depending on how it is defined, it may even collapse into neutral monism (or into panpsychism, see [104], p. 366). For instance, if dual-aspect monism is defined as the view that reality is, at a fundamental level, both physical and mental, then this comes extremely close to the view that basic entities are intrinsically both mental and physical—and hence to a version of neutral monism (see [101], sec. 8.3).

Furthermore, if the *aspect* in 'dual-aspect' is interpreted in terms of properties, such that basic entities have both mental and physical properties (as suggested by [105], p. 46), then dual-aspect theory becomes a form of property dualism—which we ruled out above.

There are versions of dual-aspect theory that explicitly refrain from defining the dual aspect in terms of property dualism (see, e.g., [106], pp. 339,342). Markovian monism is similar to dual-aspect monism (cf. [107], pp. 220–221), in that it entails that one and the same thing (i.e., internal states of a system possessing a Markov blanket) can be viewed from two perspectives. Internal states can either be regarded as states of a random dynamical system; or they can be viewed as the parameters of a probability distribution (i.e., probabilistic belief). In order to count as a dual-aspect monism, these two perspectives would have to be mutually irreducible (see [106], p. 46; [105], p. 341), we are sceptical that this would be a coherent interpretation of the dual information geometry.

As with the other two interpretations discussed above, an interpretation in terms of a dual-aspect monism would presuppose a realist view on descriptions in terms of the extrinsic information geometry. Furthermore, just as the interpretation in terms of neutral monism, it would entail that single-cell organisms have a mind. In what follows, we will sketch how Markovian monism can ground versions of reductive materialism. This physicalist interpretation of Markovian monism is the one we favour—although we admit that other interpretations cannot conclusively be ruled out.

13. Markovian Monism as Reductive Materialism

Here is what we believe is the most coherent interpretation of the formal treatment. The fact that one can associate two information geometries with systems possessing a Markov blanket reveals a continuity between simple, non-conscious systems and more complex, conscious systems such as human beings: due to the extrinsic information geometry, simple systems can be described *as if* they had beliefs about external states. For conscious systems, the perspective—afforded by such 'as if' descriptions—acquires a special status, because it will typically abstract away from many of the details inherent in the mechanistic perspective. For instance, the probabilistic beliefs ascribed to explain cognitive phenomena are typically assumed to be represented by the average activities of neuronal *populations*, which means that any differences between populations with the same average properties will be irrelevant from the perspective of the extrinsic information geometry. This squares well with the idea that causation (including mental causation) is a macroscopic phenomenon [108,109]. At the same time, these macrostates are always grounded in more fine-grained physical states, and their properties can be reductively explained in terms of physical properties.

Furthermore, the computational properties ascribed to conscious systems will be more numerous and more complex than those ascribed to non-conscious systems. There are no additional, non-reducible properties, which are necessary to explain the mind and consciousness; between some non-conscious and conscious systems, there is only a gradual difference. This entails that consciousness is a vague concept; i.e., there will be borderline cases in which the concept cannot unequivocally be applied.

In particular, this proposal rests upon a distinction between temporally deep and shallow generative models, that accompanies the distinction between conscious and unconscious inference. This distinction is vague, in the sense that any generative model of dynamics has, to a certain extent, temporal depth. For example, predictive coding, homoeostasis and thermostats can all be articulated in terms of perceptual control [46,110,111] and a reflexive form of active inference using generative models based upon differential equations. The fact that a generative model entails differential equations means that there is some inference over time. The distinction between deep and shallow then becomes a quantitative issue. Perhaps a better distinction would be between generative models that entertain a single trajectory into the future, versus multiple (counterfactual action dependent) trajectories that incur a selection problem; namely, choosing an action or planning.

Construing consciousness as a vague concept may even have relevance for the meta-problem of consciousness [112]; i.e., the problem of explaining why it seems (to many) that a physical duplicate of a conscious creature could be non-conscious. Although solving the meta-problem is not the aim of this paper,[19] we can at least contribute to an explanation: as noted above, our interpretation of Markovian monism entails that there is only a gradual difference between some non-conscious and conscious systems, and that consciousness is a vague concept. So, when people claim they can imagine a physical duplicate that is unconscious, they may in fact imagine not a complete duplicate, but a system that differs in (seemingly non-significant ways) from a conscious system. As an analogy, consider a heap of sand. A heap of sand is constituted by grains of sand. But, one could object, a heap of sand cannot be *just* a collection of grains of sand, because I can imagine a collection of grains of sand (say, three grains) that does not count as a heap. Hence, there seems to be a crucial difference between collections of grains of sand and *heaps* of sand—just adding a grain of sand to something that is not a heap does not turn it into a heap. Similarly, just adding a bit more structure and function to a non-conscious system does not turn it into a conscious system. Hence, it would seem as if consciousness requires more than just the right structure and functions, and the hard problem arises. But if consciousness is a vague concept (as suggested by our interpretation of Markovian monism), then the right structure and functions can be metaphysically sufficient for consciousness, even if adding just a bit of structure and function to any uncontroversially non-conscious system does not make it conscious.

Furthermore, the very existence of the meta-problem implies a certain kind of Bayesian belief that entails some puzzlement about 'our capacity to have subjective experiences of a quantitative sort'. But 'qualia' and accompanying 'puzzlement' are just Bayesian beliefs that imply an extrinsic information geometry. So, is there anything special about Bayesian beliefs about Bayesian beliefs? The answer is yes: beliefs about beliefs (in a mathematical sense) require a hierarchical generative model. But, a hierarchical generative model requires hierarchically deployed Markov blankets to introduce the necessary conditional independencies (which make it hierarchical). We therefore conclude that phenomenally conscious systems for which a hard problem exists must possess a certain kind of statistical structure; namely, Markov blankets within Markov blankets [4].

Although we believe that there are only gradual differences between non-conscious and conscious systems, if one merely considers the probabilistic beliefs that can be ascribed to such systems, there are still categorical differences that can be described in terms of more high-level properties, such as intentionality and computation (note that this does not imply a "phase transition" between unconscious and conscious systems).

In particular, one can make a threefold distinction between (i) systems that behave only 'as if' they implemented computations over probabilistic beliefs, (ii) systems for which the "computational stance" [114] provides added explanatory value, and (iii) systems that can usefully be described as not only computational, but also as representational systems.[20] While not speaking against a continuity between life and mind [103], this threefold distinction could be used to establish a discontinuity between life and consciousness.

Specifying the difference between (i) and (ii) would require defending a particular account of computation, which is beyond the scope of this paper. The step from (ii) to (iii)—i.e., from a computational to a representational system—requires ascribing content to internal states of the system. Representationalist interpretations of the free-energy principle refer to computations that are implemented (or approximated) by systems that minimize free energy (see e.g. [117], pp. 571–572).

[19] For a recent response to the meta-problem, see [113]. The authors identify three features that are central for conscious systems: (1) depth (including temporal depth); (2) responsiveness to "interoceptive information concerning the agent's own bodily states and self-predicted patterns of future reaction"; (3) "the capacity to keep inferred, highly certain mid-level sensory re-codings fixed while imaginatively varying top-level beliefs." ([113], p. 31). Note that these are all *gradual* features, in line with the view that consciousness is a vague concept.

[20] Note that this treatment departs from the classical conception of computation, according to which there is "no computation without representation" [115]. According to many proponents of this view, representation is prior to computation. In other words, a physical system only performs a computation if it has genuine representational states. Mechanistic conceptions of computation and representation reject this view. Accordingly, physical systems can perform computations without representation (see [116].).

Such computations are defined with respect to exactly the types of probabilistic beliefs encoded by systems with an extrinsic information geometry.

Although Markovian monism, interpreted as a form of reductive materialism, is not a theory of consciousness, it refers to properties that may ground mental properties (including phenomenal properties). As such, it provides a foundation for various physicalist approaches to consciousness and the mind, most notably representationalism and (computational) functionalism.[21]

14. Consciousness and Integrated Information

There have been previous attempts to use information theory to describe conscious processing. Perhaps the most notable is integrated information theory [118,119]. One might ask about the relationship between the free energy principle (FEP) and integrated information theory (IIT)? At the time of writing, there is a gap between these theoretical approaches. First, the FEP is a 'first principle' account that uses variational principles to build upon the Langevin formulation of random dynamical systems. In contrast, IIT is an 'axiomatic' approach that starts with some assumptions about what information processing must look like to be a contender for explaining conscious experience. The formal distinction between the FEP and IIT is that the free energy principle is articulated in terms of probabilistic beliefs *about* some (external) thing, while integrated information theory deals with probability distributions *over* the states of some system. In other words, IIT does not commit to an extrinsic information geometry (the "geometry of integrated information" is an intrinsic information geometry, see [120]). This is not necessarily a problem, in so far as IIT offers a normative (i.e., measurable, in principle) description of systems that comply with axioms, which inherit pre-theoretical notions of consciousness. On the other hand, both the free energy FEP and IIT can be cast in terms of information theory and in particular functionals (e.g., variational free energy and 'phi'). Furthermore, they both rest upon partitions (e.g., Markov blankets that separate internal from external states and complexes that constitute conscious entities and can be distinguished from other entities). This speaks to the possibility of, at least, numerical analyses that show that minimising variational free energy maximises 'phi' and *vice versa*.

Although integrated information theory does not commit to a Markovian information geometry of experience (i.e., conscious or unconscious inference about something), it is possible to establish some kind of construct validity between the FEP and IIT in terms of the axioms upon which IIT is predicated. In other words, one can establish—at least heuristically—that the FEP features the essential properties of experience that constitute the axiomatic basis of IIT. There are five axioms; namely, intrinsic existence, composition, information, integration and exclusion. In brief:

- **Intrinsic existence**—consciousness exists: each experience is actual and exists from its own intrinsic perspective. This is a necessary consequence of Bayesian mechanics under the free energy principle because the dynamics underlying inference are physically realised and are, by construction, intrinsic in the sense of pertaining to internal states.
- **Composition—**consciousness is structured: with multiple phenomenal distinctions. Again, this is a necessary aspect of Bayesian mechanics, which is defined in terms of the structure implicit in conditional independencies. Indeed, from a statistical perspective, minimising variational free energy is synonymous with structure learning [59,121,122].
- **Information**—consciousness is unique: each experience is the particular way it is, thereby differing from other possible experiences (i.e., differentiation). Again, this is a fundament of Bayesian mechanics under the free energy principle; in the sense that any information geometry implies a particular point on a statistical manifold (of internal or intrinsic states) maps to a particular

[21] Of course, functionalism itself is ontologically neutral, in that it identifies mental states with functional states that could be realised by different substrates.

probability or belief state with phenomenal support (i.e., an extrinsic belief distribution over the external states).

- **Integration**—consciousness is unified: each experience is irreducible to disjoint subsets of phenomenal distinctions (i.e., integration). Again, this is a necessary aspect of the information geometry that underwrites the free energy principle. This follows because for each point on the internal statistical manifold, there is a single probabilistic belief (i.e., variational density). In other words, although this density could be very high dimensional, it is just one probabilistic belief that cannot be dissembled or reduced. Another aspect of the axiom of integration is that "every part of the system has both causes and effects within the rest of the system" ([123], p. 3). This is true for systems possessing a Markov blanket, because the gradient flows of internal states (and associated belief updating) are, by definition, conditionally dependent.
- **Exclusion**—consciousness is definite: each experience is characterised by what it is (neither less no more than) and flows at the speed it flows (neither faster nor slower). Again, this is a necessary consequence of the density dynamics that underwrites the free energy principle. In other words, flows on the extrinsic (statistical) manifold are unique and entail particular probabilistic beliefs about external states, i.e., precise beliefs about being in a particular (external) state but not another. Furthermore, each probabilistic belief has its own sufficient statistics that exclude the possibility of other sufficient statistics. For example, beliefs about my temperature can be stipulated with an expectation that my temperature is such and such. This precludes the possibility that I expect to my temperature to be anything else. In contrast to the exclusion axiom, however, the existence of a Markov blanket at one spatiotemporal scale does not exclude the existence of (e.g., nested) Markov blankets at other spatiotemporal scales.

In summary, on an informal review, the information geometry and density dynamics implied by Markov blankets appear to possess the qualities—or conform to the essential criteria—that constitute the axiomatic basis of integrated information theory.

The important result of this section, from our perspective, is that at least some properties associated with consciousness are already entailed by Bayesian mechanics under the free energy principle. This supports the (speculative) hypothesis that adding further constraints on generative models—entailed by systems possessing a Markov blanket—might enable us to say which systems are conscious, and which are not. Unconscious systems do not perform active inference in a way that entails that characteristic features of consciousness are instantiated, whereas conscious systems do. Specifying the constraints on generative models that underpin active inference of the sort that entails characteristic features of consciousness can lead to a unitary concept of consciousness ([124], as opposed to a bundle of feature descriptions; see [125]). In other words, a sufficiently specified sort of active inference may describe computational processes that account for clusters of features that are characteristic for consciousness—and thereby show *why* these features cluster together (cf. the natural kind approach sketched in [126], p. 7).

15. Information Geometry and Altered States of Consciousness

To recap, the information geometry above—and attending free energy principle—rest upon a separation of external from internal states by blanket states. This move is crucial for elaborating a physics of sentience, in which physical dynamics entail probabilistic beliefs about something. In this sense, it takes us beyond existing formalisms in the physical and philosophical sciences—revealing some key issues. For example, quantum treatments generally rely upon some specification of a Schrödinger potential. But where did this potential come from? Similarly, for statistical thermodynamics, where did the 'heat bath' (i.e., thermal reservoir) come from and what contains the heat bath? In short, there would be no quantum or statistical mechanics in the absence of Markov blankets (i.e., Schrödinger potentials and heat baths). The same questions can be posed to things like integrated information theory: what is this information about, in the absence of a Markovian (belief-based) information geometry? What principles explain the emergence and maintenance of partitions induced by complexes?

The point here is that a Markovian monism (or information geometry) necessarily requires some notion of duality or conjugacy, here afforded by the Markov blanket. On this reading of self-evidencing to nonequilibrium steady state, some pressing questions arise. For example, what would happen if internal and external states were statistically sequestered. In other words, is there a sentient physics for isolated systems, such as those considered in statistical mechanics. From a neurobiological perspective, this speaks to altered states of consciousness that ensue with physiological or pharmacological quenching of blanket states. There are many examples that we could pursue here; including states of consciousness associated with psychedelic and psychomimetic drugs, or, indeed, the false inference associated with psychopathology (e.g., hallucinations and delusions). However, we will focus on a canonical example; namely, sleep and dreaming. So, what does sleep physiology tell us about conscious or unconscious inference?

If, for simplicity, we assume that the state of sleep corresponds to a sequestering of internal states from blanket states, we have an interesting preparation of a neuronal system that is temporarily—and repeatedly—isolated from the sensorium. This simplification is easily substantiated by many neurophysiological and neurochemical aspects of sleep physiology [127]. For us, the key question is: what happens to the Markovian information geometry and Bayesian mechanics of the internal (neuronal) states? At first glance, the notion of self-evidencing as an explanation for internal dynamics simply goes away. This is because the Lyapunov or potential function driving dynamics ceases to exist in the absence of blanket states (technically, the gradients that underwrite gradient flows disappear). However, at nonequilibrium steady-state, periods of disconnection from blanket states must themselves be transient and repetitive; i.e., be part of the itinerant dynamics that have a pullback attractor. This means Bayesian mechanics must still apply, even during the suspension of any coupling with blanket states. We will consider a physiological case (of sleep) in which in autonomous dynamics are still in play. In this setting, the variational free energy gradients are driven by the part of free energy that does not depend upon blanket states. This part is the complexity term of Equation (5), where removing blanket states discloses a description of complexity (i.e., redundancy) resolving internal dynamics:

$$\begin{aligned} f_\mu(\mu) &\approx (Q_{\mu\mu} - \Gamma_{\mu\mu})\nabla_\mu F(\mu) \\ F(\mu) &= \underbrace{D[q_\mu(\eta) \| p(\eta)]}_{complexity} \end{aligned} \tag{13}$$

In other words, neuronal dynamics during sleep will appear to minimise the complexity of the generative model (i.e., minimise the divergence between the posterior beliefs and prior beliefs—in the absence of sensory evidence). This is precisely the argument put forward in statistics for optimising models in the absence of new statistical data—by removing redundant model parameters [128]. In neurophysiology, this is the argument that we have made previously to explain the very existence of sleep phenomenology—and in particular, the role of dreaming [129–131]. In short, physiological states of altered consciousness, such as sleep, may offer an important empirical handle on theoretical notions—notions that arise from the variational principles of sentience.

In summary, an extrinsic information geometry can exist in the (temporary) absence of blanket states, in virtue of prior beliefs held by internal states. These prior beliefs underwrite proto-consciousness [127] and are necessary to generate virtual or fictive realities [132] in states such as dreaming [129–131]. There are many fascinating issues here; for example, the complexity term of the free energy functional above provides a compelling metaphor for the housekeeping that we may enjoy during sleep [71,133]. This complexity minimisation itself has formal links with both machine learning [128] and universal computation [88,90,134]—and physiology in the form of synaptic homoeostasis [133,135].

16. Conclusions

In conclusion, we have rehearsed some of the cornerstones of statistical physics and information theory to show how the very existence of things (i.e., Markov blankets) necessarily induces an information geometry with two aspects. First, the dynamics of physical (internal) states of any sentient particle or creature is equipped with an information geometry, in terms of time dependent changes in probability distributions over internal states. We have called this an *intrinsic* information geometry. At the same time, there is a conjugate information geometry, which pertains to probability densities over external states parameterised by internal states. We have called this an *extrinsic* information geometry (because it is predicated upon probabilistic beliefs about external states). Crucially, the two are formally and fundamentally linked—in that the dynamics of internal states can always be expressed as a gradient flow on a variational free energy functional of belief (protophenomenal) states. This construction is entirely consistent with forces cast in terms of stochastic thermodynamics, with the appropriate constant of proportionality (i.e., Boltzmann's constant and the temperature).

Second, we have considered the time it takes for a particle or creature to return to its attracting manifold (i.e., nonequilibrium steady state) from an initial state. When treated in the form of a path integral or fluctuation theorem, this temporal aspect may distinguish among different kinds of creatures; depending on how deeply their generative model (entailed by internal states) considers the future; c.f., counterfactual depth [111,136]. This is functionally equivalent to the temporal depth or extent of policies; namely, courses of action, and internally consistent with the notion of planning as inference [85,87,137].

Another technical formulation of information-processing—that is closely related to information geometry—is the use of gauge theories (e.g., the celebrated theory of general relativity). Our own work in this area [138] focused on gauge theories associated with information geometry and the Fisher information metric. Recall that the Fisher information metric that equips the belief space or statistical manifold (here, afforded by internal states) with a geometry has a number of revealing interpretations. First, the Fisher information metric is simply the curvature of the variational free energy as one moves on the internal (statistical) manifold. This is the same as the conditional precision or confidence placed in beliefs about external states. From a psychological perspective, this curvature or precision plays a key role in predictive processing (i.e., Bayesian brain) accounts of attentional selection and, a particularly important role in interoceptive inference [67,68,76,139,140]. We emphasise this seamless connection from gauge theories—through information geometry and variational inference—to precision for a special reason. The central role of precision and confidence in mediating consciousness is exactly the endpoint of the phenomenological and neuropsychological analysis of conscious processing and selfhood offered by Mark Solms [107]. Furthermore, the 'paper trail' from gauge theory to attention endorses pre-theoretical notions about their intimate relationship [141,142]. One could develop this story even further, in terms of the predictive processing of precision *per se*—and how this may underwrite mental action and a sense of agency [111,139].

In terms of the philosophy of science, perhaps the most tenable way of treating a dual aspect information geometry is under structural realism. We mean this in the sense that the mathematical and geometric form (i.e., structure)—afforded by the mathematical analysis above—allows one to say something about the relationship between (probabilistic) beliefs and the (statistical) physics of internal states that 'hold' or 'represent' those beliefs. Structural realism takes the pressure off any strong ontological commitments to the mapping between information structures and their content. However, this information structure implies a lawful dependency of probabilistic beliefs (about external states) and parameterised probability distributions (over internal states), in the following sense. Any movement on the internal statistical manifold will necessarily be accompanied by a movement in belief space, as measured by the information length or distance between the beliefs that are parameterised by expected internal states. Furthermore, because these internal states lie upon a statistical manifold of conditional expectations, they must play the role of thermodynamic variables. It follows that belief updating and statistical thermodynamics both supervene on the same internal manifold. Note, the claim here is that physics (i.e., statistical thermodynamics) supervenes on the same statistical manifold as belief updating. This supervenience—on the same statistical manifold—from which both information geometries inherit their structure could be read as the philosophical formulation of the mathematical conjugacy implied by intrinsic and extrinsic information geometries.

In terms of the ontological commitments beyond this structural (realism) argument, any claims would have to be argued much more carefully. It is tenable to associate physics (in the sense of quantum, statistical and classical) mechanics with the intrinsic information geometry. Indeed, this is common parlance in statistical physics [62,64,143]. The more delicate issues arise in terms of commitments to—or interpretation of—the second (extrinsic) sort of information geometry that underwrites Bayesian mechanics. One can avoid any strong ontological commitments here and simply note that should there be any philosophical sentience (i.e., 'qualia') in play, they are more likely to be an attribute of belief updating—and therefore part of Bayesian mechanics. We have approached this issue by suggesting Markovian monism entails a gradual difference between non-conscious and conscious entities, and—in this sense—consciousness is a vague concept.

Author Contributions: Conceptualization, K.J.F, W.W. and J.A.H.; formal analysis, K.J.F and W.W.; writing—original draft preparation, K.J.F. and W.W.; writing—review and editing, K.J.F, W.W. and J.A.H.; supervision, J.A.H.. All authors have read and agreed to the published version of the manuscript.

Funding: K.J.F was funded by a Wellcome Trust Principal Research Fellowship (Ref: 088130/Z/09/Z)

Conflicts of Interest: The authors declare no conflict of interest.

Glossary of Terms and Expressions

Expression	Description	Units
Variables		
$x[\tau] = \{x(t) : t \in (0, \tau)\}$	Trajectory or path through state space	a.u. (m)
$\omega(\tau)$	Random fluctuations	a.u. (m)
$x = \{\eta, s, a, \mu\} \in X$	Markovian partition into external, sensory, active, and internal states	a.u. (m)
$x = \frac{dx}{dt}$	Time derivative (Newton notation)	m/s
$\alpha = \{a, \mu\} \in A$	Autonomous states	a.u. (m)
$b = \{s, a\} \in B$	Blanket states	a.u. (m)
$\pi = \{b, \mu\} \in P$	Particular states	a.u. (m)
$\eta \in E$	External states	a.u. (m)
$\Gamma = \mu_m k_B T$	Amplitude (i.e., half the variance) of random fluctuations	J·s/kg
Q	Rate of solenoidal flow	J·s/kg
$\mu_m = \frac{1}{k_B T}\Gamma$	Mobility coefficient	s/kg
T	Temperature	K (Kelvin)
$\ell = \int d\ell : d\ell^2 = g_{ij} d\lambda^j d\lambda^i$	Information length	nats
$\boldsymbol{\tau} : d\ell(\tau \geq \boldsymbol{\tau}) \approx 0$	Critical time	s
$g_{ij} = E\left[\frac{\partial \Im}{\partial \lambda^i} \frac{\partial \Im}{\partial \lambda^j}\right]$	Fisher (information metric) tensor	a.u.

Functions, functionals and potentials		
$f(x)$	The expected flow of states from any point in state space. This is the expected temporal derivative of x, averaging over random fluctuations in the motion of states.	
$E[x] = E_p[x] = \int x p_\lambda(x)dx$	Expectation or average	
$p_\lambda(x) : \Pr[X \in A] = \int_A p_\lambda(x)dx$	Probability density function parameterised by sufficient statistics λ	
$q_\mu(\eta)$	Variational density—an (approximate posterior) density over external states that is parameterised by internal states	
$\mathcal{A}(x[\tau]) \equiv \Im(x[\tau])$	Action: the surprisal of a path, i.e., the path integral of the Lagrangian	
$U(\pi) = k_B T \cdot \Im(\pi) + \ln Z$	Thermodynamic potential	J or kg m^2/s^2
$F(\pi) \geq \Im(\pi)$	Variational free energy free energy—an upper bound on the surprisal of particular states	nats
$G(\alpha[\boldsymbol{\tau}]) \geq \mathcal{A}(\alpha[\boldsymbol{\tau}]\|\pi_0)$	Expected free energy free energy—an upper bound on the (classical) action of an autonomous path	nats
Operators		
$\nabla_x \Im(x) = \frac{\partial \Im}{\partial x} = \left(\frac{\partial \Im}{\partial x_1}, \frac{\partial \Im}{\partial x_2}, \ldots\right)$	Differential or gradient operator (on a scalar field)	
$\nabla_{xx} \Im(x) = \frac{\partial^2 \Im}{\partial x^2}$	Curvature operator (on a scalar field)	
Entropies and potentials		
$\Im(x) = -\ln p(x)$	Surprisal or self-information	nats
$D[q(x)\|\|p(x)] = E_q[\ln q(x) - \ln p(x)]$	Relative entropy or Kullback–Leibler divergence	nats

(arbitrary units (a.u.), e.g., metres (m), radians (rad), etc.).

References

1. Friston, K. A free energy principle for a particular physics. *arXiv* **2019**, arXiv:1906.10184.
2. Pearl, J. *Probabilistic Reasoning In Intelligent Systems: Networks of Plausible Inference*; Morgan Kaufmann: San Fransisco, CA, USA, 1988.
3. Parr, T.; Da Costa, L.; Friston, K. Markov blankets, information geometry and stochastic thermodynamics. *Philos. Trans. A Math. Phys. Eng. Sci.* **2020**, *378*, 20190159. [CrossRef] [PubMed]
4. Kirchhoff, M.; Parr, T.; Palacios, E.; Friston, K.; Kiverstein, J. The Markov blankets of life: Autonomy, active inference and the free energy principle. *J. R. Soc. Interface* **2018**, *15*. [CrossRef] [PubMed]
5. Palacios, E.R.; Razi, A.; Parr, T.; Kirchhoff, M.; Friston, K. Biological Self-organisation and Markov blankets. *bioRxiv* **2017**, arXiv:10.1101/227181.
6. Clark, A. How to Knit Your Own Markov Blanket. In *Philosophy and Predictive Processing*; Metzinger, T.K., Wiese, W., Eds.; MIND Group: Frankfurt, Germany, 2017.
7. Friston, K.J.; Kahan, J.; Razi, A.; Stephan, K.E.; Sporns, O. On nodes and modes in resting state fMRI. *Neuroimage* **2014**, *99*, 533–547. [CrossRef]
8. Pellet, J.P.; Elisseeff, A. Using Markov blankets for causal structure learning. *J. Mach. Learn. Res.* **2008**, *9*, 1295–1342.
9. Friston, K. Life as we know it. *J. R. Soc. Interface* **2013**, *10*, 20130475. [CrossRef]
10. Sekimoto, K. Langevin Equation and Thermodynamics. *Prog. Theor. Phys. Suppl.* **1998**, *130*, 17–27. [CrossRef]
11. Ao, P. Emerging of Stochastic Dynamical Equalities and Steady State Thermodynamics. *Commun. Theor. Phys.* **2008**, *49*, 1073–1090. [CrossRef]
12. Seifert, U. Stochastic thermodynamics, fluctuation theorems and molecular machines. *Rep. Prog. Phys. Phys. Soc.* **2012**, *75*, 126001. [CrossRef]
13. Crauel, H.; Flandoli, F. Attractors for Random Dynamical-Systems. *Probab. Theory Rel.* **1994**, *100*, 365–393. [CrossRef]

14. Birkhoff, G.D. *Dynamical Systems*; American Mathematical Society: New York, NY, USA, 1927.
15. Tribus, M. *Thermodynamics and Thermostatics: An Introduction to Energy, Information and States of Matter, with Engineering Applications*; D. Van Nostrand Company Inc.: New York, NY, USA, 1961.
16. Jaynes, E.T. Information Theory and Statistical Mechanics. *Phys. Rev. Ser. II* **1957**, *106*, 620–630. [CrossRef]
17. Jones, D.S. *Elementary Information Theory*; Clarendon Press: Oxford, UK, 1979.
18. MacKay, D.J.C. *Information Theory, Inference and Learning Algorithms*; Cambridge University Press: Cambridge, UK, 2003.
19. Kerr, W.C.; Graham, A.J. Generalized phase space version of Langevin equations and associated Fokker-Planck equations. *Eur. Phys. J. B* **2000**, *15*, 305–311. [CrossRef]
20. Frank, T.D.; Beek, P.J.; Friedrich, R. Fokker-Planck perspective on stochastic delay systems: Exact solutions and data analysis of biological systems. *Phys. Rev. E Stat. Nonlin Soft Matter Phys.* **2003**, *68*, 021912. [CrossRef]
21. Frank, T.D. *Nonlinear Fokker-Planck Equations: Fundamentals and Applications. Springer Series in Synergetics*; Springer: Berlin, Germany, 2004.
22. Tomé, T. Entropy Production in Nonequilibrium Systems Described by a Fokker-Planck Equation. *Braz. J. Phys.* **2006**, *36*, 1285–1289. [CrossRef]
23. Kim, E.-j. Investigating Information Geometry in Classical and Quantum Systems through Information Length. *Entropy* **2018**, *20*, 574. [CrossRef]
24. Yuan, R.; Ma, Y.; Yuan, B.; Ping, A. Bridging Engineering and Physics: Lyapunov Function as Potential Function. *arXiv* **2010**, arXiv:1012.2721v1.
25. Friston, K.; Ao, P. Free energy, value, and attractors. *Comput. Math. Methods Med.* **2012**, *2012*, 937860. [CrossRef]
26. Sutton, R.S.; Precup, D.; Singh, S. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artif. Intell.* **1999**, *112*, 181–211. [CrossRef]
27. Kauder, E. Genesis of the Marginal Utility Theory: From Aristotle to the End of the Eighteenth Century. *Econ. J.* **1953**, *63*, 638–650. [CrossRef]
28. Fleming, W.H.; Sheu, S.J. Risk-sensitive control and an optimal investment model II. *Ann. Appl. Probab.* **2002**, *12*, 730–767. [CrossRef]
29. Haken, H. *Synergetics: An Introduction. Non-Equilibrium Phase Transition and Self-Selforganisation in Physics, Chemistry and Biology*; Springer Verlag: Berlin, Germany, 1983.
30. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*; MIT Press: Cambridge, MA, USA, 1998.
31. Todorov, E.; Jordan, M.I. Optimal feedback control as a theory of motor coordination. *Nat. Neurosci.* **2002**, *5*, 1226–1235. [CrossRef] [PubMed]
32. Kappen, H.J. Path integrals and symmetry breaking for optimal control theory. *J. Stat. Mech. Theory Exp.* **2005**, *11*, 11011. [CrossRef]
33. Barlow, H. Possible principles underlying the transformations of sensory messages. In *Sensory Communication*; Rosenblith, W., Ed.; MIT Press: Cambridge, MA, USA, 1961; pp. 217–234.
34. Tschacher, W.; Haken, H. Intentionality in non-equilibrium systems? The functional aspects of self-organised pattern formation. *New Ideas Psychol.* **2007**, *25*, 1–15. [CrossRef]
35. Bernard, C. *Lectures on the Phenomena Common to Animals and Plants*; Charles C Thomas: Springfield, IL, USA, 1974.
36. Kass, R.E.; Raftery, A.E. Bayes Factors. *J. Am. Stat. Assoc.* **1995**, *90*, 773–795. [CrossRef]
37. Hohwy, J. The Self-Evidencing Brain. *Noûs* **2016**, *50*, 259–285. [CrossRef]
38. Friston, K.; Da Costa, L.; Parr, T. Some interesting observations on the free energy principle. *arXiv* **2020**, arXiv:2002.04501.
39. Bossaerts, P.; Murawski, C. From behavioural economics to neuroeconomics to decision neuroscience: The ascent of biology in research on human decision making. *Curr. Opin. Behav. Sci.* **2015**, *5*, 37–42. [CrossRef]
40. Linsker, R. Perceptual neural organization: Some approaches based on network models and information theory. *Annu. Rev. Neurosci.* **1990**, *13*, 257–281. [CrossRef]
41. Optican, L.; Richmond, B.J. Temporal encoding of two-dimensional patterns by single units in primate inferior cortex. II Information theoretic analysis. *J. Neurophysiol.* **1987**, *57*, 132–146. [CrossRef]

42. Friston, K.; Kilner, J.; Harrison, L. A free energy principle for the brain. *J. Physiol.* **2006**, *100*, 70–87. [CrossRef] [PubMed]
43. Nicolis, G.; Prigogine, I. *Self-Organization in Non-Equilibrium Systems*; John Wiley: New York, NY, USA, 1977.
44. Kauffman, S. *The Origins of Order: Self-Organization and Selection in Evolution*; Oxford University Press: Oxford, UK, 1993.
45. Conant, R.C.; Ashby, W.R. Every Good Regulator of a system must be a model of that system. *Int. J. Syst. Sci.* **1970**, *1*, 89–97. [CrossRef]
46. Ashby, W.R. Principles of the self-organizing dynamic system. *J. Gen. Psychol.* **1947**, *37*, 125–128. [CrossRef] [PubMed]
47. MacKay, D.J. Free-energy minimisation algorithm for decoding and cryptoanalysis. *Electron. Lett.* **1995**, *31*, 445–447. [CrossRef]
48. Helmholtz, H. Concerning the perceptions in general. In *Treatise on Physiological Optics*; Dover: New York, NY, USA, 1962.
49. Gregory, R.L. Perceptions as hypotheses. *Philos. Trans. R. Soc. Lond. B* **1980**, *290*, 181–197.
50. Dayan, P.; Hinton, G.E.; Neal, R.M.; Zemel, R.S. The Helmholtz machine. *Neural Comput.* **1995**, *7*, 889–904. [CrossRef]
51. Beal, M.J. Variational Algorithms for Approximate Bayesian Inference. Ph.D. Thesis, University College London, London, UK, 2003.
52. Dauwels, J. On Variational Message Passing on Factor Graphs. In Proceedings of the 2007 IEEE International Symposium on Information Theory, Nice, France, 24–29 June 2007; pp. 2546–2550.
53. Suh, S.; Chae, D.H.; Kang, H.G.; Choi, S. Echo-State Conditional Variational Autoencoder for Anomaly Detection. In Proceedings of the 2016 International Joint Conference on Neural Networks, Vancouver, BC, Canada, 24–29 July 2016; pp. 1015–1022.
54. Roweis, S.; Ghahramani, Z. A unifying review of linear gaussian models. *Neural Comput.* **1999**, *11*, 305–345. [CrossRef]
55. Hinton, G.E.; Zemel, R.S. Autoencoders, minimum description length and Helmholtz free energy. In Proceedings of the 6th International Conference on Neural Information Processing Systems, Denver, CO, USA, November 1994; pp. 3–10.
56. Ikeda, S.; Tanaka, T.; Amari, S.-I. Stochastic reasoning, free energy, and information geometry. *Neural Comput.* **2004**, *16*, 1779–1810. [CrossRef]
57. Knill, D.C.; Pouget, A. The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends Neurosci.* **2004**, *27*, 712–719. [CrossRef]
58. Yedidia, J.S.; Freeman, T.; Weiss, Y. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Trans. Inf. Theory* **2005**, *51*, 2282–2312. [CrossRef]
59. Amari, S. Natural gradient works efficiently in learning. *Neural Comput.* **1998**, *10*, 251–276. [CrossRef]
60. Ay, N. Information Geometry on Complexity and Stochastic Interaction. *Entropy* **2015**, *17*, 2432. [CrossRef]
61. Caticha, A. The basics of information geometry. *AIP Conf. Proc.* **2015**, *1641*, 15–26. [CrossRef]
62. Crooks, G.E. Measuring thermodynamic length. *Phys. Rev. Lett.* **2007**, *99*, 100602. [CrossRef] [PubMed]
63. Herbart, J.F. *Lehrbuch zur Psychologie*, 2nd ed.; Unzer: Königsberg, Prussia, 1834.
64. Holmes, Z.; Weidt, S.; Jennings, D.; Anders, J.; Mintert, F. Coherent fluctuation relations: From the abstract to the concrete. *Quantum* **2019**, *3*, 124. [CrossRef]
65. Van Gulick, R. Reduction, Emergence and Other Recent Options on the Mind/Body Problem. A Philosophic Overview. *J. Conscious. Stud.* **2001**, *8*, 1–34.
66. Hobson, J.A.; Friston, K.J. A Response to Our Theatre Critics. *J. Conscious. Stud.* **2016**, *23*, 245–254.
67. Brown, H.; Adams, R.A.; Parees, I.; Edwards, M.; Friston, K. Active inference, sensory attenuation and illusions. *Cogn. Process.* **2013**, *14*, 411–427. [CrossRef]
68. Clark, A. The many faces of precision. *Front. Psychol.* **2013**, *4*, 270.
69. Hohwy, J. *The Predictive Mind*; Oxford University Press: Oxford, UK, 2013.
70. Seth, A. The cybernetic brain: From interoceptive inference to sensorimotor contingencies. In *Open MIND*; Metzinger, T., Windt, J.M., Eds.; MIND Group: Frankfurt a.M., Germany, 2014.
71. Hobson, J.A.; Friston, K.J. Waking and dreaming consciousness: Neurobiological and functional considerations. *Prog. Neurobiol.* **2012**, *98*, 82–98. [CrossRef]

72. Friston, K.; Buzsaki, G. The Functional Anatomy of Time: What and When in the Brain. *Trends Cogn. Sci.* **2016**, *20*, 500–511. [CrossRef] [PubMed]
73. Adams, R.A.; Shipp, S.; Friston, K.J. Predictions not commands: Active inference in the motor system. *Brain Struct. Funct.* **2013**, *218*, 611–643. [CrossRef] [PubMed]
74. Ungerleider, L.G.; Haxby, J.V. 'What' and 'where' in the human brain. *Curr. Opin. Neurobiol.* **1994**, *4*, 157–165. [CrossRef]
75. Shigeno, S.; Andrews, P.L.R.; Ponte, G.; Fiorito, G. Cephalopod Brains: An Overview of Current Knowledge to Facilitate Comparison With Vertebrates. *Front. Physiol.* **2018**, *9*, 952. [CrossRef]
76. Seth, A.K. Interoceptive inference, emotion, and the embodied self. *Trends Cogn. Sci.* **2013**, *17*, 565–573. [CrossRef]
77. Arnold, L. *Random Dynamical Systems (Springer Monographs in Mathematics)*; Springer: Berlin/Heidelberg, Germany, 2003.
78. Kleeman, R. A Path Integral Formalism for Non-equilibrium Hamiltonian Statistical Systems. *J. Stat. Phys.* **2014**, *158*, 1271–1297. [CrossRef]
79. Jarzynski, C. Nonequilibrium Equality for Free Energy Differences. *Phys. Rev. Lett.* **1997**, *78*, 2690–2693. [CrossRef]
80. Pezzulo, G.; Rigoli, F.; Friston, K. Active Inference, homeostatic regulation and adaptive behavioural control. *Prog. Neurobiol.* **2015**, *134*, 17–35. [CrossRef]
81. Sterling, P.; Eyer, J. Allostasis: A new paradigm to explain arousal pathology. In *Handbook of Life Stress, Cognition and Health*; John Wiley & Sons: Hoboken, NJ, USA, 1988; pp. 629–649.
82. Balleine, B.W.; Dickinson, A. Goal-directed instrumental action: Contingency and incentive learning and their cortical substrates. *Neuropharmacology* **1998**, *37*, 407–419. [CrossRef]
83. Ramsay, D.S.; Woods, S.C. Clarifying the Roles of Homeostasis and Allostasis in Physiological Regulation. *Psychol. Rev.* **2014**, *121*, 225–247. [CrossRef]
84. Stephan, K.E.; Manjaly, Z.M.; Mathys, C.D.; Weber, L.A.E.; Paliwal, S.; Gard, T.; Tittgemeyer, M.; Fleming, S.M.; Haker, H.; Seth, A.K.; et al. Allostatic Self-efficacy: A Metacognitive Theory of Dyshomeostasis-Induced Fatigue and Depression. *Front. Hum. Neurosci.* **2016**, *10*, 550. [CrossRef] [PubMed]
85. Attias, H. Planning by Probabilistic Inference. In Proceedings of the 9th International Workshop on Artificial Intelligence and Statistics, Key West, FL, USA, 3–6 January 2003.
86. Toussaint, M.; Storkey, A. Probabilistic inference for solving discrete and continuous state Markov Decision Processes. In Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, 25–29 June 2006; pp. 945–952.
87. Botvinick, M.; Toussaint, M. Planning as inference. *Trends Cogn Sci.* **2012**, *16*, 485–488. [CrossRef] [PubMed]
88. Schmidhuber, J. Curious model-building control systems. *Proc. Int. Jt. Conf. Neural Netw. Singap.* **1991**, *2*, 1458–1463.
89. Schmidhuber, J. Formal Theory of Creativity, Fun, and Intrinsic Motivation (1990–2010). *IEEE Trans. Auton. Ment. Dev.* **2010**, *2*, 230–247. [CrossRef]
90. Sun, Y.; Gomez, F.; Schmidhuber, J. Planning to Be Surprised: Optimal Bayesian Exploration in Dynamic Environments. In Proceedings of the Artificial General Intelligence: 4th International Conference, AGI 2011, Mountain View, CA, USA, 3–6 August 2011; Schmidhuber, J., Thórisson, K.R., Looks, M., Eds.; Springer: Berlin/Heidelberg, Germany, 2011; pp. 41–51.
91. Gibson, J.J. The theory of affordances. In *Perceiving, Acting, and Knowing: Toward an Ecological Psychology*; Erlbaum: Hillsdale, NJ, USA, 1977; pp. 67–82.
92. Bruineberg, J.; Rietveld, E. Self-organization, free energy minimization, and optimal grip on a field of affordances. *Front. Hum. Neurosci.* **2014**, *8*, 599. [CrossRef]
93. Parr, T.; Friston, K.J. Working memory, attention, and salience in active inference. *Sci. Rep.* **2017**, *7*, 14678. [CrossRef]
94. Friston, K.; Mattout, J.; Kilner, J. Action understanding and active inference. *Biol. Cybern.* **2011**, *104*, 137–160. [CrossRef]
95. Rizzolatti, G.; Craighero, L. The mirror-neuron system. *Annu. Rev. Neurosci.* **2004**, *27*, 169–192. [CrossRef]
96. Kilner, J.M.; Friston, K.J.; Frith, C.D. Predictive coding: An account of the mirror neuron system. *Cogn. Process.* **2007**, *8*, 159–166. [CrossRef]

97. Gallese, V.; Goldman, A. Mirror neurons and the simulation theory of mind-reading. *Trends Cogn. Sci.* **1998**, *2*, 493–501. [CrossRef]
98. Friston, K.; Adams, R.A.; Perrinet, L.; Breakspear, M. Perceptions as hypotheses: Saccades as experiments. *Front. Psychol.* **2012**, *3*, 151. [CrossRef] [PubMed]
99. Dennett, D.C. True believers: The intentional strategy and why it works. In *Scientific Explanation: Papers Based on Herbert Spencer Lectures Given in the University of Oxford*; Heath, A.F., Ed.; Clarendon Press: Oxford, UK, 1981; pp. 150–167.
100. Chalmers, D.J. Panpsychism and panprotopsychism. In *Consciousness in the Physical World: Perspectives on Russellian Monism*; Oxford University Press: New York, NY, USA, 2015; pp. 246–276.
101. Stubenberg, L. Neutral monism. In *The Stanford Encyclopedia of Philosophy*; Zalta, E.N., Ed.; Metaphysics Research Lab, Stanford University: Stanford, CA, USA, 2018.
102. Howell, R. The Russellian Monist's Problems with Mental Causation. *Philos. Q.* **2015**, *65*, 22–39. [CrossRef]
103. Kirchhoff, M.D.; Froese, T. Where There is Life There is Mind: In Support of a Strong Life-Mind Continuity Thesis. *Entropy* **2017**, *19*, 169. [CrossRef]
104. Skrbina, D. Minds, objects, and relations. Toward a dual-aspect ontology. In *Mind that Abides. Panpsychism in the New Millenium*; Skrbina, D., Ed.; John Benjamins Publishing Company: Amsterdam, The Netherlands; Philadelphia, PA, USA, 2009; pp. 361–397.
105. Velmans, M. Reflexive Monism. *J. Conscious. Stud.* **2008**, *15*, 5–50.
106. Benovsky, J. Dual-Aspect Monism. *Philos. Investig.* **2016**, *39*, 335–352. [CrossRef]
107. Solms, M.; Friston, K. How and Why Consciousness Arises. Some Considerations from Physics and Physiology. *J. Conscious. Stud.* **2018**, *25*, 202–238.
108. Woodward, J. Mental Causation and Neural Mechanisms. In *Being Reduced: New Essays on Reduction, Explanation, and Causation*; Hohwy, J., Kallestrup, J., Eds.; Oxford University Press: Oxford, UK, 2008.
109. Papineau, D. Causation is Macroscopic but Not Irreducible. In *Mental Causation and Ontology*; Gibb, S.C., Lowe, E.J., Ingthorsson, R.D., Eds.; Oxford University Press: Oxford, UK, 2013; pp. 126–151.
110. Mansell, W. Control of perception should be operationalized as a fundamental property of the nervous system. *Top. Cogn. Sci.* **2011**, *3*, 257–261. [CrossRef]
111. Seth, A.K. Inference to the Best Prediction. In *Open MIND*; Metzinger, T.K., Windt, J.M., Eds.; MIND Group: Frankfurt, Germany, 2015.
112. Chalmers, D.J. The Meta-Problem of Consciousness. *J. Conscious. Stud.* **2018**, *25*, 6–61.
113. Clark, A.; Friston, K.; Wilkinson, S. Bayesing Qualia. Consciousness as Inference, Not Raw Datum. *J. Conscious. Stud.* **2019**, *26*, 19–33.
114. Schweizer, P. Triviality Arguments Reconsidered. *Minds Mach.* **2019**, *29*, 287–308. [CrossRef]
115. Fodor, J. The mind-body problem. *Sci. Am.* **1981**, *244*, 114–123. [CrossRef] [PubMed]
116. Piccinini, G. Computation without Representation. *Philos. Stud.* **2006**, *137*, 205–241. [CrossRef]
117. Gładziejewski, P. Predictive coding and representationalism. *Synthese* **2016**, *193*, 559–582. [CrossRef]
118. Tononi, G. An information integration theory of consciousness. *BMC Neurosci.* **2004**, *5*, 42. [CrossRef]
119. Tononi, G. Consciousness as Integrated Information: A Provisional Manifesto. *Biol. Bull.* **2008**, *215*, 216–242. [CrossRef]
120. Balduzzi, D.; Tononi, G. Qualia: The Geometry of Integrated Information. *PLoS Comput. Biol.* **2009**, *5*, 1–24. [CrossRef]
121. van Leeuwen, C. Perceptual-learning systems as conservative structures: Is economy an attractor? *Psychol. Res.* **1990**, *52*, 145–152. [CrossRef]
122. Tervo, D.G.; Tenenbaum, J.B.; Gershman, S.J. Toward the neural implementation of structure learning. *Curr. Opin. Neurobiol.* **2016**, *37*, 99–105. [CrossRef]
123. Tononi, G.; Boly, M.; Massimini, M.; Koch, C. Integrated information theory: From consciousness to its physical substrate. *Nat. Rev. Neurosci.* **2016**, *17*, 450–461. [CrossRef]
124. Wiese, W. Toward a Mature Science of Consciousness. *Front. Psychol.* **2018**, *9*, 693. [CrossRef]
125. Wiese, W. *Experienced Wholeness. Integrating Insights from Gestalt Theory, Cognitive Neuroscience, and Predictive Processing*; MIT Press: Cambridge, MA, USA, 2018.
126. Bayne, T. On the axiomatic foundations of the integrated information theory of consciousness. *Neurosci. Conscious.* **2018**, *2018*, niy007. [CrossRef] [PubMed]

127. Hobson, J.A. REM sleep and dreaming: Towards a theory of protoconsciousness. *Nat. Rev. Neurosci.* **2009**, *10*, 803–813. [CrossRef] [PubMed]
128. Hinton, G.E.; Dayan, P.; Frey, B.J.; Neal, R.M. The "wake-sleep" algorithm for unsupervised neural networks. *Science* **1995**, *268*, 1158–1161. [CrossRef] [PubMed]
129. Hobson, J.A. *Dreaming as Delirium*; The MIT Press: Cambridge, MA, USA, 1999.
130. Hobson, J.A.; Friston, K.J. Consciousness, Dreams, and Inference The Cartesian Theatre Revisited. *J. Conscious. Stud.* **2014**, *21*, 6–32.
131. Friston, K.J.; Lin, M.; Frith, C.D.; Pezzulo, G.; Hobson, J.A.; Ondobaka, S. Active Inference, Curiosity and Insight. *Neural Comput.* **2017**, *29*, 2633–2683. [CrossRef]
132. Metzinger, T. *Being No One: The Self-Model Theory of Subjectivity*; MIT Press: Cambridge, MA, USA, 2003.
133. Tononi, G.; Cirelli, C. Sleep function and synaptic homeostasis. *Sleep Med. Rev.* **2006**, *10*, 49–62. [CrossRef]
134. Hochreiter, S.; Schmidhuber, J. Flat minima. *Neural Comput.* **1997**, *9*, 1–42. [CrossRef]
135. Gilestro, G.F.; Tononi, G.; Cirelli, C. Widespread changes in synaptic markers as a function of sleep and wakefulness in Drosophila. *Science* **2009**, *324*, 109–112. [CrossRef]
136. Palmer, C.J.; Seth, A.K.; Hohwy, J. The felt presence of other minds: Predictive processing, counterfactual predictions, and mentalising in autism. *Conscious. Cogn.* **2015**, *36*, 376–389. [CrossRef]
137. Kaplan, R.; Friston, K.J. Planning and navigation as active inference. *Biol. Cybern.* **2018**, *112*, 323–343. [CrossRef]
138. Sengupta, B.; Tozzi, A.; Cooray, G.K.; Douglas, P.K.; Friston, K.J. Towards a Neuronal Gauge Theory. *PLoS Biol.* **2016**, *14*, e1002400. [CrossRef] [PubMed]
139. Limanowski, J.; Friston, K. 'Seeing the Dark': Grounding Phenomenal Transparency and Opacity in Precision Estimation for Active Inference. *Front. Psychol.* **2018**, *9*, 643. [CrossRef] [PubMed]
140. Fotopoulou, A.; Tsakiris, M. Mentalizing homeostasis: The social origins of interoceptive inference—Replies to Commentaries. *Neuropsychoanalysis* **2017**, *19*, 71–76. [CrossRef]
141. Dehaene, S.; Naccache, L. Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition* **2001**, *79*, 1–37. [CrossRef]
142. Cavanna, A.E.; Trimble, M.R. The precuneus: A review of its functional anatomy and behavioural correlates. *Brain* **2006**, *129*, 564–583. [CrossRef]
143. Still, S.; Sivak, D.A.; Bell, A.J.; Crooks, G.E. Thermodynamics of prediction. *Phys. Rev. Lett.* **2012**, *109*, 120604. [CrossRef]