First principles in the life sciences: The free-energy principle, organicism, and mechanism

Matteo Colombo & Cory Wright

Abstract The free-energy principle states that all systems that resist a tendency to physical disintegration must minimize their free energy. Originally proposed to account for perception, learning, and action, the free-energy principle has been applied to the evolution, development, morphology, and function of the brain, and has been called a *postulate*, an *unfalsifiable principle*, a *natural law*, and an *imperative*. While it might afford a theoretical foundation for understanding the relationship between environment, life, and mind, its epistemic status and scope are unclear. Also unclear is how the free-energy principle relates to prominent theoretical approaches to life science phenomena, such as organicism and mechanicism. This paper clarifies both issues, and identifies limits and prospects for the free-energy principle as a first principle in the life sciences.

Keywords: adaptation; free energy; life; mechanism; organicism

1 Introduction

According to the free-energy principle (FEP), all systems that resist a tendency to physical disintegration must minimize their free energy. Originally proposed to explain how sensory cortex infers the causes of its inputs and learns causal regularities, FEP has been used to elucidate the function of action, perception, and attention, and to account for organisms' evolution and development (Friston 2003, 2009, 2010a, 2013; Friston et al. 2006; Friston & Stephan 2007). Advocates have claimed FEP offers a "framework within which to explain the constitutive coupling of the brain to the body and the environment," which provides "a normative, teleological essence to the synthesis of biology and information," and which may illuminate the continuity between life and mind (Allen & Friston 2018: 2476).

Advocates present FEP as a "mandatory principle" or "imperative" for biological systems, and as a principle enjoying a "fundamental status" in neuroscience (Friston et al. 2006: 71; Friston

& Stephan 2007). The principle purportedly "applies to any biological system [...] from single-cell organisms to social networks" (Friston 2009: 293). These bold ascriptions have attracted attention in philosophy and the life sciences. However, FEP's epistemic status remains opaque, along with its exact role in biological and neuroscientific theorizing.

Regarding its role in theorizing, FEP seemingly conflicts with tenets of two of the most prominent contemporary theoretical approaches in the life sciences: organicism (Gilbert & Sarkar 2000; Soto et al. 2016) and mechanism (Brandon 1984; Bechtel & Richardson 1993/2010). Regarding its epistemic status, one worry is that FEP lacks explanatory power because it "is divorced from the biophysical reality of the nervous system" (Fiorillo 2010: 605). The principle's unifying power has been called into question, too. Some have suggested that FEP provides an implausible model of the functional roles of perception and action (Gershman & Daw 2012; Colombo 2017; Klein 2018); others have argued that FEP doesn't deliver a grand unifying theory, and a plurality of modeling approaches is preferable for explaining complex neurophysiological and cognitive phenomena (Marblestone et al. 2016; Colombo & Wright 2017). Also controversial is how FEP illuminates the continuity between life and mind, synthesizing biology and information—especially if FEP is committed to some form of cognitivism (Hohwy 2016; Kirchhoff & Froese 2017). Most basically, the inferential steps in the reasoning leading to FEP need still be laid out clearly and accessibly to allow for adequate evaluation.

According to organicists (aka 'holists'), phenomena studied in the life sciences should be explained by appealing to whole organisms. Organicists often appeal to the principles of biological autonomy and adaptivity (Varela, Maturana, & Uribe 1974; Di Paolo & Thompson 2014; Moreno & Mossio 2015), organization as closure of constraints (Maturana 1975; Mossio et al. 2016), and variation as extended criticality (Montévil et al. 2016). While FEP borrows the formalism of random dynamical systems to explain the behavior and nature of organisms, organicists maintain that the formalisms borrowed from current theories in physics are not apt for the representation of life science phenomena (Longo et al. 2012).

According to mechanists, phenomena studied in the life sciences should be explained by appealing to the component parts and operations of mechanisms, where a mechanism is a spatiotemporally-organized composite system producing a phenomenon (Bechtel & Richardson 1993/2010; Darden 2006). While FEP is used to logically derive claims about structural and functional properties of the brain "a priori, on the basis of purely theoretical considerations" (Friston 2003: 1325), mechanistic philosophers argue that laws and logical deduction do not adequately capture explanation in the life sciences (Glennan 2017).

In this paper, we clarify the epistemic status of FEP by answering two sets of questions. First: what are the inferential steps leading to FEP? And how do they relate to one another? Second: what's the relationship between FEP and central tenets of the organicist and mechanistic philosophies when it comes to adequate scientific representation of the phenomena of life?

To answer these questions, we provide a transparent reconstruction of the reasoning leading to FEP (§2) and then deploy a contrastive strategy to bring out salient assumptions of FEP. We argue that FEP is inconsistent with a tenet of organicism, specifically that the formalism and concepts of current physical theories are not apt for the scientific representation of organisms and their behaviors (§3). We also show that FEP is inconsistent with mechanistic approaches in the life sciences, which eschew laws and theories and take the explanatory power of scientific representations to be dependent on the degree of relevant biophysical detail they include. After we clarify the status of FEP as a first principle (§4), we conclude by suggesting that understanding phenomena in the life sciences should allow for incompatible approaches (§5). In particular, the axiomatic, idealizing, first-principles methods favored by free-energy theorists should be pursued in the life sciences alongside the synthetic methods favored by organicists, and the analytic methods favored by mechanists.

2 The transcendental argument

Free-energy theorists formulate FEP in various ways. Depending on the exact formulation, the scope of application of FEP varies. One formulation has FEP focused on brains, aiming to clarify the functional significance of their activity and structural connectivity (Friston 2003, 2009; Friston et al. 2006). Another has FEP ranging over any biological entity, process, or complex system—including brainless organisms like single cells and plants, evolutionary processes by natural selection, and ecosystems—and aiming to explain how biological systems can maintain their physical integrity in a changing environment (Friston 2013; Hobson & Friston 2016). In yet another formulation, FEP applies to any complex adaptive system, including physical information systems and non-biological systems like social networks and artefacts (Friston 2009: 293).

Under any formulation, the reasoning leading to FEP has the form of a transcendental argument for the conclusion that FEP is a condition on the very possibility of systems maintaining their physical integrity and displaying adaptive behavior. Beginning with the observation that some systems behave adaptively, resisting a tendency to disorder, the transcendental deduction of FEP involves five main steps organized in two batches:

(1) If a system Σ acts selectively on the environment to avoid phase transitions and is in a non-equilibrium steady-state, then Σ behaves adaptively.[1]

(2) Σ behaves adaptively only if Σ preserves its physical integrity by maintaining its "characteristic" variables within homeostatic bounds despite environmental fluctuations.[2]

[1] As explained in §2.1, *equilibrium* refers to thermodynamic equilibrium. A system is in a steady state if the variables defining the behavior of the system are unchanging with time. An equilibrium state is a special case of a steady state. (A system in equilibrium is one in a steady state; the converse isn't necessarily true.)

[2] Values of these characteristic variables would be observable properties of "the extended phenotype of the organism—its morphology, physiology, behavioural patterns, cultural patterns, and designer environments," and would be phase functions of an attracting set of the organism's states (Ramstead et al. 2017: 3). The idea is that any living system possesses an 'attracting set', i.e., a set of states towards which it will tend to evolve, for a wide variety of values of its initial states, and "this set of attracting states can be interpreted as the extended phenotype of the organism" (ibid). Free-energy theorists' transcendental deduction assumes that an attracting set in a system open to its external milieu is formally equivalent to a steady-state solution that is far from equilibrium.

(3) Σ acts selectively on the environment to avoid phase transitions and is in a non-equilibrium steady-state just in case Σ preserves its physical integrity by maintaining its "characteristic" variables within homeostatic bounds despite environmental fluctuations.

(4) Σ preserves its physical integrity by maintaining its "characteristic" variables within homeostatic bounds despite environmental fluctuations just in case Σ minimizes the informational entropy (*average surprise*) of its possible sensory states.

(5) If Σ minimizes the informational entropy of its possible sensory states, Σ minimizes the free energy of its possible sensory states.[3]

∴ (6) for any system Σ to maintain its physical integrity and behave adaptively despite environmental fluctuations, Σ must minimize its free energy.

The first three claims comprise the first batch, where (1) states a sufficient condition on adaptive behavior, (2) states a necessary condition, and (3) functions as a bridge principle that connects physical and biological predicates. The bridge is derivable from the conjunction of its left-to-right conditional, which follows directly from (1) and (2), and the stipulation of its suppressed right-to-left converse. The next two claims comprise the second batch, where (4) stipulatively connects homeostasis to information-theoretic quantities, and (5) states that free-energy minimization is a necessary condition on minimizing informational entropy.[4] Claim (6) expresses FEP, and follows from the two batches of claims.

[3] According to some formulations, any system that conserves its boundaries "can be described as" modeling its external milieu or "can be cast as" minimizing free energy (Hobson & Friston 2016: 246; Ramstead et al. 2017: 2). In other formulations, the system itself is said to be the modeler (Friston 2011). We address this confusion in §3.

[4] Premise 5 should be treated with care. First, in machine learning, free energy is a variational bound on informational entropy that often plays a central role in variational Bayesian inferences. Where X is an observed variable (e.g., a sensory sample) and Z a hidden random variable of interest (e.g., an environmental cause of a sensory sample), the idea behind variational Bayesian inference is to find some approximation distribution density $q(Z)$ that is tractable and that is as close as possible to the true posterior distribution $p(Z|X)$. As the approximation distribution can have its own parameters $q(Z|\theta)$, the problem of inferring the value of Z given the value of X can be understood as an optimization problem, in which one aims to find the parameter values that minimize some objective function and make q as close as possible to the posterior of interest. To measure the closeness of the two distributions $q(Z)$ and $p(Z|X)$, a common metric is the Kullback-Leibler divergence. Second, free-energy theorists argue that organisms cannot minimize the informational entropy of their sensory states (i.e., 'surprisal') directly, because this involves the intractable problem of

The modal force of FEP, as well as its claim to being an a priori first principle, depend on free-energy theorists' understanding of the predicate *being an adaptive system*, on assumptions about the semantic equivalence between concepts from different disciplines, and on their interpretation of probability. We'll consider their understanding of adaptive living systems in §2.2, and put into sharper focus the other two assumptions underlying their reasoning in §3.

2.1 Thermodynamics

The argument for FEP begins with the observation that some systems maintain their physical integrity and display adaptive behavior amid a changing environment. Such systems are thermodynamically open, persisting far from their thermodynamic equilibrium state (Friston et al. 2006: 71–72; Friston & Stephan 2007: 421–422). Systems that are thermodynamically open exchange matter and energy with their surroundings. For example, both snowflakes and bacteria exchange chemicals and energy with their surroundings. Snowflakes acquire and lose matter and heat under the causal pressures of their environment, and bacteria allow matter and energy to cross their cytoplasmic membranes. Closed systems have boundaries that matter from their surroundings cannot cross. Earth, for example, exchanges energy with its surroundings, but hardly any matter. Isolated systems exchange neither energy nor matter with their surroundings. The universe and a closed thermos bottle are examples of isolated systems.

---

computing marginal probabilities. Instead, they argue that organisms can tractably minimize the informational entropy of their sensory states indirectly, by minimizing a bound called 'variational free energy'. As Buckley et al.'s (2017) helpful mathematical review makes clear, this claim requires qualification, since variational free energy is a tight bound on surprisal only when the organism's current 'best guess' of the causes of its sensory input (represented by a 'recognition density function' $q(Z)$) is identical with the posterior density of environmental states given the organism's sensory input $p(Z|X)$. Finding an optimal recognition density that is identical with the posterior is non-trivial, and requires further assumptions about its form, and about the form of the dynamics in the environment. "Furthermore, while this process furnishes the organism with an approximation of surprisal it does not minimise it. Instead the organism can minimise VFE [variational free energy] further by minimising surprisal indirectly by acting on the environment and changing sensory input" (Buckley et al. 2017: 59). Third and finally, as one reviewer suggested, free-energy theorists may be inclined for technical reasons to endorse, not premise 5, but its converse. If they are right and minimizing free energy is always sufficient for minimizing informational entropy, then the transcendental deduction will be invalid. We'll return on some of these points in §2.4.

Rather than properties of Σ—like being thermodynamically open, or being in equilibrium—it may be useful to speak of (sets of) states. The term *state* here refers to a property instance at a time. Velocities, for example, are states: different objects can have different velocities over time but only a single velocity at any given time. A system's states are specified by the values of sets of measurable macroscopic variables, including geometric, thermal, mechanical, and chemical properties at a given time scale, and can typically be associated with mathematical representations, e.g., vectors in a vector- or state-space. Such variables are related to one another in law-like ways, and so provide a way to deduce changes in the system's behavior over time.

When systems are left to themselves and external conditions are unchanging, their states change until there are no net flows of matter or energy either within the system or between it and its surroundings. This state is thermodynamic equilibrium. Although most systems found in nature aren't in equilibrium, there are important differences across systems. In comparison to snowflakes left to themselves, bacteria are far from their equilibrium state: while their chemical and metabolic properties change, the bacteria can acquire energy and nutrients, and maintain a non-equilibrium steady state to avoid thermodynamic equilibrium.

When systems change states, they undergo a process. The succession of a system's states defines the process's path or trajectory. Some processes involve abrupt discontinuous transitions between solid, liquid, or gaseous states of matter. These transitions are called *phase transitions*. When systems undergo phase transitions, some of their physical properties change—often resulting from changes in temperature, pressure, or other surrounding conditions. For example, snowflakes will phase transition into water droplets when external temperature reaches its melting point. Many biological systems also undergo phase transitions, where their morphological and metabolic profiles change dramatically. These transitions depend on phenotypic plasticity, which is the ability of biological systems with one genotype to change their phenotype in response to changing environmental conditions. Examples include temperature-dependent sex determination in some fish and reptiles, eye-spot formations in some butterflies, and insect metamorphosis.

Because of the second law of thermodynamics, the path of the process undergone by a system isolated from all external influences eventually goes into a state of thermodynamic equilibrium. There are several formulations of the law (Uffink 2001). One formulation refers to thermodynamic entropy, which is a state variable measuring the amount of "disorder" (or randomness) in a system, and implies that entropies of isolated systems not in equilibrium typically increase over time, approaching maximum value at equilibrium. It is a function of a system's state, and captured by the formula: $S = k \log(W)$. This formula describes the entropy S of a system in terms of the logarithm of the number of possible microstates $W = \{w_1, w_2, ..., w_n\}$ that are consistent with the macroscopic states of the system, where k is the Boltzmann constant.

Comparing snowflakes with water vapor offers one intuitive way of visualizing the notion of "disorder" associated with thermodynamic entropy. Water vapor in a container can have many possible arrangements of individual molecules consistent with the macroscopic properties of the gas like its volume and pressure. Because snowflakes' molecules are constrained by crystalline bonds, the number of possible configurations of individual molecules consistent with the macroscopic properties of snowflakes is smaller. Snowflakes have less entropy than water vapor: they are less "disordered" (random).

2.2 From thermodynamics to homeostasis

Free-energy theorists claim that biological systems apparently "resist" or "violate" the second law of thermodynamics because they maintain their physical integrity in the face of random fluctuations in the environment (Friston & Stephan 2007: 421–422; Ramstead et al. 2017: 2 ff.). Biological systems are open systems that maintain "order" (or thermodynamic entropy) by exchanging energy and matter with their surroundings. Upon considering all such exchanges, the total entropy of the system and its environment increases over time in ways that can be described by thermodynamical laws and principles. It's this capacity for "negative entropy"—acting selectively upon their

environments and metabolizing food, which distinguishes living from non-living systems (Collier 1986; Schrödinger 1992; Collier & Hooker 1999; Morowitz & Smith 2007; Bailly & Longo 2009).

A system's capacity to act selectively upon its environment enables the system to maintain its physical integrity amid changing environmental conditions. Friston & Stephan (2007) illustrate this point with a fictitious example. They compare regular snowflakes with winged snowflakes that can act on their environment. Regular snowflakes are passively pushed around by environmental forces until their temperatures reach a certain threshold, where they undergo phase-transitions, losing their integrity and turning into water droplets. Because winged snowflakes can fly and keep certain altitudes, they can maintain their temperature within bounds and away from their melting point. Maintaining their temperature within certain bounds is a necessary condition for the winged snowflakes to keep a non-equilibrium steady state so that they may avoid a phase transition and disintegration. So, winged snowflakes can maintain a "relatively constant milieu" despite environmental changes—that is, maintain homeostasis—via behaviors that adjust to new or changing conditions while maintaining their macroscopic properties within bounds. Because their behaviors enable winged snowflakes to maintain homeostasis, those behaviors, unlike regular snowflakes, are said to be adaptive.

We have two conditions on a system's adaptive behavior. One is in terms of thermodynamics: if systems act selectively upon their environments to preclude phase transitions and stay away from thermodynamic equilibrium, then they behave adaptively. And another is in terms of homeostasis: if systems behave adaptively, then they change their relationships with their environments to maintain vital physiological variables within certain bounds.

Free-energy theorists link adaptivity directly to viability, understood in terms of homeostasis, similarly to Ashby's (1960: 58) account of adaptation as ultrastability did. For free-energy theorists, "characteristics for phase-dependent measurement function"—or as Ashby called them, *essential variables*—must be kept within viable limits to prevent the system from dying rapidly or disintegrating from phase transitions. Like Ashby (1960), free-energy theorists relate

thermodynamic and biological formulations of adaptive processes (see e.g., Friston & Stephan 2007). Specifically, free-energy theorists assume that any living system possesses a random dynamical attractor—a set of states towards which a dynamical system tends to evolve for a wide variety of initial conditions of the system's state. This attractive set is interpreted as the system's extended phenotype, which includes characteristics defining a kind of biological system. Under appropriate conditions, any system possessing a random dynamical attractor can be shown to be formally equivalent to any system at a steady state far from equilibrium, where the system's "characteristic" variables are within homeostatic bounds (Friston 2012; Ramstead et al. 2017).[5] In other words, the paths of the processes of adaptive (living) systems fall within a specific, relatively narrow region of all possible states in their phase space. For Friston, no less than Ashby, survival is equivalent to the system's being in that narrow region.

With these physical and biological formulations articulated, the argument's next step is to deploy the mathematics of random dynamical systems theory and information theory to answer the following question. What characteristics must biological systems possess to maintain their path within a specific (homeostatic) region that precludes phase transitions?

2.3 From homeostasis to surprise

Physical systems can be represented as sets of variables. Different values of these variables pick out different states of the systems. The set of all possible states of a system can be represented as a state space, which allows one to describe the system and its changes in time. A phase space is a continuous state space described with a smooth manifold. A space's dimensionality depends on how many variables are needed to completely describe the target system and its dynamics. Each state of the system is represented with a point in the state space. Given the state of the system at any

[5] Equating adaptation and viability renders Ashby's concept of essential variables unclear. If essential, in what sense can its values be transgressed without causing death? Or, if such transgression is possible, then in what sense are they essential? The construct ALLOSTASIS may offer, here, a way to understand how organisms can operate outside of normal set-points and approximate equilibria values, without leading to the total cessation of all physiological function.

moment, one can use an evolution rule, which can be either deterministic or stochastic, to describe the next states of the system in state space.

Friston (2012, 2013) represents biological systems as random dynamical systems, with state spaces partitioned into external and internal states. External states correspond to environmental causes that generate sensory samples (also known as *sensory input*, *sensory outcomes*, *sensory data*, or *evidence*), which affect the system's internal state. A subset of states (known as *Markov blanket*—more on this momentarily) can ground a separation between the internal states of the system and the external states of the environment. A further subset of these separating (blanket) states is distinguished as its so-called *active states*, where different values of an active state determine different positions of the system in the environment.

The justification for partitioning these states appeals to the construct MARKOV BLANKET (Friston 2013). Pearl (1988) introduced this concept in relation to Markov networks for representing probabilistic knowledge. Roughly, given a set of random variables N, the Markov blanket for a variable $x \in N$ is the subset M containing all random variables that "shield" x from all the other variables in N. Fixing the values of the variables in M leaves x conditionally independent of all other random variables; hence, the Markov blanket of a random variable is the only knowledge one may need to predict the behavior of that variable (Pearl 1988: 97 ff.).

In machine learning, Markov blankets help address problems with constructing sets of causal models from sample data that are as small as possible, and help search these sets to find a true causal model (Spirtes et al. 2000). Friston (2013), however, supposes further that Markov blankets are objective features of the real world separating the states internal to biological systems from those external to them (more on this reification in §3). His mathematical representation of system's exchanges with their environments involves four basic types of quantities (for a helpful review of the mathematical details, see Buckley et al. 2017):

1. a time-varying parameter $\Psi = \{\psi_1, \ldots, \psi_n\}$ standing for environmental states that cause sensory samples and vary nonlinearly over time;
2. a time-varying parameter $A = \{a_1, \ldots, a_n\}$ that changes the way the system samples the environment;
3. a variable $D = \{y_1, \ldots, y_n\}$, defined as a function of the system's active state a_n and environmental state ψ_n, that denotes the set of possible sensory samples that influence the physical state of the system;
4. a statistical model M—which, if Gaussian, can be defined with time-varying parameters $\theta = \{\mu, \sigma^2\}$—of how environmental causes Ψ generate sensory samples D.

For biological systems, free-energy theorists state that M represents an organism's phenotype, defined as "the repertoire of physiological and sensory states in which an organism can be" (Friston 2010: 127), and add that an organism's internal states should be formally representable using the time-varying parameter μ of a generative probability density function M (Friston & Stephan 2007: 424). More specifically, internal states μ are represented by a posterior probability distribution over environmental causes of sensory samples; unlike μ, biological systems would not "encode" or "represent" M, which is (somewhat confusingly) said to be "entailed" by the organism's phenotype (ibid; see also Friston 2012). In this context, *entailment*-talk plausibly means that M should be inferred from a suitable interpretation of the organism's internal states, along with identifying the characteristics defining its kind. Given how free-energy theorists define *phenotype*, the generative model M provides the probability that a certain kind of system obtains any possible state in state space. In short, according to free-energy theorists, for any phenotype, there is a generative model M that renders the internal states of the phenotype as the sufficient statistics of posterior densities of external states under M.

In the winged snowflake example, external states include ambient temperature, wind direction, and other environmental factors that generate sensory samples (or inputs) influencing the

snowflake's internal state. Internal states include its temperature and the strength of its microcrystals' electrostatic bonds, but also active states like its local position in the environment at a time. The snowflake's active state changes its position in the environment (i.e., this active state just is the state of velocity of the snowflake), so that it receives different sensory samples. Given the kind of system a winged snowflake happens to be, it's improbable that states where its temperature is higher than 0° Celsius will obtain.

With respect to cells, external states include ambient temperature and pH. These states generate sensory samples (i.e., energy arrays impinging on organisms' sensory surfaces) that influence the state of transmembrane receptors. The cell's internal states include the concentration of intracellular metabolites, but also active states, like the motion of flagella at times, that change how the cell's environment influences the cell's receptors. Given the kind of system a cell happens to be, it's improbable that states where its temperature is higher than 50° Celsius will obtain.

Given the four basic quantities (for external, sensory, active, and internal states of a biological system) and their relationships of probabilistic conditional (in)dependence, one can examine what characteristics a random dynamical system must possess so that its physical states "are confined to a bounded subset of states and remain there indefinitely" (Friston 2012: 2106). This subset of states in state space is a random dynamical attractor corresponding to a set of (non-equilibrium) steady states, where levels of variables representing the sensory states of the system—like temperature, pH, glucose, blood oxygenation, etc.—are within homeostatic bounds. To learn what characteristics biological systems must possess if they are to maintain their path within a specific (homeostatic) region that precludes phase transitions, free-energy theorists ask what dynamics such systems must exhibit for a random dynamical attractor to obtain.

If one then stipulates that "biological systems move around in their state space, but revisit a limited number of states" that correspond to homeostatic steady states far from equilibrium (Friston 2013: 11), then, of all possible obtaining states, there's a small number that they will achieve in

their lifetime with a high probability. All other possible states will be obtained with an exceedingly low probability.

To capture this idea formally, free-energy theorists use the information-theoretic concept SURPRISE. The average surprise of sampling some outcome corresponds to Shannon's entropy, which is formally similar to the thermodynamic concept ENTROPY. Specifically, the surprise of sampling some sensory outcome can be represented with the negative log probability: $-\log p(Y = y_{t+1} \mid a_t, M)$. This measure quantifies the improbability that a system M samples a sensory outcome y_n, given internal state μ_t and its action a_t. If the sampled sensory outcome is "incompatible" with M and a_t, then the sensory sample y_{t+1} is surprising. If there's a high probability that biological systems are found at any point in their lifetime in homeostatic states, then environmentally-generated sensory samples will be unsurprising. Sensory samples generated by all other states in the environment will be highly surprising.

2.4 From surprise to free energy

If one stipulatively defines *adaptively-behaving system* as any system whose behavior minimizes the average surprise of its possible states, then actual systems behaving adaptively must sample unsurprising sensory outcomes. That is, the system's objective is to maximize model evidence $p(Y = y_{t+1} \mid M)$, or minimize surprise $-\log p(Y = y_{t+1} \mid a_t, M)$. This means, formally, that the system must select actions that optimize the function

$$G(a_t) = \log \int p(Y = y_{t+1}, \Psi = \psi_{t+1} \mid a_t)\, d\Psi$$

where $p(Y = y_{t+1}, \Psi = \psi_{t+1} \mid A = a_t)$ is the joint density of sensory samples Y and their generating causes Ψ in the environment, conditioned on action and on a phenotype.[6] This density factors into a

[6] We have simplified by dropping the dependence on M.

likelihood $p(Y = y_{t+1} \mid \Psi = \psi_{t+1}, A = a_t)$ and prior density $p(\Psi = \psi_{t+1})$, which jointly specify the generative model "entailed by" the system's phenotype. Optimizing $G(a_t)$ requires changes in actions or in the parameter μ that represents the internal state of systems with phenotype M.

However, optimizing $G(a_t)$ involves an intractable marginalization over (hidden) environmental states Ψ. To overcome computational intractability, a variational (or ensemble) density $q(\Psi, \mu)$ can be introduced to define another quantity that is greater than surprise. This quantity—called *free energy*—provides a bound on the integral mentioned above, and is a function of sensory samples and internal states of the system. It is defined thus:[7]

$$F(y_{t+1}, \mu_{t+1} \mid a_t) = -\log \langle p(Y = y_{t+1}, \Psi = \psi_{t+1} \mid A = a_t) \rangle_q + \log \langle q(\Psi = \psi_{t+1}; \mu_{t+1}) \rangle_q$$

Under certain assumptions (cf., Dayan et al 1995; MacKay 1995), optimizing $F(y_{t+1}, \mu_{t+1} \mid a_t)$ is computationally tractable. In the context of FEP, this optimization involves changes only to internal parameters μ or to the action parameter a, which can be controlled by the system.

Because the free-energy function F is greater than G, by acting on the environment to minimize the free energy of their sensory samples, biological systems would indirectly avoid surprising sensory states. If they avoid surprising sensory states, biological systems may attain a homeostatic state; and by selecting actions that attain homeostatic states, biological systems will thereby behave adaptively. And by FEP, any system that minimizes the free energy of its sensory states with respect to action, or its internal parameters, would avoid phase transitions, which would make its physical disintegration unlikely.

3 FEP and organicism

[7] $\langle . \rangle_q$ represents the expectation under the variational density q.

Having articulated the steps involved in the reasoning to FEP, let's examine two dimensions along which FEP and organicism appear inconsistent: their understanding of how, adaptive organisms should be represented, and their interpretation of probability.

3.1 How to represent organisms?

Free-energy theorists formulate FEP using the traditional modeling tools of random dynamical systems in thermodynamics, and represent organisms' adaptive dynamics as trajectories through attractive non-equilibrium states in phase space. For organicists, however, organisms' adaptive dynamics cannot be adequately represented with this tool predefined over the "characteristic" variables individuating kinds of biological systems (Longo et al. 2012).

Organicists have different options for justifying this claim. They may deny the existence of an organism's characteristics, or deny that characteristic variables (if they exist) can be reliably identified for any kind of biological system interacting with the environment. Another option is to emphasize that the mathematical tools used to represent and explain biological phenomena are merely that—abstract mathematical tools. For instance, Chater & Oaksford argue, albeit in a different context, that imputing these modeling tools to the phenomena themselves by requiring organisms to perform these calculations mischaracterizes how such principles are used to explain behavior:

> the theory of aerodynamics is a crucial component of explaining why birds can fly. But clearly birds know nothing about aerodynamics, and the computational intractability of aerodynamic calculations does not in any way prevent birds from flying. Similarly, [systems] do not need to calculate their optimal behavior functions to behave adaptively. They simply have to use successful algorithms; they do not have to be able to make the calculations that would show that these algorithms are successful. (2000: 110)

The suggestion is that these tools aren't part of organisms' biological or cognitive equipment. If so, this would bear directly on the posit of free energy, which free-energy theorists introduce because the optimization problem described in §2.4 is intractable. Yet, if there's no imperative for organisms to compute solutions to the optimization problem, then the motivation to posit free energy dissipates.

Free-energy theorists assume that organisms are ergodic: their phase averages (i.e., average values of specified functions of their microscopic states) are identical to averages over time of quantities measurable from microscopic states. For ergodic systems, the dynamics of a system's microstates—e.g., its biomolecular kinematics—are sufficiently random, and the coupling between the system and the external states in its environment is sufficiently slow, such that the microscopic dynamics of the system's states can be replaced by a random sample from the ensemble density of the microscopic states. This ensemble density assigns a probability to each possible microscopic configuration, and can be used to derive predictions about macroscopic properties of the system as those expected in the ensemble. To assign probabilities to all microscopic configurations, the system is assumed to start in any microscopic state and traverse all possible microscopic states in phase space. Given ergodicity, for any region of phase space, the average time the system spends in that region is proportional to the region's size. Ergodic systems will repeatedly revisit the neighborhoods of attracting states. So, if organisms are ergodic, they can be represented as random dynamical attractors in phase space. If such ergodic dynamical systems also possess a Markov blanket, then "[they'll] appear to actively maintain their structural and dynamical integrity" (Friston 2013: 10). Under certain conditions, such systems necessarily minimize their free energy.

Organicists will be quick to note that organisms, at all levels of organization above the level of molecules, cannot explore all possible paths. "Not only will we not make all possible proteins of length 200 or 2000, we will not make all possible organs, organisms, social systems, [... t]here is an indefinite hierarchy of non-ergodicity as the complexity of the objects we consider increases"

(Kauffman 2013: 167). For organicists, ergodicity is biologically irrelevant simply because organisms are non-ergodic (Longo et al. 2012).

If ergodicity were biologically irrelevant, and the average of any measure of the state of an organism doesn't converge over a sufficient period of time, then it is misleading to represent an organism's phenotype with an invariant, ergodic ensemble density that specifies the probability, for any possible microstate, that the organism is in a certain macrostate. And it would also be biologically irrelevant to note that, under ergodic assumptions, the long-term average of surprise is entropy. For organicists, FEP would thus be misleading and biologically irrelevant.

Free-energy theorists, however, would counter by asserting that the notion of an attracting set itself implies ergodicity. So, if all living systems possess an attracting set, then they must be ergodic too; and if those systems are ergodic, then they will possess characteristic measureable properties. For free-energy theorists, the dynamics of such systems will appear to place an upper bound on their informational entropy, and to maximize the evidence for a model M of external states "entailed" by their characteristic properties. This behavior—they would conclude—can be expressed as approximate Bayesian (active) inference about the causes of sensory input in terms of minimizing variational free energy.[8]

On the other hand, organicists have argued that organisms constitute historically grounded constraints on energy flows, where their phase space continually changes, and so organicists deny that living systems are aptly represented with a classical phase space. Since attracting sets are subsets of classically predefined phase spaces, organicists deny the assumption that all living systems' characteristic behavior is aptly represented with an attracting set. Hence, for organicists, the implication from attracting sets to ergodicity is a red herring.

[8] This inference-optimization bridge, which we also mentioned in fn. 4, is a powerful feature of variational methods that treat statistical inference problems as optimization problems. Minimizing variational free energy can thus be understood as approximate Bayesian inference, in the sense that minimizing free energy is formally equivalent to optimizing a variational bound on Bayesian model evidence.

Representing organisms and their dynamics while idealizing away from their non-ergodic status would prevent taking seriously the historical considerations of lineage that are essential to understanding what organisms are, and how they change over time. If historical considerations and lineage matter to understanding organisms and their dynamics, then biological systems should be represented as "specific" and their trajectories as "generic." Instead, free-energy theorists get it backward: physical systems are "generic," while their trajectories "specific" (Longo & Montévil 2013: ch. 7).

If physical systems were "generic," then different types of systems could be individuated by properties such as mass, charge, temperature, or momentum. Position and momentum, for example, define mechanical systems. The mathematical representation of physical systems of the same type will preserve formal symmetries between their defining characteristics.[9] Each symmetry in mathematical representation implies that certain physical properties of the system, such as total kinetic energy, are conserved and remain unchanged as the system evolves over time (Gross 1996). However, for tokens of the same type of physical system, trajectories (or paths) would be "specific" since they are uniquely determined by the rules of the system's evolution given its initial state.

In free-energy theorists' representations, organisms are "generic." Their representation in terms of actions, internal states, and generative models preserves symmetries in formulations of FEP. Trajectories in the phase space of organisms would be "specific" according to FEP, because solutions of FEP for a given kind of organism yield a unique trajectory in phase space given the initial state of the organism.[10] Properties of organisms like lineage and heritability would just be an expression of a specific trajectory on a generic manifold, namely the attracting set.

[9] For example, in Newtonian mechanics, given two bodies with the same mass starting from rest and moving in opposite directions with different velocities along the same axis, the total kinetic energy of the systems comprised of the two bodies remains the same if the velocities are interchanged. That is, the solution of the equation yielding the total kinetic energy will be the same if the velocities of the two bodies are interchanged.

[10] Friston (2012) relates this result to the principle of least action (PLA). In Hamilton's formulation, *action* is defined as the integral along possible paths of a system's process connecting two specified states. According to PLA, the actual path of the process between initial and final states in a specified time is a dynamical system's trajectory in phase space, which is found "by imagining all possible trajectories that the system could conceivably take, computing the action for each of these trajectories, and selecting one that makes the action locally stationary" (Gray 2009).

For Longo & colleagues (2012), organisms should instead be represented as "specific," while their trajectories "generic." Because of their historicity and materiality, organisms wouldn't possess general characteristics that allow for mathematically invariant representations. They write,

> In biology, symmetries at the phenotypic level, are continually changed, beginning with the least mitosis, up to the "structural bifurcations" which yield speciations in evolution. Thus, there are no biological symmetries that are a priori preserved. […] There are no sufficiently stable mathematical regularities and transformations to allow an equational and lawlike description entailing the phylogenetic and ontogenetic trajectories. (2012: 1390)

So, from an organicist's perspective, organisms cannot—unlike non-living physical systems—be represented with a fixed phase space and rules of evolution predefined over some set of "characteristic," mathematically invariant variables.

The disagreement between free-energy theorists and organicists cuts deeper still. Free-energy theorists take an adaptationist and selectionist standpoint in biology, whereby "selection explains how biological systems arise and the only outstanding issue is what characteristics they must possess" (Friston & Stephan 2007: 423). Central to the organicist approach is instead a critical rejection of the adaptationist and selectionist perspective, which—they argue—is insufficient for explaining the autonomy of living beings, and their capacities to regulate their processes in relation to environmental conditions registered as viable or unviable, improving or deteriorating (Di Paolo & Thompson 2014; Moreno & Mossio 2015).

Furthermore, following Ashby (1940), FEP emphasizes homeostatic stability as the core feature of organisms. Accordingly, organisms should be represented as random dynamical attractors. Contemporary organicists like Longo & colleagues emphasize that organisms are fundamentally ever-changing processes, maintained in relatively stable conditions by further processes. Longo & Montévil (2013), for instance, advance the idea that organisms should be

represented as attaining extended critical phase transitions. Free-energy theorists generally construe undergoing a phase transition as equivalent to biological disintegration and death—although there are several examples of phase transitions in biology, such as metamorphosis, that are consistent with an attracting set, and, arguably, with the FEP too (Clark 2017).

Even so, Longo & Montévil (2013) argue that organisms' adaptive behavior should be represented as a continuous critical transition from one phase to another, whereby organisms are continuously reconstructed with variations in their dynamic couplings with ecosystems. The basic idea is that ancestry continuously co-constitutes and reshapes organisms as well as by their interactions with ecosystems. This idea fits a growing wealth of evidence concerning phenotypic plasticity, whereby organisms with the same genotype can generate differing phenotypes through their interactions with the environment, which recreate novel conditions of existence passed on to their descendants (Montévil et al. 2016).

Free-energy theorists would again note that evolution, and natural selection in particular, is also a free-energy minimizing process (Hobson & Friston 2016; Ramstead et al. 2017). But, for organicists, selectionism, and the fixation on surprise-minimizing processes obscure the historically-grounded, environmentally co-constituted nature of biological adaptivity.

One last dimension of disagreement concerns the interpretation of the probabilities involved in scientific representations of changing organisms. In free-energy theorists' accounts, probability plays a central role. Adaptive organisms are said to be embodied generative models of their environments, where a generative model *M* specifies the probability that a certain external state, sensory input, and internal state occur together (Friston 2013). Organisms are said to have Markov blankets separating their internal states from the external environment, defined over a set of random variables and a probability measure (Kirchhoff et al. 2018). Furthermore, Friston (2009, 2010) claims FEP entails the Bayesian brain hypothesis, which implies that nervous systems represent probability distributions, store generative models and prior probabilistic knowledge about the world, and that neural networks can perform statistical inferences based on these probabilities

(Knill & Pouget 2004; Colombo & Seriès 2012). Finally, free energy itself is a bound on surprise (or information entropy), and surprise is a probabilistic measure of the uncertainty of sampling some sensory data, given a generative model.

Neglect of both the differing scope of distinct formulations of FEP (as applied to non-biological adaptive systems, organisms, or brains at a certain time scale) and the interpretations of the probabilities in different formulations of FEP has generated disagreement—particularly about whether FEP is committed to positing mental representations and to an essentially inferential picture of cognition. To move the debate forward, at least two questions should be distinguished. First, should we understand FEP as a modeler's tool to characterize and predict adaptive behavior, or should it be understood as an objective feature of target systems? Second, in determining the scope of FEP, how should the probabilities it posits be interpreted? Should they be understood epistemically or physically?

Some researchers in the organicist tradition seem to interpret the probabilities involved in FEP physically, as the frequency or propensity of the occurrence of some event (e.g., Bruineberg et al. 2018; Kirchhoff & Froese 2017). When FEP targets brains, these researchers reject the idea that brains literally represent probabilities and draw inferences. Other researchers have interpreted the probabilities involved in FEP epistemically, as rational degrees of belief in the occurrence of some event and the willingness to act on this belief. When FEP targets brains, these researchers tend to suggest that brains literally represent probabilities, possess Markov blankets, and draw inferences (e.g., Hohwy 2016). For his part, Friston (2013) seems to interpret the probabilities involved in FEP as objective features of real-world systems; they aren't just modeling tools. When FEP targets brains, he claims that brains represent probabilities about the occurrence of events, and that brains make inferences about the causes of their sensory inputs, based on neurally encoded statistical models (Friston 2011, 2013).

Let's consider the first idea. In formulating FEP, one need initially specify a system's phase space (or sample space) and a desired equilibrium state distribution over it. One can then optimize

the equilibrium state distribution by minimizing the free energy of samples generated by the environment with respect to actions and a generative model. The system—claims Friston (2011, 2013)—will then appear to sample its environment *as if* it were aiming at maximizing the evidence for its own existence. However, the construction of explanations thereof may fall short when the explanatory goal is construed as specifying the real nature of biological explananda; and where the functional capacities attributed to those systems prove intractable, resorting to 'as if' explanation won't circumvent the problem, even when the computations involved are construed as subsymbolic, offline, heuristic, or approximate (van Rooij et al. 2018).

The desired steady-state distribution is, by definition, just the organism's evolved equilibrium state distribution. This means that minimizing free energy would be equivalent to maximizing expected adaptive value: surprising states would just be, by definition, maladaptive. One way to support this definition is to note that organisms, and brains, are immersed in a "statistical bath" of energy arrays from the environment impinging on their sensory surfaces. These energy arrays would sculpt phenotypes, associated with a certain equilibrium steady-state distribution, which the organism would update as a function of its sensory samples.

However, organicists will argue that defining *adaptivity* as unsurprisingness is misguided and relies on implausible assumptions. With Longo & colleagues (2012), one may note that a desired equilibrium state distribution is only definable on a predefined phase space. If it's not possible to predefine a phase space for biological systems, the definition will be undercut—and with it, the justification for believing that the probabilities involved in FEP are not mere modeling tools but are objective features of living systems.

Furthermore, leaving on the side free-energy theorists' selectionist perspective, one may find implausible that surprising events should always be maladaptive. Using Gershman & Daw's examples, "[s]hould the first amphibian out of water dive back in? If a wolf eats deer not because he is hungry, but because he is attracted to the equilibrium state of his ancestors, would a sudden bonanza of deer inspire him to eat only the amount to which he is accustomed?" (2012: 306). If

adaptivity is stipulatively defined in terms of minimization of surprise, then some instances of adaptive behavior will consist in seeking out novel situations that may provide systems the opportunity to resolve expected surprise.

Let's assume that the probabilities involved in FEP aren't simply modelers' tools, and turn to the second question about how to interpret these probabilities. In its maximal scope formulation, the probabilities involved in FEP should be interpreted physically, as objective propensities or frequencies. (After all, it makes little sense to say that an amoeba has a certain degree of belief that a bacterium is in the premises.) Physical interpretations of probabilities cohere both with how probabilities are generally understood in statistical mechanics, as well as with the suggestion that inference talk is inapposite to the behavior of organisms like an amoeba or bacterium, understood as dynamical systems coupled with their environment (Bruineberg et al. 2018).[11] Finally, a propensity-based interpretation clarifies the sense in which FEP isn't a tautology, just like it isn't a tautology to say that dice produce odd numbers more often than threes—more on tautology and FEP in §4.1. Propensities of free-energy minimizing organisms to survive, or of dice to fall equally often on each side, permit fallible predictions about their behaviors.

Yet, a propensity interpretation of the probabilities involved in FEP raises several issues. One is that it's opaque what sort of property a propensity is. If propensities are causal tendencies,

[11] The notion of inference in FEP also requires clarification. Bruineberg et al. claim that, "[w]ithin the Free Energy framework, the notion of 'inference' is much more minimal and does not involve any propositions: any dynamical system *A* coupled with another *B* can be said to "infer" the "hidden cause" of its "input" (the dynamics of *B*) when it reliably co-varies with the dynamics of *B* and it is robust to the noise inherent in the coupling" (2018: 2436). They invoke Huygens's case of the synchronization of two pendulum clocks to illustrate this "minimal" sense of inference. Three points in response. First, whether or not the concept INFERENCE in FEP involves propositions depends how to understand the probabilities featuring in the free energy framework. If these probabilities are modelers' tools without actual counterparts in target systems, then INFERENCE may involve propositions, viz., those propositions entertained by scientists when they make inferences about target systems on the basis of the mathematics of random dynamic systems theory. Second, reliable covariance robust to noise falls within the extension, not of the concept INFERENCE, but of some successor notion; so Bruineberg et al. have not illustrated a "minimal sense" so much as they've just changed the subject; and doing so may do an injustice to actual scientific practice, where the topic of statistical inference for dynamical systems is studied extensively across several fields. Third, Huygens observed phase/-opposition coupling between two pendulum clocks hanging from a beam or from a board sitting on two chairs. But FEP is formulated within the mathematics of random dynamical systems, and Huygens's case has not been aptly represented as a problem in parameter estimation in dynamical systems—a problem that FEP is supposed to help solve (see Oliveira & Melo 2015). So, the question of how INFERENCE should be understood within the free-energy framework is not settled by Bruineberg et al.'s claim.

then they should be asymmetric and diachronic relationships just like causal relationships. But then they cannot be probabilities, since conditional probabilities are symmetric. If propensities are long-run relative frequencies, then a reference class should be defined. For biological organisms, such a reference class would correspond to a phase space; but if a phase space cannot be predefined for an organism, the probabilities involved in FEP cannot be defined.

Suppose instead that the probabilities in FEP are understood epistemically as rational degrees of belief. This interpretation is most plausible when FEP is restricted to whole, cognitively-sophisticated animals, i.e., to creatures possessing mental representations and capacities for rational inference. Organisms would thus act adaptively by minimizing a free-energy bound defined over internal, graded, epistemic states representing events in the environment.

While this interpretation wouldn't mitigate the problem of defining a phase space (or sample space) for organisms, Friston and colleagues surmise that "sustained exposure to environmental inputs causes the internal structure of the brain to recapitulate the causal structure of those inputs. In turn, this enables efficient perceptual inference" (Friston et al. 2006: 77). This suggestion risks obfuscating that adaptive pressures on biological observers are generally unrelated to the accuracy of their epistemic states. The suggestion is also at odds with orthodox Bayesianism, since it dovetails the assumption that Bayesian inference is "efficient" or optimal just in case one's prior beliefs match the actual statistics of the environment (Feldman 2017).

4 FEP and mechanism

We now turn to mechanistic philosophy, and argue that FEP is inconsistent with mechanism along two dimensions of representation: the dependence of explanatory force on describing mechanisms and the rejection of the idea that life science phenomena can be adequately explained through an axiomatic, physics-first approach.

4.1 How to explain life phenomena?

Free-energy theorists appeal to FEP to attempt explanations of various phenomena, including Hebb's rule and spike-timing dependent plasticity, the multiplicity and hierarchical organization of cortical layers, their reciprocal connection with distinct feedforward and feedback properties, and the existence of adaptation and repetition suppression (Friston 2010a). Such explanations are thought of axiomatically, as logical deductions from sets of axioms and formulae (Friston 2012). Stipulative definitions, like *living system* as an attracting set in a phase space or *adaptive* behavior as behavior that reduces average surprise, provide the bridge principles that connect theoretical predicates from different disciplines, and that allow free-energy theorists to attempt the deductions needed to claim reductions of other principles to FEP. In their sweeping attempt to explain these and other phenomena, and to reduce their theories and principles to FEP, free-energy theorists have also claimed for their theory another virtue: a grand theoretical unification (Friston 2010a; Hohwy 2014).

In presenting these features, FEP is apparently at odds with mechanists' emphasis that life science phenomena should be explained by appeal to mechanisms, and that adequate strategies for explanation in the life sciences should involve decomposing these mechanisms into component parts and operations and providing an account of how these parts and operations work together to produce the phenomenon.

FEP and mechanism are related by their common emphasis on function. Both free-energy theorists and mechanists can agree that FEP provides an idealized functional principle. Where they diverge is on the issues of whether and when functional accounts of biological phenomena suffice for adequate explanation, or whether structural details of component parts and mechanistic organization are also necessary for such functional accounts to have explanatory power.

Progress in the life sciences—and especially the cognitive and behavioral sciences—often begins by empirically identifying and adequately describing a functional capacity. Descriptions of that capacity then allow for the method of functional analysis, wherein that capacity is functionally decomposed in terms of its constituent properties, processes, and components. Functional properties

of a system have traditionally been individuated by their relations to inputs, outputs, and other internal properties of the system under investigation—that is, by their causal role. By redescribing systems' capacities in terms of their functional properties and dispositions, functional analysis offers scientists a way to tackle the target phenomenon. But it also offers the potential for prediction and explanation, as empirical applications of FEP demonstrate (e.g., Bastos et al 2012).

FEP provides a functional analysis of adaptive behavior, in the sense that it derives from a stipulation that such behaviors must be surprise-minimizing. In the abstract, FEP requires an interplay between four sets variables: sensory samples D that influence a system Σ's internal states, and active states A that influence the external states of the local environment Ψ. The mathematical dependencies between $<\Sigma (D, A), \Psi>$ define free energy over Σ as a function of D and an approximate probabilistic representation of its causes—the thought being that a system's capacity to adapt to its environment can be functionally analyzed in terms of suppression of free energy, via internal representations that readjust state changes in its transducers (to maintain or improve perceptual fidelity) or effectors (to maintain or promote successful control).

For their part, mechanists have argued at length that functional analyses lack explanatory power, as they are mechanism sketches. The notion of a mechanism sketch is that of an incomplete representation of (the function of) a mechanism, in which some relevant aspects—either component parts or their operations, or their organization—are omitted from the explanation. Biological or cognitive functions may be decomposed into their constituent properties, processes, or subroutines; and decomposition may detail a capacity as a nexus of functional relationships between variables standing for component operations, without thereby specifying how the capacity is actually realized. Mechanists acknowledge that functional decompositions of capacities into modeled causal and non-causal operations within a mechanism constrain the possible structures and configurations that might perform those operations; but they are equally keen to emphasize that structural decompositions into modeled components within a mechanism can also constrain the possible functions and configurations performed. Details about the relevant physiological and anatomical

components of mechanisms are necessary to filling in mechanism sketches by localizing each operation to its respective component part (Bechtel & Richardson 1993/2010).

Free-energy theorists' functional analyses appeal to states and dynamics that idealize away from the biophysical details of the structural complexity of actual systems. While functional analyses may be construed as mechanism sketches, FEP itself doesn't provide in any obvious sense a sketch of a mechanism, since it swings free of the need to supply any microstructural, biophysical, or anatomical detail. Mechanists would therefore conclude that FEP lacks explanatory power (Kaplan & Craver 2011).

Free-energy theorists may reply that FEP defines a class of process models that provide hypotheses about how spatiotemporally organized components and operations in biological systems might carry out free-energy minimization. For example, some have suggested that predictive coding is one mechanism by which FEP works, in which hierarchically-structured neuronal assemblies engage in message-passing operations (Friston 2009; Bastos et al. 2012). Higher-order neuronal assemblies would output predictions of the states of lower-order assemblies, which are then compared with the actual states of the lower-order assemblies to form prediction errors that are passed back up the hierarchy to update the predictions from higher-order neural assemblies. The recurrent exchange of signals between adjacent neural assemblies resolves prediction error at each level, resulting in hierarchically deep, neurally-encoded "accounts" of sensory inputs.

While it may go some way toward meeting the demands of filling out a mechanism sketch, this suggestion is inadequate for mechanists. First, predictive coding is only one of several possible algorithms that may be used to optimize energy functions. Since many of them might be empirically adequate but difficult to disentangle, concerns of underdetermination may emerge. Second, while predictive coding has been used to model some aspects of visual perception (Rao & Ballard 1999), "the experimental evidence for it seems currently inconclusive in the sense that it does not rule out Bayesian inference with a direct variable code, potentially in combination with a variety of non-probabilistic processes" (Aitchison & Lengyel 2017: 224). Third, because FEP is intended to

generalize beyond organisms with a nervous system, predictive coding would need to be a mechanism by which all adaptive systems work—from bacteria, to winged snowflakes and bladder cells, to plants and social networks. Obviously, for such a wide array of systems, appealing to message passing in neuronal hierarchies as a relevant mechanism by which they operate is insufficiently general. Fourth and finally, even if issues of underdetermination were put aside and the experimental evidence for predictive coding were overwhelming, and even if it were shown to be the generalized mechanism by which all adaptive systems work, mechanists would be positioned to claim that filling out the mechanism sketch is what matters: appeals to the mechanism of predictive coding—not FEP—are what provides explanatory depth. For mechanists, functional descriptions of capacities that require mechanistic analysis to achieve this depth can be important principles, but not foundational ones that do the heavy lifting in ultimately explaining biological phenomena.

Free-energy theorists may simply reject the idea that adequate scientific representation of life science phenomena must target the component parts and operations and internal organization of mechanisms. They may refer to Chirimuuta's (2017) work, which contends that several explanatorily adequate models in computational neuroscience are non-mechanistic. The models Chirimuuta considers would be instances of "efficient coding explanation," which, abstracting away from biophysical specifics, would answer why certain neuronal systems should behave in the ways described by the models. Based on design principles informing the model, such explanations would thus identify the functional utility of general patterns of behaviors instantiated by neural systems.

Similarly, free-energy theorists may argue that FEP is an optimality principle in the life sciences (cf., Rice 2015). FEP, along with the class of process models it defines, provides us with idealized, coarse-grained descriptions of certain factors and functional variables, which leverage varieties of realization types against the drive to detail lower-level biomechanical structures essentially involved in producing target explananda-phenomena. FEP, and the class of process models it defines, would omit causal-mechanical detail and make biologically unrealistic

assumption to focus attention on very general observable patterns displayed by biological systems, but also by any other system behaving adaptively. Given that aim—they might claim—it's misguided to charge FEP for not providing us with mechanistic information.

4.2 FEP as a first principle?

As we have seen, explanations derived from FEP abstract away and distort most of the mechanistic features of their target phenomena. But if FEP doesn't aim at uncovering mechanisms or difference-makers, then what's its epistemic status vis-à-vis mechanistic explanation, exactly?

Unfortunately, FEP's epistemic status is muddled. It has been called an *unfalsifiable platitude*, an *imperative*, a *tautology*, a *stipulative definition*, *paradigm*, *law of the life sciences*, *law of nature*, an *a priori first principle*, a *unifying explanation*, and a *simple postulate or axiom*. Wiese & Metzinger assert that "FEP can be regarded as the fundamental theory, which can combine the different features of predictive processing described above within a single, formally rigorous framework" (2017: 12). Friston and collaborators contend that "free energy minimization may be an imperative for all self-organizing biological systems" (2012: 2117), and that "the whole point of [FEP] is to unify all adaptive autopoietic and self-organizing behavior under one simple imperative; avoid surprises and you will last longer […]", which is a principle so basic that "there is no need to recourse to any other principles" (Friston et al. 2012), and "The tautology here is deliberate, it appeals to exactly the same tautology in natural selection (Why am I here? – because I have adaptive fitness: Why do I have adaptive fitness? – because I am here). Like adaptive fitness, the free-energy formulation is not a mechanism or magic recipe for life; it is just a characterization of biological systems that exist" (ibid.). For his part, Allen (2018: 19) characterizes FEP as *a normative theory, an axiomatic, self-evidently true natural law,* and *a tautologically true axiom,* likening it to a *paradigm*, *framework,* and a *research programme* as well.

This rhetorical jumble makes it harder to understand the status of FEP as a first principle, and thus the contrast between FEP and mechanism. So, some clarification is called for. To begin,

one should not follow Allen or Wiese & Metzinger in confusing *principles* with *theories*, *paradigms*, and *research programmes*, since these technical terms refer to different species of scientific representations, with different properties and scientific and philosophical purposes. Likewise, that any self-conserving system, via environmental exchanges, must, as necessary condition on the possibility of maximizing its adaptivity, minimize an information-theoretic bound on a negative log probability is an intriguing thought—but not one that qualifies as a platitude or a truism, under any normal understanding of those concepts.

One plausible thought is that FEP is a first principle because it's an axiom or postulate: "[t]he free-energy principle is a simple postulate that has complicated ramifications" (Friston 2011: 91), and again, "FEP derives [is?] a normative, a priori first principle from a provable definition of living systems" (Allen & Friston 2018: 2473).[12] And in fact, as mentioned, free-energy theorists spin off an enormous variety of derivations from FEP. In that sense, FEP may be said to play the role of a first principle. But while the free-energy theorists in the life sciences utilize an axiomatic approach grounded in the mathematics of theoretical physics (Friston 2012), FEP is a principle that is itself derived from other statements and definitions. So, it is not an underived axiom or postulate, strictly speaking. By implication, principles like FEP need not be underived axioms to play the role of first principle in the life sciences.

Free-energy theorists often characterize FEP as a principle in the life sciences because it's "tautological," even "unfalsifiable." Such claims are not easily interpreted. If FEP were a tautology, then free energy theorists may have a triviality problem. As Klein remarks, "[a]ppeal to apparent tautologies should trouble you. For whatever tautologies do, they don't explain why things happen" (2018: 2552). Tautologies may offer a starting point for explaining the end goal states of optimal systems, but the empirical adequacy of models with added biophysical causal detail is far removed

[12] It's unclear in what sense the definition of *living system* as an attracting set is "provable." Definitions are commonly restatements that do not stand in need of proof, or are even capable of being proved. More plausibly, Friston & colleagues want to claim that the definition they propose is legitimate—though it's not obvious what criteria of legitimacy should be in place to evaluate this claim.

from triviality. And the attempted explanations relying on FEP do not involve anything like the decomposition and localization of biophysical mechanisms underlying adaptive behavior.

Similarly, falsifiability is normally treated as a hallmark of any scientific claim; so if principles must be falsifiable to be scientific and FEP is unfalsifiable, then FEP is not a scientific principle. What's intended cannot be that FEP is unfalsifiable because it's merely stipulated; for while there are stipulative definitions involved in the transcendental argument, FEP is not one of them. Similarly, what's intended cannot be that FEP is unfalsifiable because it fails to be truth-apt, since it would then not be a law-like generalization, and could not serve as the conclusion of a transcendental argument. Presumably, then, FEP is like all other scientific principles in being truth-apt, such as Archimedes's principle describing basic relationships in fluid dynamics. But unlike other principles such as Galileo's principle describing the periodicity of pendula, which subsequently enjoyed more accurate and precise formulations, it seems that what's intended is that principles like FEP or Hamilton's PLA survive all scrutiny of their pedigree and have truth-values that cannot be improved upon. In that sense, FEP might be a constraint that mechanist explanations in the life sciences must honor, in so far as life scientists aim to determine which values of some energy functional constitute the best available solution—given certain design constraints—to the problem of maintaining the path of a target biological system in state space within a specific (homeostatic) region that precludes phase transitions. Claims of tautologousness and unfalsifiability, in addition to being interpretively difficult, just lead to further questions about its epistemic status, such as whether free-energy theorists, in taking FEP to be a necessary condition on the possibility of adaptive behavior, thereby take FEP to be some kind of necessary truth.

Unlike what advocates have claimed (e.g., Hohwy 2014), FEP-based theorizing is not consistent with mechanists' idea that the power of an explanation depends on its capacity to uncover the mechanism of a target phenomenon. But unlike what mechanists will reply, the response is not obviously just to claim, "so much the worse for FEP". This apparent inconsistency is methodologically good: the boom of research relying on FEP just highlights there is room for

deductive systematization and physics-first approaches in life science theorizing (see also Gurova 2011). Rather than aiming to represent difference-makers of life science phenomena, FEP instead aims to represent, in the language of information theory and random dynamical systems theory, what characteristics complex systems must possess for self-maintenance and self-regulation.

5 Conclusion: a plurality of principles

Living organisms are complex, adaptive systems that present both robust regularities but also constant variation. Understanding their regularities and variations requires conceptual frameworks in which knowledge from physics, chemistry, biology, ecology, and ethology can be synthesized, using tools from mathematics and computational theory. The diversity of expertise involved in understanding brains and organisms, and the fragmentation in present-day neuroscience and biology, highlights the need for principles that could afford a common intellectual framework for researchers from different communities to work together to answers questions of common concern.

FEP is an impressive candidate for one such first principle aiming to ground general, idealized models for tracking one fundamental pattern underlying the robust regularities and constant variation displayed by complex and diverse phenomena in the life sciences. FEP symbolizes what a first-principle, physics-first, axiomatic approach to the life sciences can look like, while it has crystallized the notion of prediction-error minimization in philosophical and scientific debates as a central theoretical posit to understanding life and mind.

In this paper, we have identified apparent disagreements between FEP and basic tenets of organicism and mechanicism concerning the scientific representation of life phenomena. The incompatibility between these different approaches should not suggest that only one of these approaches can aptly represent the phenomena of life. These different approaches are meant to fulfill different epistemic aims of different communities of life scientists; and these aims may be best pursued separately, with a diverse array of tools for piecemeal modeling, prediction, and understanding of target phenomena.

Motivated by the kind of pragmatic, epistemic pluralism endorsed by fundamental inquiries like Smith & Morowitz (2016) and by philosophers of the life sciences like Mitchell (2002), we conclude with a note of caution against indulging in metaphysical speculation on the basis of scientific tools for modeling and representations of the phenomena of life (cf., Potochnik 2017: §7.2). While we may legitimately and productively argue about the aptness of a tool for its intended purposes, the risk of confusion is high when philosophers and scientists directly reads off metaphysical conclusions about the nature of life and mind from usage of epistemic tools.

Acknowledgments

...

References

Aitchison, L. & Lengyel, M. (2017). With or without you: predictive coding and Bayesian inference in the brain. *Current Opinion in Neurobiology*, *46*, 219–227.

Allen, M. & Friston, K. (2018). From cognitivism to autopoiesis: towards a computational framework for the embodied mind. *Synthese*, 195, 2459–2482.

Ashby, W. (1960). *Design for a Brain*. Chapman & Hall.

Ashby, W. (1940). Adaptiveness and equilibrium. *British Journal of Psychiatry*, *86*, 478–483.

Bailly, F. & Longo, G. (2009). Biological organization and anti-entropy. *Journal of Biological Systems*, *17*, 63–96.

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

Bechtel, W. & Richardson, R. (1993/2010). *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. MIT Press.

Brandon, R. (1984). Grene on mechanism and reductionism: more than just a side issue. *PSA 1984*, *2*, 345–353.

Bruineberg, J., Kiverstein, J., & Rietveld, E. (2018). The anticipating brain is not a scientist: the free-energy principle from an ecological-enactive perspective. *Synthese*, 195, 2417–2444.

Buckley, C., Kim, C., McGregor, S., & Seth, A. (2017). The free energy principle for action and perception: a mathematical review. *Journal of Mathematical Psychology, 81*, 55–79.

Chater, N. & Oaksford, M. (2000). The rational analysis of mind and behavior. *Synthese, 122*, 93–131.

Chirimuuta, M. (2017). Explanation in computational neuroscience: causal and non-causal. *British Journal for the Philosophy of Science*, axw034.

Clark, A. (2017). How to knit your own Markov blanket. In T. Metzinger & W. Wiese. (eds), *Philosophy and Predictive Processing*. Open MIND Group.

Collier J. (1986). Entropy in evolution. *Biology and Philosophy, 1*, 5–24.

Collier, J. & Hooker, C. (1999). Complexly organized dynamical systems. *Open Systems and Information Dynamics*, *6*: 241–302.

Colombo, M. (2017). Social motivation in computational neuroscience. Or if brains are prediction machines, then the Humean theory of motivation is false. In J. Kiverstein (ed.) *Routledge Handbook of Philosophy of the Social Mind* (320–340). Routledge.

Colombo, M. & Wright, C. (2017). Explanatory pluralism: an unrewarding prediction error for free energy theorists. *Brain and Cognition, 112*, 3–12.

Colombo, M., & Seriès, P. (2012). Bayes in the brain—on Bayesian modelling in neuroscience. *The British Journal for the Philosophy of Science, 63*, 697–723.

Darden, L. (2006). *Reasoning in Biological Discoveries: Mechanisms, Interfield Relations, and Anomaly Resolution*. Cambridge University Press.

Dayan, P., Hinton, G. E., Neal, R., & Zemel, R. (1995). The Helmholtz machine. *Neural Computation, 7*, 889–904.

Di Paolo, E. & Thompson, E. (2014). The enactive approach. In L. Shapiro (ed.), *Routledge Handbook of Embodied Cognition* (68–78). New York: Routledge.

Feldman, J. (2017). What are the 'true' statistics of the environment? *Cognitive Science*, *41*, 1871–1903.

Fiorillo, C. (2010). A neurocentric approach to Bayesian inference. *Nature Reviews Neuroscience*, *11*, 605.

Friston, K. (2013). Life as we know it. *Journal of the Royal Society Interface*, *10*, 20130475.

Friston, K. (2012). A free energy principle for biological systems. *Entropy*, *14*, 2100–2121.

Friston, K. (2011). Embodied inference: or I think therefore I am, if I am what I think. In W. Tschacher & C. Bergomi (eds.), *The Implications of Embodiment* (89–125). Imprint Academic.

Friston, K. (2010a). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, *11*, 127–138.

Friston, K. (2010b). Is the free-energy principle neurocentric? *Nature Reviews Neuroscience*, *11*, 605.

Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences*, *13*, 293–301.

Friston, K. (2003). Learning and inference in the brain. *Neural Networks*, *16*, 1325-1352.

Friston, K., Kilner, J., & Harrison, L. (2006). A free energy principle for the brain. *Journal of Physiology Paris*, *100*, 70–87.

Friston, K. & Stephan, K. (2007). Free-energy and the brain. *Synthese*, *159*, 417–458.

Friston, K., Thornton, C., & Clark, A. (2012). Free-energy minimization and the dark-room problem. *Frontiers in Psychology*, *3*, 130.

Gershman, S. & Daw, N. (2012). Perception, action and utility: the tangled skein. In M. Rabinovich, K. Friston, & P. Varona (eds.), *Principles of Brain Dynamics: Global State Interactions* (293–312). MIT Press.

Gilbert, S. & Sarkar, S. (2000). Embracing complexity: organicism for the 21st century. *Developmental Dynamics*, *219*, 1–9.

Glennan, S. (2017). *The New Mechanical Philosophy*. Oxford University Press.

Gray, C. (2009). Principle of least action. *Scholarpedia, 4*, 8291.

Gross, D. (1996). The role of symmetry in fundamental physics. *Proceedings of the National Academy of Sciences, 93*, 14256–14259.

Gurova, L. (2011). Principles versus mechanisms in cognitive science. In V. Karakostas & D. Dieks (eds.), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science, 2*, 393–403.

Hobson, J. & Friston, K. (2016). A response to our theatre critics. *Journal of Consciousness Studies, 23*, 245–254.

Hohwy, J. (2016). The self-evidencing brain. *Noûs, 50*, 259–285.

Hohwy, J. (2014). The neural organ explains the mind. In T. Metzinger & J. Windt (eds.), *Open Mind*. Open MIND Group.

Kaplan, D. M. & Craver, C. (2011). The explanatory force of dynamical and mathematical models in neuroscience: a mechanistic perspective. *Philosophy of Science, 78*, 601–627.

Kauffman, S. (2013). Evolution beyond Newton, Darwin, and entailing law. In C. Lineweaver, P. Davies, & M. Ruse (eds.), *Complexity and the Arrow of Time* (162–190). Cambridge University Press.

Kirchhoff, M., Parr, T., Palacios, E., Friston, K., & Kiverstein, J. (2018). The Markov blankets of life: autonomy, active inference and the free energy principle. *Journal of The Royal Society Interface*, 15(138), 20170792.

Kirchhoff, M. & Froese, T. (2017). Where there is life there is mind: in support of a strong life-mind continuity thesis. *Entropy, 19*, 169.

Klein, C. (2018). What do predictive coders want? *Synthese, 195*, 2541–2557.

Knill, D. & Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences, 27*, 712–719.

Longo, G. & Montévil, M. (2013). *Perspectives on organisms: biological time, symmetries and*

singularities. Springer.

Longo, G., Montévil, M., & Kauffman, S. (2012). No entailing laws, but enablement in the evolution of the biosphere. *Proceedings of the 14th International Conference on Genetic and Evolutionary Computation Conference Companion* (1379–1392).

Marblestone, A., Wayne, G., & Kording, K. (2016). Towards an integration of deep learning and neuroscience. *Frontiers in Computational Neuroscience*, *10*, 1–41.

Maturana, H. (1975). The organization of the living: a theory of the living organization. *International Journal of Man-Machine Studies*, *7*, 313–332.

Mitchell, S. (2002). Integrative pluralism. *Biology and Philosophy*, *17*, 55–70.

Montévil, M., Mossio, M., Pocheville, A., & Longo, G. (2016). Theoretical principles for biology: variation. *Progress in Biophysics and Molecular Biology*, *122*, 36–50.

Moreno A. & Mossio, M. (2015). *Biological Autonomy: A Philosophical and Theoretical Enquiry*. Springer.

Morowitz, H. & Smith, E. (2007). Energy flow and the organization of life. *Complexity*, *13*, 51–59.

Mossio, M., Montévil, M., & Longo, G. (2016). Theoretical principles for biology: organization. *Progress in Biophysics and Molecular Biology*, *122*, 24–35.

Oliveira, H., & Melo, L. V. (2015). Huygens synchronization of two clocks. *Scientific Reports*, *5*, 11548.

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

Potochnik, A. (2017). *Idealization and the Aims of Science*. University of Chicago Press.

Ramstead, M. J. D., Badcock, P., & Friston, K. (2017). Answering Schrödinger's question: a free-energy formulation. *Physics of Life Reviews*.

Rao, R. & Ballard, D. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*, 79–87.

Rice, C. (2015). Moving beyond causes: optimality models and scientific explanation. *Noûs*, *49*, 589–615.

Schrödinger, E. (1992). *What is Life?* Cambridge University Press.

Smith, E. & Morowitz, H. (2016). *The Origin and Nature of Life on Earth: The Emergence of The Fourth Geosphere*. Cambridge University Press.

Soto, A., Longo, G., Miquel, P., Montévil, M., …, & Sonnenschein, C. (2016). Toward a theory of organisms: three founding principles in search of a useful integration. *Progress in Biophysics and Molecular Biology*, *122*, 77–82.

Spirtes, P., Glymour, C. & Scheines, R. (2000). *Causation, Prediction, and Search*, 2nd edition. MIT Press.

Uffink, J. (2001). Bluff your way in the second law of thermodynamics. *Studies in History and Philosophy of Modern Physics*, *32*, 305–394.

van Rooij, I., Wright, C., Kwisthout, J., & Wareham, T. (2018). Rational analysis, intractability, and the prospects of 'as if'-explanations. *Synthese*, *195*, 491–510.

Varela, F., Maturana, H., Uribe, R., (1974). Autopoiesis: the organization of living systems, its characterization and a model. *Biosystems*, *5*, 187–196.

Wiese, W. & Metzinger T. (2017). Vanilla PP for philosophers: a primer on predictive processing. In T. Metzinger & W. Wiese (eds.). *Philosophy and Predictive Processing*. Open MIND Group.