The Markov Blanket Trick: On the Scope of the Free Energy Principle and Active Inference

Vicente Raja,[1] Dinesh Valluri,[2] Edward Baggs,[1] Anthony Chemero,[3,4] and Michael L. Anderson[1,5,6]

[1]Rotman Institute of Philosophy, Western University (Canada)
[2]Department of Computer Science, Western University (Canada)
[3]Department of Philosophy, University of Cincinnati (USA)
[4]Department of Psychology, University of Cincinnati (USA)
[5]Department of Philosophy, Western University (Canada)
[6]Brain and Mind Institute, Western University (Canada)

Abstract

The free energy principle (FEP) has been presented as a unified brain theory, as a general principle for the self-organization of biological systems, and most recently as a principle for a theory of every *thing*. Additionally, active inference has been proposed as the process theory entailed by FEP that is able to model the full range of biological and cognitive events. In this paper, we challenge these two claims. We argue that FEP is not the general principle it is claimed to be, and that active inference is not the all-encompassing process theory it is purported to be either. The core aspects of our argumentation are that (i) FEP is just a way to generalize Bayesian inference to all domains by the use of a Markov blanket formalism, a generalization we call the Markov blanket trick; and that (ii) active inference presupposes successful perception and action instead of explaining them.

1. Introduction: Getting to the Free Energy Principle

In the early 1990s, machine learning researchers developed a formalism analogous to Helmholtz free energy and began using it as an objective function for artificial neural networks in the context of parameter optimization through expectation-maximization algorithms (Dayan et al. 1995; Hinton & van Camp 1993; Hinton & Zemel 1993; Neal & Hinton 1998; for early similar formulations, see

Csiszár & Tusnády 1983; Hathaway 1986; Hinton & Sejnowski 1983). The formalism was based on a quantity known as *variational free energy* with roots in the variational methods first developed by Richard Feynman (1972) and allowed researchers to solve otherwise intractable inferences when trying to guess the true distribution of partial sample data. In short, minimizing variational free energy is equivalent to optimizing the evidential lower bound (ELBO) in Bayesian inference (Fox & Roberts 2012). Using the minimization of variational free energy as the objective function for a neural network ensures that, after the training process, the network will instantiate a close-to-optimal model of the true distribution of the partial sample data—i.e., its parameters will define a distribution of data very similar to the real distribution.

In the context of machine learning, algorithms based on the minimization of variational free energy and/or the equivalent ELBO maximization were used to optimize systems able to discover the latent/hidden variables (i.e., the true distribution) that generated the observed variables (i.e., the partial sample data)—e.g., what variational autoencoders do nowadays (Kingma & Welling 2014, 2019). An interesting question that follows is whether this method for discovering hidden variables from observable variables is useful as a model of perception (Dayan et al. 1995). This would be consistent with the common understanding of perception as a kind of inference to discover the environment (i.e., hidden variables) out of sensations (i.e., observable variables).[1] Around the turn of the century this idea permeated existing Bayesian brain hypotheses in the form of a set of proposals that described brain activity as a matter of predicting sensory inputs under a model of the relationship between hidden, observable, and internal variables. The general term for this set of proposals is *predictive processing* (e.g., Friston 2002; Clark 2015; Hohwy 2013; Rao & Ballard 1999).

[1] This idea is as old as the sciences of the mind, so this move was not novel in any meaningful way. It was just a way of applying new tools to an old framework. It was even called "Helmholtz machine" (Dayan et al. 1995).

Karl Friston (2002, 2003, 2005) pioneered a specific form of the general predictive processing framework explicitly based on the minimization of variational free energy. Later, Friston presented this version of the framework as a general principle for the functioning of the brain. Not only perception but all types of activity carried out by brains (e.g., learning, action, etc.) were cast in terms of the minimization of variational free energy (Friston et al. 2006; Friston 2010; Friston et al. 2017). The *free energy principle* (hereafter simply FEP) was born. In the following years, FEP was expanded beyond brains and was proposed, first, as a general principle for the self-organization of biological systems (Friston 2013; Ramstead et al. 2018) and, most recently, as a general principle for a theory of every *thing* (Friston 2019). The most recent formulation of FEP is:

> ***FEP*** = *Any ergodic random dynamical system with an attractor and a Markov blanket behaves as if it were minimizing the variational free energy of its particular states.*

In the case of physical systems, FEP speaks to the way that a system's dynamics unfold over time and relax towards an attractor. In the case of biological systems, it speaks to their self-organization to maintain homeostasis. And in the case of cognitive systems, FEP speaks to their epistemological contact with their environment in terms of accurate perception and proper actions. In all these cases, FEP entails a specific Bayesian process of minimization of free energy labelled as *active inference.*

In this paper, we will defend two simple theses. First, that FEP lacks the resources to be the principle it is claimed to be and should be understood as one modeling framework among many others. And second, that active inference fails to account for perception and action; rather, it presupposes them. To defend these two theses, we unpack the general definition of FEP in section 2 and analyze some of its shortcomings in Section 3. Then, we illustrate active inference with a simple perception-action model and address its main problems in Section 4.

A note on some things that we will include and some things that we will leave out of the paper. First, we will not shy away from mathematics when needed. The FEP is at heart a mathematical formulation, and in general it is important to engage with the mathematical details. However, we also want to be as general and clear as possible. We thus confine the bulk of the mathematical formalism to section 2, in which we draw heavily on mathematical formulas presented in recent work (e.g., Buckley et al. 2017; Friston 2019; Friston et al. 2020; Friston et al. 2021; Ramstead et al. 2020a).[2]

Secondly, we will avoid definitional discussions of terms such as surprise/surprisal, inference, belief, or model in the literature on FEP. We acknowledge that the use of these terms has been criticized on various grounds (e.g., Anderson & Chemero 2019; van Es 2020). We will not problematize them in this paper unless it is strictly necessary for our argument.

Thirdly, we will not be discussing empirical evidence in support of FEP. Direct experimental support for predictive processing in general, and for FEP and active inference in particular is, at best, scarce (e.g., Walsh et al. 2020). To date, the best existing empirical evidence for FEP amounts to no more than a set of simulations that bear, perhaps, metaphorical resemblance to actual biological and cognitive processes—see, e.g., Friston et al. (2017, p. 36). Given the growing interest in FEP both within the life sciences and in philosophy, this lack of empirical evidence is an important gap in the existing literature.

Finally, we are not arguing against FEP and active inference as research programs. To do so would be to provide a Hegelian argument (Chemero 2009), named after Hegel's infamous argument that, as a matter of logic, there could only be seven planets in the solar system. Hegelian arguments are attempts to shut down empirical research using conceptual means alone. We are not offering a Hegelian argument here, and we are fully in favor of continuing research on active inference and FEP, especially, given the considerations just mentioned, empirical research. What follows can be read as a

[2] In many cases we will refer to pre-prints as an important part of the current discussions on the FEP is in them.

series of suggestions for issues that proponents of FEP and active inference will need to pay further attention to, especially if they wish to maintain the more ambitious claims they have made.

2. The Free-Energy Principle

FEP has recently been proposed as the foundation for a theory of every *thing* (Friston 2019). In a nutshell, the proponents of FEP claim that, once we have committed to self-evident notions of what it is to be a thing and what it means for a thing to be different from its environment, minimizing free energy becomes a physical and biological imperative for every such *thing*. Note that we use *thing* (in italics) to refer to things as described in FEP. The scope of FEP as a principle depends on the supposedly self-evident nature of such a notion of *thing* and, more concretely, on the extent to which physical, biological, and cognitive systems can be characterized in that way. The latter will be our criterion for evaluating FEP.[3]

Under FEP, a *thing* is a system that evolves in time towards a set of relatively stable states. In order to be a *thing*, such a temporal evolution is maintained far from thermodynamic equilibrium during a nontrivial period of time—otherwise the *thing* will simply disintegrate into its environment. The *thing* is therefore a measurable system—i.e., it has some stable states—and can be formalized in terms of its random dynamics using a Langevin formulation:

$$\dot{x} \ = f(x) + \omega \tag{1}$$

In this formulation, the temporal evolution of the random dynamical system $\dot{x}$ is described in terms of a deterministic function of the states x and some stochastic fluctuations ω. If the system evolves towards a stable set of states, it is said to have a *pullback attractor*; namely, the dynamics of the system

[3] This section offers a succinct presentation of the formalism of FEP. The reader already familiar with it may prefer to skip ahead to section 3. Similarly, readers of a less mathematical persuasion may as well prefer to skip ahead to section 3 where an understanding of the formalism is helpful—and indeed highly recommended—but is not a requirement.

are 'attracted' to some set of stable states and attain what is known as a *nonequilibrium steady state* (NESS). The existence of the NESS is what allows us to identify the system as a *thing*—i.e., a persisting process that does not disintegrate towards equilibrium. In the case of biological systems, the NESS is taken to be the *phenotype* (Ramstead et al. 2018).

The Langevin dynamics formulation is fairly general. Under ergodic assumptions, some statistical properties can be associated to the dynamics it describes.[4] Concretely, the states of the system can be associated with a probability density, $p(x)$, whose dynamics can be described in terms of the Fokker-Plank equation:

$$\dot{p}(x) = \nabla \cdot (\Gamma\nabla - f)p(x) \tag{2}$$

When the system has a NESS, we can solve the Fokker-Plank equation as:

$$\dot{p}(x) = 0 \Leftrightarrow f(x) = (Q - \Gamma) \cdot \nabla\Im(x) \tag{3}$$

$$\Im(x) = -lnp(x)$$

Equations (2) and (3) provide a statistical description of the dynamics of a random dynamical system with a NESS in terms of two orthogonal gradient flows on surprisal, $\Im(x)$—also known as self-information or Shannon information—via the Helmholtz decomposition: one sinusoidal flow, $Q\nabla\Im(x)$, that contours around surprisal and arises from the Q-matrix, and one curl-free flow, $-\Gamma\nabla\Im(x)$, that depends upon random fluctuations Γ and performs a descent on surprisal to effectively counter those fluctuations. Thus, equations (2) and (3) describe a probability density, $p(x)$, in which every state x has a probability such that the states in which the system typically is (i.e., the states of the attractor)

[4] Langevin dynamics, however, are not adequate to describe discrete dynamical systems. In this sense, it is not clear that they are general enough as to support a theory of every *thing* (Friston 2019). Additionally, the assumption of ergodicity—i.e., that all the states of the system are equiprobable in the long term—is a very strong assumption and the fact FEP needs it may be concerning.

have a higher probability than atypical states. If we measure this typicality with $\mathfrak{I}(x)$, we notice that typical states are less surprising than atypical states. Therefore, random dynamical systems with a NESS counter random fluctuations that would make them disintegrate into their environment (i.e., the system would stop being a *thing*) in a way that can be expressed in terms of a gradient descent on (i.e., a reduction) of surprisal. As the average of surprisal in time is *entropy*, this formulation of a *thing* simply states that systems maintain their state far from equilibrium by reducing their entropy (see Figure 1).

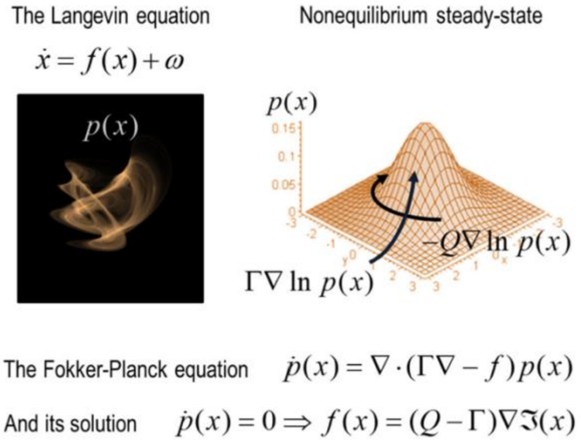

The Langevin equation
$\dot{x} = f(x) + \omega$
$p(x)$
Nonequilibrium steady-state
$p(x)$
0.15
0.1
0.05
0
$-Q\nabla \ln p(x)$
$\Gamma\nabla \ln p(x)$
The Fokker-Planck equation $\dot{p}(x) = \nabla \cdot (\Gamma\nabla - f)p(x)$
And its solution $\dot{p}(x) = 0 \Rightarrow f(x) = (Q - \Gamma)\nabla\mathfrak{I}(x)$

Figure 1. Two ways to represent a random dynamical system with a NESS. The left panel represents its trajectory through the phase space. You can associate a probability density, $p(x)$, with that trajectory. The right panel represents the shape of the dynamics on $p(x)$ in terms of two flows, the sinusoidal one depending on Q and the curl-free one depending on Γ. These two representations are connected by the Fokker-Plank equation associated with the Langevin formalization of the dynamics of x. (Figure from Friston et al. 2020, p. 6; used under Creative Common Attribution License CC BY.)

This is the FEP formalization of what it is to be a *thing*; it is expressed in terms of random dynamical systems that have attained a NESS using statistical properties and, more concretely, by appealing to surprisal, $\mathfrak{I}(x)$. This formalization is compatible with the idea that things counter entropic forces and, in the concrete case of biological things, with the idea that they maintain homeostasis. At this point, proponents of FEP introduce their second commitment regarding *things*: a *thing* must be distinguishable from its environment. Once we select a *thing* of interest, we must be able to distinguish its internal states from external states (i.e., its environment). Such a distinction is understood in terms of statistical independence between internal and external states relative to a third set of states known as the blanket states. This is the point at which *Markov blankets* get into the formalism.

Markov blankets were first proposed by Judea Pearl (1988) in the context of graphical models. These models express the statistical dependencies (edges) between different factors or states (nodes). Given a state of the model, its Markov blanket is the set of sufficient states of the network needed to predict that state. In the concrete case of Bayesian networks (the ones used in FEP), the Markov blanket of a state *S* is comprised by the states that directly influence *S*, the states directly influenced by *S* (a.k.a., children), and the other states that directly influence *S*'s children. Furthermore, if we take the state *S* to be the internal state $\boldsymbol{\mu}$ of a system, we can classify the states $\boldsymbol{b}$ of the Markov blanket as sensory states $\boldsymbol{s}$, which are the ones that are not influenced by $\boldsymbol{\mu}$, and active states $\boldsymbol{a}$, which are the ones that are not influenced for the rest of external states in the network, $\boldsymbol{\eta}$ (see Figure 2).

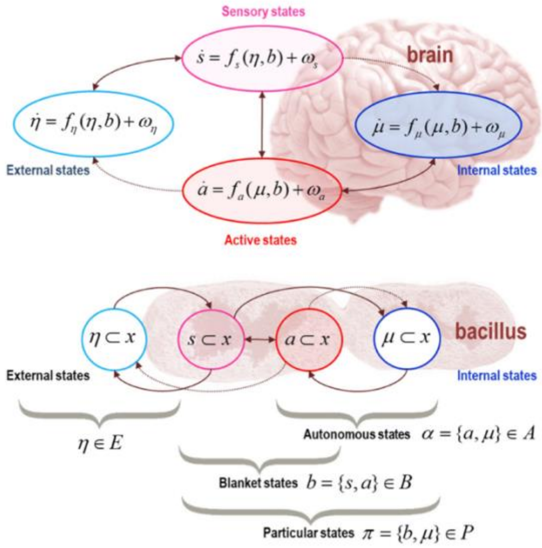

Sensory states
$\dot{s} = f_s(\eta, b) + \omega_s$
brain
$\dot{\eta} = f_\eta(\eta, b) + \omega_\eta$
$\dot{\mu} = f_\mu(\mu, b) + \omega_\mu$
External states
Internal states
$\dot{a} = f_a(\mu, b) + \omega_a$
Active states
$\eta \subset x$
$s \subset x$
$a \subset x$
$\mu \subset x$
bacillus
External 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 2. Example of a partition between internal and external states, formalized in terms of Makov blankets, in two random dynamical systems with a NESS: brain (top) and bacillus (bottom). All the different states and their combinations are explicit in the figure. More details (including the mathematical formulation) in the main text. (Figure from Friston et al. 2020, p. 6; used under Creative Common Attribution License CC BY.)

Given the presence of a Markov blanket, the statistical independence between internal and external states of a graphical model can be formalized as:

$$\eta \perp \mu \;\Leftrightarrow\; p(\eta, \mu \,|\, b) = p(\eta \,|\, b)p(\mu \,|\, b) \tag{4}$$

$$b = \{s, a\}$$

$$\pi = \{b, \mu\}$$

Equation (4) expresses the joint probability between internal and external states mediated by blanket states, $p(\boldsymbol{\eta},\boldsymbol{\mu}\,|\,\boldsymbol{b})$. Additionally, equation (4) defines blanket states, $\boldsymbol{b}$, as the set of sensory and active states, and defines *particular states*, $\boldsymbol{\pi}$, as the set of internal and blanket states.[5] Under the Markov blanket formalism, the notion of *state* is pretty general and, although it doesn't need to be directly mapped to structural properties of a system, when applied to biological and cognitive systems, states are usually mapped to those properties. In cells, for instance, internal states are the states within a cell, blanket states are states of the cellular membrane, and external states are those environmental states beyond the cellular membrane. Similarly, in brains, the states of the brain are internal states, blanket states are divided between the states of sensory receptors (sensory states) and the states of motor receptors (active states), and the states of the surrounding environment are external states.

As the FEP story goes, once we understand the dynamics of systems in terms of their statistical properties and, more concretely, in terms of gradients over $\Im(x)$, and we separate them from the external dynamics with a Markov blanket, we find a very interesting observation: the dynamics of the particular states of the system $\boldsymbol{\pi}$ can be seen as descending a gradient over $\Im(x)$ defined in terms of external states $\boldsymbol{\eta}$. Namely, particular states behave *as if* they were minimizing the surprisal with regard

[5] This version of the Markov blanket formalism is the most recent one (e.g., Friston 2019; Friston et al. 2020; Parr et al. 2020). Previous versions of the Markov blanket formalism under FEP have been criticized as mathematically inaccurate (Biehl et al. 2021; see also Friston et al. 2020).

to external states. This is possible because, when a Markov blanket is applied to a system as the one described in equations (1) to (3), the dynamics $\dot{x}$ can be formalized as (see also figure 2 above):

$$\dot{x} = [\dot{\eta} = f_\eta(\eta, b) + \omega_\eta;\ \dot{s} = f_s(\eta, b) + \omega_s; \dot{\mu} = f_\mu(\mu, b) + \omega_\mu; \dot{a} = f_a(\mu, b) + \omega_a] \quad (5)$$

$b = \{s, a\}$ (blanket states)

$\alpha = \{\mu, a\}$ (autonomous states)

$\pi = \{s, \mu, a\}$ (particular states)[6]

Notice that equation (5) takes the dynamics $\dot{x}$ described in equation (1) and partitions it in four different flows with the same Langevin formulation: external dynamics ($\dot{\eta}$), sensory dynamics ($\dot{s}$), internal dynamics ($\dot{\mu}$), and action dynamics ($\dot{a}$). All these dynamics are respectively the product of a deterministic function *f_i* and random noise ω_i. Additionally, equation (5) re-states the notions of blanket states (***b***) and particular states (***π***), and introduces the notion of autonomous states (**α**), which are the states that are not directly influenced by external states.

The Markov blanket partition of random dynamical systems that have attained a NESS provides a way to express the dynamics of those systems in terms of orthogonal gradients over surprisal $\Im$, as specified in equations (2) and (3). Given this, it can be shown that:

$$\begin{aligned} f_\eta(\eta, b) &= (Q_{\eta\eta} - \Gamma_{\eta\eta})\nabla_\eta \Im(\eta, b) \\ f_s(\eta, b) &= (Q_{ss} - \Gamma_{ss})\nabla_s \Im(\eta, b) + Q_{sa}\nabla_a \Im(\eta, b) \\ f_\mu(\mu, b) &= \left(Q_{\mu\mu} - \Gamma_{\mu\mu}\right)\nabla_\mu \Im(\mu, b) \\ f_a(\mu, b) &= (Q_{aa} - \Gamma_{aa})\nabla_a \Im(\mu, b) + Q_{as}\nabla_s \Im(\mu, b) \end{aligned} \quad (6)$$

Equation (6) shows the solution of the Fokker-Plank equation applied to all partitions of the dynamics (***η***, ***s***, ***μ***, ***a***). The dynamics of the system can still be understood in terms of particular gradients over

[6] Notice that particular states encompass autonomous states, so $\pi = \{s, \alpha\}$ is an equivalent way to define particular states.

surprisal that depend on the statistical independences entailed by the blanket. In the case of autonomous states ($\boldsymbol{\alpha}$), the last two lines of equation (6) can be summarized as:

$$f_{\alpha}(\pi) = (Q_{\alpha\alpha} - \Gamma_{\alpha\alpha})\nabla_{\alpha}\Im(\pi) \tag{7}$$
$$\Im(\pi) = \Im(\mu, b)$$

This can be read as expressing that the flow of dynamics of the states that are under control of the system (i.e., internal and action states) evolve on a gradient of particular surprisal: the surprisal of particular states, $\Im(\boldsymbol{\pi})$. Notice that equations (6) and (7) are just the mathematical consequence of combining Langevin dynamics with the statistical formulation of the Fokker-Plank equation and the Markov blanket formalism.

At this point, we zoom out a little bit for a better sense of where we are. So far, we have formalized a simple but central observation of FEP: when systems are characterized in terms of being measurable, their dynamics can be understood as gradients over surprisal in the sense that, as those systems exist, they can be described as tending towards low-surprise (i.e., high probability) states and, therefore, as minimizing $\Im(x)$. Under the Markov blanket formalism, $\Im(x)$ is understood as $\Im(\boldsymbol{\eta}, \boldsymbol{b})$ and $\Im(\boldsymbol{\pi})$ depending on the specific dynamics within the partition. With this, we now have all the concepts we need to understand the FEP definition provided above:

> ***FEP*** = *Any ergodic random dynamical system with an attractor and a Markov blanket behaves as if it were minimizing the variational free energy of its particular states.*

We have provided the Langevin characterization of the random dynamical system with NESS and combined it with the Markov blanket formalism. We also know what 'particular states' means, and we have described a process of minimization (of surprisal). This process of minimization is not of variational free energy, but as variational free energy is defined as an upper bound on surprisal, the minimization of variational free energy entails the minimization of surprisal.

In making the relationship between variational free energy and surprisal explicit, we will show why FEP is becoming relevant in the biological and cognitive sciences. When systems are described in terms of FEP, their internal states can be described as having Bayesian beliefs (i.e., instantiating probability densities) about external states given the blanket states. This fact opens a possibility to understand the way systems (e.g., brains) epistemically relate to external states (e.g., environments), even when they are conditionally independent given other states (e.g., sensory states), in terms of a (Bayesian) inferential process.

Due to the partition formalized in equation (5), it is straightforward to see that for every blanket state $\boldsymbol{b}$ there is one associated internal state $\boldsymbol{\mu}(\boldsymbol{b})$ and one associated external state $\boldsymbol{\eta}(\boldsymbol{b})$.[7] This leads to a possible mapping $\boldsymbol{\sigma}$ between internal states, external states, and their dynamics:

$$\eta(b) = \sigma\big(\mu(b)\big) \qquad\qquad \dot{\eta}(b) = \nabla_{\mu}\sigma\dot{\mu}(b) \tag{8}$$

Under Laplacian assumptions, the relationships of equation (8) can be formalized in terms of a family of densities over external states parametrized by internal states, $q_{\mu}(\boldsymbol{\eta})$, such that:

$$p(\eta|b) = p(\eta|\pi) \approx q_{\mu}(\eta) = N(\sigma(\mu), \textstyle\sum(\mu)^{-1}) \tag{9}$$

$$\textstyle\sum(\mu)^{-1} \triangleq \nabla_{\sigma\sigma}\mathfrak{J}(\sigma(\mu), b)$$

$$\mathfrak{J}(\sigma(\mu), b) = \mathfrak{J}(\mu, b) = \mathfrak{J}(\pi)$$

That is, that the probability of a particular state to be associated to an external state, given the blanket, $p(\boldsymbol{\eta}|\boldsymbol{\pi})$, can be described in terms of the density dynamics of internal states, $q_{\mu}(\boldsymbol{\eta})$, as each of them is associated with a probability density parametrized by their mapping to external states, $\boldsymbol{\sigma}(\boldsymbol{\mu})$, and their gradient on surprisal, $(\boldsymbol{\mu})^{-1}$. As the dynamics of the system of interest descend through the gradient of $\mathfrak{J}(\boldsymbol{\pi})$, $q_{\mu}(\boldsymbol{\eta})$ becomes a better model of $p(\boldsymbol{\eta}|\boldsymbol{\pi})$. Notice that, in terms of (approximate) Bayesian

[7] Assuming the relationship between $\boldsymbol{b}$ and $\boldsymbol{\mu}(\boldsymbol{b})$ is injective (Parr et al. 2019).

inference, p($\boldsymbol{\eta}$| $\boldsymbol{\pi}$) can be understood as the *true posterior* and q$_{\mu}$($\boldsymbol{\eta}$) can be understood as a density of Bayesian beliefs also known as *recognition density*. This interpretation will become relevant soon.[8]

The relationship in the first line of equation (9) is granted by the dual information geometry ensued by the Markov blanket partition. Such partition allows for the density dynamics of internal states to be described in two complementary ways: as changing—*g* is a measure of that change—with respect to time (intrinsic) and as changing in terms of the parametrization of external states (extrinsic):

$$g(\tau) = \nabla_{\tau'\tau'} D[p_{\tau'}(\mu)||p_{\tau}(\mu)]|_{\tau'=\tau} \qquad \text{(intrinsic)} \qquad (10)$$

$$g(\mu) = \nabla_{\mu'\mu'} D[q_{\mu'}(\eta)||q_{\mu}(\eta)]|_{\mu'=\mu} \qquad \text{(extrinsic)}$$

The relationships made explicit in equations (8), (9), and (10), allow us to redescribe the dynamics of external and internal states as performing a descent of the same surprisal of particular states, $\mathfrak{I}$($\boldsymbol{\pi}$), when mediated by blanket states:

$$\dot{\eta}(b) = (Q_{\eta\eta} - \Gamma_{\eta\eta})(\nabla_{\mu}\sigma)^{-}\nabla_{\mu} I(\pi) \qquad (11)$$

$$\dot{\mu}(b) = -\Gamma_{\sigma\sigma}\nabla_{\mu}\mathfrak{I}(\pi)$$

$$\Gamma_{\sigma\sigma} \triangleq (\nabla_{\mu}\sigma)^{-}\Gamma_{\eta\eta}(\nabla_{\mu}\sigma)^{-}$$

Notice that the second line of equation (11) is analogous to equation (7) as it relates one set of autonomous states (internal states in this case) to a gradient on $\mathfrak{I}$($\boldsymbol{\pi}$). However, the fluctuations $\Gamma_{\sigma\sigma}$ of that gradient are defined in terms of external states and, as the descent over is $\mathfrak{I}$($\boldsymbol{\pi}$) exactly countering those fluctuations, this descent depends on external states.

[8] Given the equivalence p($\boldsymbol{\eta}$| $\boldsymbol{b}$) = p($\boldsymbol{\eta}$| $\boldsymbol{\pi}$) in equation (9), there are two compatible ways to proceed in describing Bayesian inference from this point. One way is to define *variational free energy* in terms of particular states ($\boldsymbol{\pi}$) as Friston et al. (2020, p. 9) do—see also Friston (2019, p. 84 & ff.). The other one is to define *variational free energy* in terms of blanket states ($\boldsymbol{b}$) as Ramstead et al. (2020a, p. 22) do. The relationship between these two options is somewhat expressed in Parr et al. (2019, p. 7). We will describe *variational free energy* in terms of particular states ($\boldsymbol{\pi}$) to maintain the consistency of our notation.

The final step towards understanding FEP is the relationship between $\Im(\pi)$ and *variational free energy*. We have seen that as *things* counter entropic tendencies and remain in a non-equilibrium state they are minimizing the surprisal of their internal and blanket states with respect to external states. In this sense, surprisal $\Im(\pi)$ is just an index of the dynamics of systems as they maintain some set of stable states. In biological systems, for instance, $\Im(\pi)$ indexes homeostasis. Crucially, the minimization of surprisal can also be understood in terms of Bayesian inference. Following equation (9), the internal states of the system can be interpreted as instantiating a density, $q_{\mu}(\eta)$, that parametrizes Bayesian beliefs over external states given particular states, $p(\eta|\pi)$. How $q_{\mu}(\eta)$ approximates $p(\eta|\pi)$ through minimizing $\Im(\pi)$ can be straightforwardly understood as the way a *recognition density* ($q_{\mu}(\eta)$) approximates a *true posterior* ($p(\eta|\pi)$) given a *generative model* ($p(\eta, \pi)$).[9] The way $q_{\mu}(\eta)$ approximates $p(\eta|\pi)$ can be expressed in terms of the Kullback-Liebler Divergence, D_{KL}, between the two probability densities. D_{KL} is a measure of the divergence between two probability densities that is always positive and where a zero value means the two densities are the same. Given this, it is possible cast the relationship between $q_{\mu}(\eta)$ and $p(\eta|\pi)$ as:

$$D_{KL}(q_{\mu}(\eta)||p(\eta|\pi)) = \int d\eta\, q_{\mu}(\eta) ln\, \frac{q_{\mu}(\eta)}{p(\eta|\pi)} \qquad (12)$$

$$= \int d\eta\, q_{\mu}(\eta)\big[ln\, q_{\mu}(\eta) - ln\, p(\eta|\pi)\,\big]$$

Equation (12) makes use of the posterior $p(\eta|\pi)$ to calculate D_{KL}. However, given the Markov blanket, the internal states of the system have no access to the external states and, therefore, have no access to the true posterior—e.g., brains have access to sensory data but not to its environmental causes. A way to overcome this issue is to define the true posterior in terms of a *generative model* $p(\eta, \pi)$ defined as the

[9] In machine learning, variational methods are used in cases in which the true posterior is unknown and, therefore, *exact Bayesian inference* is most likely intractable. In these cases, we can try to approximate the true posterior given a generative model—i.e., we can use variational methods to perform *approximate Bayesian inference* (e.g., Kingma & Welling 2019).

joint probability between external and particular states. With this in mind, we can describe the true posterior as:

$$p(\eta|\pi) = \frac{p(\eta,\pi)}{p(\pi)} \tag{13}$$

Given equation (13) and including the requirement that the total probability adds to 1, equation (12) can be re-written as:

$$D_{KL}(q_{\mu}(\eta)||p(\eta|\pi)) = \int d\eta\, q_{\mu}(\eta) ln\, \frac{q_{\mu}(\eta)}{p(\eta,\pi)} + ln\, p(\pi) \tag{14}$$

Equation (14) does not depend on the true posterior anymore, but just on the recognition density, the generative model, and a surprisal term. At this point, we define *variational free energy*, ***F***, as:

$$F \equiv \int d\eta\, [q_{\mu}(\eta) ln\, \frac{q_{\mu}(\eta)}{p(\eta,\pi)}] \qquad (15)^{10}$$

And then we have:

$$D_{KL}(q_{\mu}(\eta)||p(\eta|\pi)) = F + ln\, p(\pi) \tag{16}$$

Including the definition of surprisal ($\Im$) to reorganize equation (16) and given that D_{KL} must be always positive, it is easy to see that variational free energy, ***F***, is an upper bound on $\Im(\boldsymbol{\pi})$:

$$\Im(x) = -ln\, p(x) \;\; \Rightarrow \;\; \Im(\pi) = -ln\, p(\pi) \qquad (17)^{11}$$

$$F = D_{KL}(q_{\mu}(\eta)||p(\eta|\pi)) + \Im(\pi)$$

$$D_{KL}(q_{\mu}(\eta)||p(\eta|\pi)) \;\geq\; 0$$

$$\therefore F \geq \; \Im(\pi)$$

[10] Re-writing equation (15) as F = $\int d\boldsymbol{\eta}\, q_{\mu}(\boldsymbol{\eta})\, ln q_{\mu}(\eta) - \int d\boldsymbol{\eta}\, q_{\mu}(\boldsymbol{\eta})\, ln p(\boldsymbol{\eta},\boldsymbol{\pi})$ and defining an *energy* term $E(\boldsymbol{\eta},\boldsymbol{\pi}) = -ln p(\boldsymbol{\eta},\boldsymbol{\pi})$, we have $F = \int d\boldsymbol{\eta}\, q_{\mu}(\boldsymbol{\eta})\, E(\boldsymbol{\eta},\boldsymbol{\pi}) + \int d\boldsymbol{\eta}\, q_{\mu}(\boldsymbol{\eta})\, ln q_{\mu}(\boldsymbol{\eta})$. Here ***F*** is formulated as depending on an energy-like function and something similar to entropy, which is analogous to *Helmholtz free energy*, F = U - TS, where *U* is internal energy, *S* is entropy, and *T* is temperature. This analogy illustrates why ***F*** is understood as (variational) free energy.

[11] The second line of equation (17) is analogous to *evidence lower bound* (ELBO), $\boldsymbol{L}$ = lnP(x) - D_{KL}(Q||P). ELBO is also known as *negative variational free energy*.

The result of equation (17) entails that any gradient over $\Im(\boldsymbol{\pi})$ can be expressed in terms of a gradient over $\boldsymbol{F}$ given under a generative model. Alternatively, it can be said that any systems that minimized $\boldsymbol{F}$ under a generative model is implicitly minimizing $\Im(\boldsymbol{\pi})$. Given this, we can re-write (and expand) the dynamics of *autonomous states* captured by equation (7) above in terms of variational free energy:

$$
\begin{aligned}
f_\alpha(\pi) &\approx (Q_{\alpha\alpha} - \mathit{\Gamma}_{\alpha\alpha})\nabla_\alpha F \qquad (18)\\
f_\mu(\pi) &\approx (Q_{\mu\mu} - \mathit{\Gamma}_{\mu\mu})\nabla_\mu F\\
f_a(\pi) &\approx (Q_{aa} - \mathit{\Gamma}_{aa})\nabla_{\alpha a} F
\end{aligned}
$$

Equation (18) expresses FEP as it has been formulated above: any random dynamical system that has attained a NESS and has a Markov blanket behaves as if it were minimizing the variational free energy ($\boldsymbol{F}$) of its particular states ($\boldsymbol{\pi}$). This minimization is understood in terms of approximate Bayesian inference as shown in equations (12) to (17). Moreover, equation (18) shows that such approximate Bayesian inference can be accomplished in two different ways: by changing internal states ($\boldsymbol{\mu}$) and by changing active states ($\boldsymbol{a}$). This is the result that harnesses *active inference* (see Section 4).

In summary, we have shown that FEP rests on three fundamental commitments. First, that every *thing* has a set of preferred states that makes it viable (NESS). Second, that for a *thing* to be different from its environment means to have a set of states ($\boldsymbol{b}$) that makes the internal states of the system ($\boldsymbol{\mu}$) statistically independent of external states ($\boldsymbol{\eta}$). This means the *thing* has a Markov blanket. Combining the first two commitments, it can be shown that the dynamics of every *thing* can be understood in terms of the minimization of the surprisal of its particular states, $\Im(\boldsymbol{\pi})$, with respect to its environment. Thus the process of minimization of surprisal is equivalent to a process of approximate Bayesian inference where $\Im(\boldsymbol{\pi})$ is implicitly minimized by the minimization of variational free energy $\boldsymbol{F}$ under a generative model provided by the NESS of the *thing*.

Given these commitments, we think FEP is true; that is, when *things* are defined as they are under the FEP, they can be described as minimizing surprisal. And given the *right* generative model, the minimization of surprisal can be understood in terms of minimization of variational free energy. However, we will show that there are problems with the commitments of FEP. We will show the notion of Markov blanket is not well defined under the FEP and that, even if a coherent notion of Markov blanket may be developed, it greatly limits the scope of the principle. Additionally, we will show that the notion of generative model in *active inference* impedes the framework from delivering the results it promises. The next two sections are respectively devoted to developing these two ideas.

3. The Markov Blanket Trick: On the Origin and Use of Markov Blankets

FEP might be considered as a principle for a theory of every *thing* either because it applies to all physical, biological, and cognitive systems we know or because it provides the most fundamental account of physical, biological, and/or cognitive systems. And FEP is claimed to meet (at least one of) these two conditions. Self-evident answers to the questions 'What is it to be a thing?' and 'What does it mean for a thing to be different from its environment?', so the story goes, directly lead to the notion of *thing* in FEP: a random dynamical system with a NESS and a Markov blanket. In this sense, any system (physical, biological, or cognitive) may be formalized in terms of FEP. Moreover, once every *thing* is described in these terms, it is possible to show that the fundamental laws and theorems of classical mechanics, statistical mechanics, quantum mechanics, and Bayesian mechanics follow from the FEP formalism (Friston 2019).

The *universal* and *fundamental* scope of FEP depends on being able to describe any system in terms of the formalisms entailed by the principle. The first formalism is Langevin dynamics in equation (1). This is a pretty general formalism and, although it might have some limitations (e.g., accounting for past dynamics and discrete dynamics), its scope is likely wide enough. The second one is the Markov blanket formalism described in equations (4), (5), and (6). Markov blankets are actually doing most of the heavy lifting in FEP: they provide the partition required for the dual information geometry on which FEP rests and also provide the proper structure for a variational Bayesian process to occur.[12]

Despite the centrality of Markov blankets in FEP, the reason they are selected to formalize the boundary of *things* is not provided in the literature. Usually, Markov blankets are presented in FEP as simply the obvious way to formalize the boundary of *things* of interest.[13] As far as we can tell, the reason to postulate Markov blankets is that proponents of FEP take the boundary between any *thing* and its environment to be a statistical one and Markov blankets are a good tool to capture such statistical relationships. However, a closer look at the concept of the Markov blanket shows that its applicability is not as wide and straightforward as the proponents of FEP seem to assume.

One problem with Markov blankets is purely formal. Biehl et al. (2021) have shown that some formalizations of Markov blankets in the FEP literature are mathematically inadequate given the assumptions of the framework (e.g., Friston 2013). This criticism led Friston et al. (2020) to provide further specifications for the Markov blanket formalism in terms of *sparsity constraints* that, ultimately, lead to an acknowledgment of some statistical independences between the states of the system that are not captured by the Markov blanket (e.g., between sensory and internal states; see Friston et al. 2020, p. 7). The problem pointed out by Biehl et al. (2021) and the solution proposed by Friston et al. (2020) cast doubt on the ability of Markov blankets to capture *all* possible statistical independences and, more importantly, likely reduces the general applicability of the formalism.

[12] Without the Markov blanket formalism, the dual information geometry presented in equation (10) would be reduced to its intrinsic component, which is the only one needed for standard thermodynamics (Friston 2019, Friston et al. 2020).

[13] Here's an example: "Clearly, one needs to differentiate between the system and its environment—those states that constitute or are intrinsic to the system and those that are not. To do this, we have to introduce a third set of states that separates internal from external states. This is known as a Markov blanket." (Ramstead et al. 2018, p. 3-4).

This formal issue with Markov blankets is interesting, but we won't focus on it. In the following, we provide a sketch of other issues with the use of Markov blankets in FEP. These issues are: first, that Markov blankets are not applicable to everything but only to every *thing* (3.1); secondly, that it is not clear what the system partitioned by the Markov blanket is (3.2); and thirdly, that there are properties of things that cannot be accounted for under the Markov blanket formalism (3.3). These three issues cast doubt on the universality of FEP. Additionally, we will raise some questions about the fundamentality of FEP and its relationship to physics (3.4). Finally, we will ask: why Markov blankets? We will claim that the actual reason is that the *Markov blanket trick* allows for the description of any *thing* as a kind of thing that engages in Bayesian inference (3.5).

3.1 A Theory of Every Thing?

Some things or systems cannot be described in terms of a Markov blanket. This is acknowledged in the FEP literature. The canonical example of a thing that cannot possess a Markov blanket is a candle flame (Friston 2013, Friston 2019). Put simply, due to the fast pace of the generation and destruction of molecular interactions in candle flames, it is impossible to find stable states in the system that could correspond to blanket states. Therefore, it is impossible to apply the Markov blanket formalism and to say that candle flames are *things* although they are clearly things.

The deeper reason why a thing like a candle flame cannot be a *thing* is that FEP requires the partition realized by the Markov blanket to be more stable than the relevant states of the system of interest. FEP does not handle well the fact that the states "that constitute a Markov blanket can, over time, wander away or, indeed, be exchanged or renewed" (Friston 2019, p. 50). This issue might entail that the Markov blanket formalism deals well with things like cells, that maintain a clear and stable boundary with their environment, but not with many other things like clouds or groups of people, in which such a boundary is more diffuse and changes over time. This issue is highly relevant as FEP has been claimed to apply at multiple scales of reality, from fundamental particles to social ensembles (e.g., Hesp et al. 2019; Ramstead et al. 2019; Veissière et al. 2020). The inability of the framework to accommodate changing blanket states draws into question whether the FEP could possibly have such wide applicability.[14]

The take home message of this section is that, while FEP may be a principle for a theory of every *thing* (i.e., a theory of such systems as can be properly characterized within the constraints of the FEP formalisms), it is not a principle for a theory of everything. This much is generally acknowledged in the FEP literature. This does undermine some of the more outlandish claims about the scope of the FEP as a totalizing and universal theory. To claim that you have a theory of every *thing* (excluding candle flames) is less impressive than to claim that you have a theory of everything. Nevertheless, if the FEP notion of *thing* is able to accommodate biological and cognitive systems within its formal constraints, FEP would still be a powerful framework.

3.2 On Markov Blankets and States

The second concern regarding the applicability of the Markov blanket formalism has to do with the partition of the systems in which it applies. In some sense, this concern is more fundamental than the previous one as it points out that the relationship between dynamics that underlie FEP and the partition realized by Markov blankets is not well defined. Take three different systems to which the Markov blanket partition has been applied:

[14] The literature on FEP is starting to address this issue. One example is a computational simulation that, assuming the existence of all states, shows that they can change in function over time (Ramstead et al. 2020a). This might have the ability to account for some of the tricky situations in which the blanket states of a system change over time. The simulation, however, is still too abstract to decide in favor or against its applicability to real systems.

- *Brains or cognitive systems*: brain states are the internal states (***μ***) and environmental states are the external states (***η***). The blanket states are sensory receptors (sensory states, ***s***) and motor-receptors (active sates, ***a***). (See top of figure 2 above).
- *Bacillus*: intracellular states of the Bacillus are the internal states (***μ***) and environmental states are the external states (***η***). The blanket states are cell membrane (sensory states, ***s***) and the actin filaments of the cytoskeleton (active states, ***a***). (See bottom of figure 2 above).
- *Coupled pendulums*: in a system of two pendulums connected by a beam, each pendulum is considered its own internal state (***μ***) and the external state of the other pendulum (***η***). The beam is the Markov blanket (***b***), the velocity of pendulums corresponds to sensory states (***s***) and the position of pendulums corresponds to active states (***a***). (See Kirchhoff et al. 2019).

In these three examples, the Markov blanket partition is stipulated as being located exactly wherever happens to be convenient for the purposes of the model. What counts as internal, external, or blanket state bears little relationship with structural properties or boundaries of the systems of interest. For instance, active and sensory states are aspects of the "blanket" in the case of the brain—as sensory and active states can be abstractly modelled as being at both "ends" of the nervous system in the form of sensory and motor receptors—but not in the case of the coupled pendulums, where the "blanket" is the beam but active and sensory states are properties of the pendulums (position and velocity).

This observation points to the idea that the Markov blanket formalism does not follow from a simple answer to the question 'What does it mean for a thing to be different from its environment?'. In coupled pendulums, for instance, an obvious answer to this question is "the atoms in the surface of the pendulums". Actually, it seems that the beam itself has little to do with the pendulums being things or not. In this sense, there seems to be no reason to choose the beam and not the atoms in the surface of the pendulums as the blanket states of the system or *vice versa*, other than that the beam might be a better selection to model the coupling event. Namely, the selection of the Markov blanket partition in the case of coupled pendulums does not prescribe the way they count as *things*. On the contrary, the very selection of that partition shows lurking assumptions about what the *things* in the system are or about what the *things* in the system *must be* for the principle to hold—these lurking assumptions are present in even the most recent algorithms to identify blanket states (e.g., Friston et al. 2021). This might not be an issue insofar as any kind of modeling requires some commitments and assumptions, but it does make FEP fail to follow from simple responses to basic questions as the proponents usually claim.

A deeper issue in this context is that what counts as a system or as a *thing* shows some incoherence in the FEP literature. Under FEP, existing *things* are taken to be random dynamical systems (x) that have attained a NESS as described in equations (1) to (3) (see Friston 2019, p. 4 & ff.). If this is true, then brains and bacilli, for example, are not things but parts of things. From equation (5) and figure 2, it follows, for a system x, the partition realized by the Markov blanket includes the set of all possible states ($\boldsymbol{\eta}$, $\boldsymbol{\mu}$, $\boldsymbol{b}$). Given this, it is straightforward to see that external states $\boldsymbol{\eta}$ are themselves part of x. In this sense, if the *thing* is described as x, the *thing* is the brain-environment system or the bacillus-environment system, but not the brain or the bacillus themselves.

In the FEP literature, however, the *thing* seems to be the brain or the bacillus and not the brain-environment system or the bacillus-environment system. The bacillus, for instance, seems to be the one actively reducing the entropy of its internal states to remain a thing. However, if the Markov blanket partition is applied to a thing like x, the only viable interpretation is that the *thing* is the whole system and not just a part of it. A possible answer to this issue is that, under FEP, x is just a "system" and the *thing* only emerges after the partition realized by the Markov blanket. In this sense, the thing would be identified not with x but with the particular states ($\boldsymbol{\mu}$, $\boldsymbol{s}$, $\boldsymbol{a}$) after the partition. If that's the case, this issue just reduces to the previous one: as the selection of a specific partition shows lurking assumptions about what *things* are, the preconceptions and modelling needs regarding what a specific *thing* is guides the partition and not the other way around. However, this possibility seems incompatible with several claims made by the proponents of FEP, such as the claim that the NESS of x is the "-ness" of "thingness" (Friston 2019, p. 5).

Another possible answer to this problem might be that there is no specific nature or scale of things (Friston 2019, p. 4), but that everything can be understood in terms of nested Markov blankets. If this is the case, then both the bacillus and the bacillus-environment system are *things* and either of these systems can be investigated independently of the other; it's simply a matter of selecting the right Markov partition for the system of interest (see also Ramstead et al. 2019). This answer captures one important observation regarding biological and cognitive systems: both are *complex systems* and complexity can only be understood by observing the system at different scales. Thus, a framework based on nested scales/systems seems to be the right way to go. However, it is not clear how the Markov blanket formalism applies to this. For instance, a system of coupled pendulums can be taken to be a *thing* measured by the relative phase between them (Haken et al. 1985; Kelso 1995). In this situation, we can take it that relative phase measures the dynamics of x. These dynamics can then be partitioned, but what is the partition adding to the explanation of the dynamics of x other than making them amenable for use within a Bayesian formalism? The answer to this question is not obvious. Another possibility is to take relative phase to be measuring the dynamics of the internal states (μ) of the system. In this case, the Markov blanket partition would provide a set of external states (η) that would be outside the dynamics of the coupled pendulums themselves, but what does identifying these external states add to the explanation of the dynamics of x, which we have already described in terms of relative phase? The answer to this question is not obvious either.[15]

[15] Some claim that the Markov blanket formalism postulates *fictive* external states (Ramstead et al. 2020a). Namely, the external states do not need to be real states but just fictive states needed for the inferential framework to work. This seems

The take home message of this section is that the application of the Markov blanket formalism is not as straightforward as is suggested in the FEP literature. Not even in the case of "proper" *things*—i.e., things that are not candle flames. The partition realized by Markov blankets is not well defined and, effectively, it seems to depend on fundamental assumptions regarding *things* and modeling that are not explicitly acknowledged in FEP. The Markov blanket formalism does not provide a principled way to distinguish *things* from their environments. On the contrary, Markov blankets seem to work as a tool to formalize a statistical distinction between sets of states that are *already* taken to be the states of a *thing* and the states of its environment. Although this issue further erodes the status of FEP as a general principle for a theory of everything, it should not be seen as entailing a defeat of the framework. It is possible that, in the long run, FEP's assumptions regarding thingness and modeling will turn out to be right, or at least useful. Time will tell.

3.3 Your Feet Outside the Blanket!

We have seen that Markov blankets cannot be applied to everything, only to every *thing*, and that the notion of *thing* at play in FEP precedes the Markov blanket formalism. In this sense, the formalism does not provide a principled way to describe *things* but is a tool to formalize previous assumptions regarding thingness. There is, however, a different aspect of the applicability of Markov blankets that is not captured by the discussion of thingness: the *properties*, *processes*, and *events* that can be described in terms of FEP.

As already mentioned, Markov blankets were first described by Judea Pearl (1988) in the context of graphical models and, more concretely, of Bayesian networks. In these networks, nodes are states (or factors) of a process/system and edges are the statistical relationships between those states.

to be an ad-hoc move: it saves situations like the one described, but it makes unintelligible other situations in which real external states seem to be required—e.g., the bacillus-environment system.

As far as Markov blankets make sense in this context, the application of the Markov blanket formalism under FEP entails an (at least) *implicit* commitment to the idea that the relevant aspects of the systems of interest can be modeled in terms of these networks. This commitment is not properly addressed in the literature and some aspects (or measures) of *things* do not seem amenable to the description entailed by Bayesian networks. In other words, we think there are highly relevant properties of biological and cognitive systems that fall outside the Markov blanket. We focus on two of them: *relational features* and *constitutive self-organization*.

A very simple example of a property of things that seems to be difficult to capture in terms of the statistical partition required under the Markov blanket formalism takes the form "x is taller than y". We can say, for instance, "the table is taller than the chair" or "Satoshi is taller than Vicente". In both cases, we are stating some measurable property of some *things*. However, this property does not seem to be properly described in terms of statistical independences between tables and chairs or between Satoshi and Vicente, nor in terms of internal and external states: how could the property "taller than" be an internal or external state of a *thing*? It seems even less true to say that tables, chairs, Satoshi, or Vicente are engaging in any kind of inference or minimization of $\boldsymbol{F}$ with regard to this property.[16] Thus, it is not clear that the kind of framework entailed by the Markov blanket formalism accounts for all possible measurable aspects of *things* and, more concretely, it seems *relational features* fall outside the blanket.

Interestingly, we do not need to dig too deep to find examples of fundamentally relational properties of cognitive systems. For instance, the navigation of sparsely populated environments can

[16] For Satoshi and Vicente it might be said the property "Satoshi is taller than Vicente" entails some specific expectations that would make them predict the probability of their sensory inputs in some ways and not in other ways, and therefore they'd be minimizing $\boldsymbol{F}$. However, even if such a claim were true, the minimization of $\boldsymbol{F}$ would be happening with regard to some activity or experience of the two relata but not with regard to the relation itself.

be described in terms of the dynamics of the heading direction of the agent as they avoid obstacles and approach a goal (e.g., Fajen & Warren 2003; Warren 2006). In this case, it is important to note that "heading" is a relational feature in itself as the agent's heading direction is always with respect to some aspects of the agent's environment—i.e., there's not such a thing as "heading" outside a given environment. The property of "heading" seems to be difficult to describe in terms of just internal states of the agent or just external states of the environment. It might be possible to consider the position of an obstacle in the environment as an external state $\boldsymbol{\eta}_i$, the agents' will to avoid it as an internal states $\boldsymbol{\mu}_i$, and the change in heading as one active state $\boldsymbol{a}_i$, but it would require us to multiply the properties in the model to account for heading. It's also not clear what the benefit would be of multiplying the properties inside the model in this way as opposed to taking "heading" as a relational property of the organism-environment system. Thus, it is not clear that the partition entailed by the Markov blanket formalism is the best way to characterize this kind of situation.

The same can be said with regard to the notion of affordance in the cognitive science literature (Gibson 1979; Chemero 2003; Heras-Escribano 2019). Affordances are commonly understood as *relationships* between environmental features and abilities of organisms (Chemero 2003; Chemero 2009). Therefore, affordances are an example of the kind of relational feature that seems difficult to accommodate within the Markov blanket formalism. For instance, the fact that a step is climbable for a given organism does not seem to be well captured in terms of the statistical partition between organism and environment required under the Markov blanket formalism. Warren (1984) shows that to perceive a step as climbable depends on a tacit relationship between the height of the step and the leg length of the actor. Importantly, such a tacit relationship—i.e., such an affordance—either exists or does not, but it is not inferred and does not seem to be amenable to a description based on the statistical independence of internal and external states. It is difficult to imagine how the critical ratio of step height to leg length could be conceived purely as an internal state of the agent. Indeed, it is not clear that motor abilities or skills can ever be described purely in terms of internal states of the agent, given that such behavior is always organized relative to at least some structure in the environment (Baggs et al. 2020). This problem of characterizing affordances has led some proponents of FEP and active inference to re-define them in terms of action-selection preferences of the agents (e.g., Ramstead et al. 2016). However, the example of the "climbability" of a step clearly shows that affordances themselves do not have to do with selecting one action or another but with the very possibility of action given a relational property in the organism-environment system. The re-definition of affordances in terms of action-selection preferences is a good demonstration of FEP's inability to capture relational properties.

Heading and affordances illustrate the kind of *relational properties* that seem to be pervasive in cognitive activities and, to date, FEP seems unable to properly account for them. There is a deeper reason for this situation: whether cognition is an intrinsic property of agents or a relationship between agents and environments is a question as old as psychology (see Holt 1915; Raja 2019). As a framework, FEP takes cognition to be an intrinsic property of agents and, therefore, it encounters issues when accounting for aspects of cognition that seem to be better cast in terms of relations. The commitment to the intrinsicality of cognition is hardly unique to FEP. However, as with many other assumptions of the framework, when FEP is presented as a first principle for a theory of cognitive systems, this commitment is not explicit.

A different limitation of FEP is *constitutive self-organization*. It is usual to characterize living systems as self-organized, self-maintained systems—e.g., *autopoiesis* and operational closure (Maturana & Varela 1976; Varela et al. 1991; Di Paolo et al. 2017). Put simply, living systems actively and continuously create/produce themselves by maintaining a boundary that distinguishes them from their environment. The paradigmatic example of a self-organized, self-maintained system is the cell. Cells devote most of their interactions with their environment to maintaining their internal states and their cellular membrane. In this sense, cells have a very particular relationship to their cellular membranes: the membrane is not just a boundary between the internal states of a given cell and its environment, but it is also a *product* of the activities of those internal states and the interactions with the environment. Cells are self-organized, self-maintained systems by virtue of this productive character of their own boundary. We will call this process 'constitutive self-organization'.

Markov blankets are claimed to stand for the cellular membrane of the bacillus, for instance, and more generally to be capturing the fundamental autopoietic aspects of living systems (e.g., Kirchhoff et al. 2018; Wiese & Friston 2021). At the same time, Markov blankets are claimed to stand for sensory and motor receptors in the case of cognitive systems, for the connecting beam in the case of two coupled pendulums, or for changes in greenhouse/albedo effects and ocean-driven global temperature changes in the case of the biosphere (Rubin et al. 2020). This wide applicability of Markov blankets is taken to be one of the strengths of FEP. However, there are fundamental differences between cellular membranes, sensory and motor receptors, beams, and global temperatures. The very existence of cellular membranes is the *product* of the activities of cells. This is an explicit case of constitutive self-organization. Plausibly, constitutive self-organization may also be operating in the case of greenhouse effects or ocean-driven global temperature changes, but it is difficult to see how constitutive self-organization could be predicated of sensory and motor receptors. Although those receptors are somehow connected to the states of brains, it would be a stretch to say that their very existence is the product of the activity of those brains. In this sense, the blanket of cognitive systems is not the *product* of their internal states. Finally, it is even more implausible to think that the beam that connects the coupled pendulums is the *product* of the pendulums or vice versa. Therefore, we see different instances of the application of the Markov blanket formalism irrespective of their possible self-organized, self-maintained character.

The problem, as we see it, is that the Markov blanket formalism has no tools to account for constitutive self-organization. A partition in terms of a Markov blanket can be applied to living systems, of course. And sometimes Markov blankets may be identified with the natural boundaries of those systems. However, Markov blankets are applied to those boundaries *ad hoc*. There is nothing in the use of Markov blankets that accounts for the fundamental features of the boundary of self-organized, self-maintained systems.[17] This is, just like relational features, an aspect of biological systems that falls outside the blanket.

The take home message of this section is that there are some features of living and cognitive systems that cannot be accounted for within the framework provided by Markov blankets. The partition entailed by this formalism is not sensitive to constitutive self-organization and, therefore, it is blind to one foundational aspect that makes some systems living systems. Additionally, Markov blankets do not provide the best framework to account for relational properties between *things*. Some of those properties, like affordances, may be highly relevant for cognitive systems. Thus, we see again that the application of the Markov blanket formalism is not as straightforward and universal as suggested by FEP literature. We have seen that not all things are *things*. We have also seen that it is not clear what counts as a *thing*. And finally, we have seen that even undisputable *things* have properties and features that cannot be captured by Markov blankets. So, why Markov blankets? In the FEP literature, Markov blankets are presented as a non-controversial, assumption-free answer to the question 'What does it mean for a thing to be different from its environment?'. However, in the last three sections we have shown this is far from the case. This leads us to wonder why the Markov

[17] Some computational simulations show groups of particles engaging in the form of Bayesian inference granted by FEP (e.g., Friston 2013). An algorithm to identify the Markov blanket partition is then applied to the group and some of the particles are identified as blanket states. However, these simulations assume the existence of all the particles from the beginning and, therefore it cannot be said that their blanket states are the *product* of the activities of internal states.

blanket formalism is placed at the foundation of FEP. We answer this question in section 3.5, after a brief detour.

3.4 A Little Detour: On FEP and Fundamental Physics

A different way in which FEP is taken to be a principle for a theory of every *thing* has to do with the way some fundamental formalisms of physics follow from the one presented in equations (1) to (3). Put simply, by carrying out different transformations in those equations—e.g., by setting to zero either the curl-free flow ($\Gamma = 0$) or sinusoidal flow ($Q = 0$)—you get the fundamental equations of classical mechanics, statistical mechanics, quantum mechanics, and Bayesian mechanics. In this sense, FEP seems to be as fundamental as it gets. However, despite the remarkable interest of these relationships between FEP and these formalisms, it does not make FEP unique: there are other proposals from which you can get a similar level of fundamentality.

In 1957, Edwin T. Jaynes proposed the *maximum entropy principle* (MaxEnt; Jaynes 1957). Without getting into too many details, MaxEnt states that in order to make inferences with partial data, we must assume it comes from a maximally entropic distribution of data. Then:

> If one considers statistical mechanics as a form of statistical inference rather than as a physical theory, it is found that the usual computational rules, starting with the determination of the partition function, are an immediate consequence of the maximum-entropy principle. In the resulting "subjective statistical mechanics", the usual rules are that justified independently of any physical argument, and in particular independently of experimental verification; whether or not the results agree with experiment, they still represent the best estimates that could have been made on the basis of the information available. (Jaynes 1957, p. 620).

What Jaynes proposes, therefore, is that given some assumptions about what statistical mechanics are—e.g., that they are a form of statistical inference—you just need a statistical principle to derive all the fundamental formalisms in the field.[18] In this sense, MaxEnt is as fundamental as FEP and, actually, they are formally related (see Friston 2013).

A recent proposal by Stephen Wolfram (2002, 2020) is similar in terms of fundamentality. Wolfram claims that a fundamental theory of physics can be developed by using graphs or hypergraphs and a set of very simple computational rules. With just those tools, you can describe different topologies, time, energy, gravity, quantum mechanics, etc. (see Wolfram 2020 for a detailed analysis). Although these tools are different, the idea is pretty similar to MaxEnt and FEP: once you are willing to make some assumptions regarding basic aspects of physical systems, you can derive some or all fundamental formalisms of the field in a principled way. In the case of MaxEnt and FEP, these assumptions have to do with Bayesian inference and the conditions in which it may happen while, in the case of Wolfram, the assumptions have to do with computational algorithms and programs.

The question, however, is whether these principled ways to derive formalisms of physics are something more than a mathematical exercise. It is well known that MaxEnt suffers from a lack of results and empirical predictions (Dougherty 1994; Kleidon & Lorenz 2005). FEP is strongly criticized on the same grounds (Walsh et al. 2020). Wolfram's proposal likely shares this problem. Additionally, the three proposals make assumptions that have no clear physical significance. In the case of Jaynes, his proposal openly detaches statistical mechanics from its physical significance. Wolfram's proposal rests on computational rules that have no clear physical significance either. And we have shown that FEP is not completely clear in its account of different systems.

The point of this brief detour is already made: the fact that one can derive some formalisms from FEP is remarkable, but not special. There are other proposals able to do the same that,

[18] Jaynes acknowledges that this move can be taken to be just an information-theoretic/Bayesian gloss to Laplace's principle of insufficient reason (Jaynes 1957, pp. 621-622).

interestingly, seem to face similar empirical limitations. An evaluation of these proposals in term of the extent to which they provide an accurate account of physical systems is outside the scope of this paper. For now, it is enough to say that FEP is an instance of a class of proposals closely related to the foundation of physics.

3.5 The Markov Blanket Trick

We have shown so far that, in addition to several purely formal/mathematical issues (Biehl et al. 2021), the applicability of the Markov blanket formalism is neither as straightforward nor as universal as is implied in the FEP literature. We could include other reasons to challenge the use of Markov blankets. For instance, one of the consequences of the Markov blanket formalism is that the statistical independence between internal and external states *must* be mediated by a different set of states. But why should that be the case? Why couldn't internal and external states be statistically independent without being mediated by a blanket of states? It is perfectly possible to model, say, the statistical independence of the temperature of the skin of a person and the temperature of the air surrounding them without the need for modeling the skin in terms of Markov blankets. The fact that all states are somewhat independent of each other is acknowledged even by the main proponents of FEP (e.g., Friston et al. 2020, p. 7), so the need for a mediating set of states doesn't seem to be justified.

The analysis above suggests that Markov blankets are better viewed as a *tool* that allows for setting up a statistical boundary between *things* and their environments rather than as a principled way to describe and find such a boundary (see also Bruineberg et al. 2020). Why then should we use that tool and not another (e.g., *causal blankets*; Rosas et al. 2020)? We think the reason is that Markov blankets provide the kind of structure needed for Bayesian inference. Thus, by postulating Markov blankets as a fundamental assumption for "thingness", proponents of FEP ensure the applicability of a Bayesian model. By using Markov blankets, one can effectively model any *thing* as if it were a variational autoencoder (Kingma & Welling 2019) or a Helmoltz machine (Dayan et al. 1995), and Bayesian inference naturally follows from that. We refer to this move as *the Markov blanket trick*.

We have provided a derivation of Bayesian inference from FEP in section 2. Assuming a random dynamical system with a NESS and a Markov blanket, it can be shown that the particular states ($\boldsymbol{\pi}$) of the system can be described as engaging with their environment by minimizing variational free energy ($\boldsymbol{F}$) through a form of Bayesian inference. However, this derivation of Bayesian inference from FEP and its plausible application to describe some systems is not the way researchers arrived at variational free energy ($\boldsymbol{F}$) as a relevant quantity. On the contrary, FEP is a *post factum* proposal. Researchers (including Friston) were already using variational free energy ($\boldsymbol{F}$) as presented in equations (12) to (17) in the context of artificial neural networks and theoretical neuroscience more than a decade before the FEP was first proposed (see Dayan & Abbott 2001). In this sense, FEP is just re-conveying work already done in the cognitive sciences and generalizing it from the point of view of first principles. We think, however, that presenting FEP in terms of first principles obscures the real process by which Bayesian inference gets generalized under FEP.

Conceptually, the way variational free-energy ($\boldsymbol{F}$) or its analogous evidence lower bound (ELBO) are used in theoretical and computational neuroscience has to do with the identification of cognitive activities with a form of inference (e.g., Dayan et al. 1995; Friston 2002, 2003, 2005; Stone 2012). Perception, the story goes, is about inferring the world from sensations which only provide partial information about the world itself. In this sense, perception is about discovering which hidden variables ($\boldsymbol{v}$; a.k.a., the world) have caused the observable variables ($\boldsymbol{u}$; a.k.a., sensations) brains have access to. In the Bayesian jargon, what brains are trying to do is to infer the true posterior density $p(\boldsymbol{v}|\ \boldsymbol{u})$ given just $\boldsymbol{u}$. As this inference is usually intractable, one way to deal with it is allowing brains to instantiate a recognition density $q_{\boldsymbol{u}}(\boldsymbol{v})$ and make it approximate the true posterior under a generative density $p(\boldsymbol{v},\ \boldsymbol{u})$ that instantiates the prior beliefs of the system. Having access just to sensations $\boldsymbol{u}$ and the generative density $p(\boldsymbol{v}, \boldsymbol{u})$, brains can approximate $q_{\boldsymbol{u}}(\boldsymbol{v})$ to $p(\boldsymbol{v}|\,\boldsymbol{u})$ by minimizing variational free energy $\boldsymbol{F}$ or analogously by maximizing the evidence lower bound (ELBO).[19] Both $\boldsymbol{F}$ and/or ELBO are used as objective functions , for instance, in variational autoencoders (Kingma & Welling 2019).

In explaining that $\boldsymbol{F}$ has been used since the 1990s in theoretical and computational neuroscience, we have just gone through the same process made explicit in equations (12) to (17). But this use of $\boldsymbol{F}$ is prior to FEP, which was built up on the use of Bayesian inference in computational neuroscience and not the other way around. This fact is not an argument against FEP. Actually, it is possible that FEP is a principled justification of Bayesian inference even though it is built on previous work on Bayesian inference. However, understanding the precedence of the use of variational free energy $\boldsymbol{F}$ in computational neuroscience over FEP highlights a deeper reason for the use of the Markov blanket formalism in the latter. The precedence of $\boldsymbol{F}$ over FEP entails a reversal in the logic of FEP as it is currently presented. The more recent literature tells a story that goes like this: if you have random dynamical systems with a NESS and a Markov blanket, then you have FEP, and then you have a principled justification for Bayesian inference. However, the historical development of FEP suggests that the actual logical flow is: if you are able to model any system as if it were an autoencoder/a Helmholtz machine, you can describe any system as engaging in Bayesian inference; Markov blankets permit you to model almost anything as if it were an autoencoder/a Helmholtz machine; thus we can model anything as engaging in Bayesian inference; therefore FEP holds.

The way Bayesian inference works in machine learning and deep learning (Goodfellow et al. 2016) depends on a very specific structure of the data source, the data itself, the system, etc. Put simply, you need a generative density ($p(\boldsymbol{v}, \boldsymbol{u})$) to harness the inference and you need a *partition* between hidden variables ($\boldsymbol{v}$), observable variables ($\boldsymbol{u}$), and systemic variables ($q_{\boldsymbol{u}}(\boldsymbol{v})$). It is easy to see how

[19] We have used the typical notation in the field (see Dayan & Abbott 2001).

Markov blankets provide that exact *partition* for any system: hidden variables are external states ($\boldsymbol{\eta}$), observable variables are blanket states ($\boldsymbol{b}$), and system states are internal state ($\boldsymbol{\mu}$). With this partition in place, Bayesian inference is possible: given a particular generative model, p ($\boldsymbol{\eta}, \boldsymbol{\mu} | \boldsymbol{b}$), and the input received through the blanket states, $\boldsymbol{b}$, the internal states of the system can instantiate a recognition density $q_{\mu}(\boldsymbol{\eta})$ that approximate the true posterior $\mathrm{p}(\boldsymbol{\eta} | \boldsymbol{\mu}; \boldsymbol{b})$ by minimizing $\boldsymbol{F}$. Notice, again, that this is what is described in equations (12) to (17) and, more importantly, that given this partition from the Markov blanket formalism, one just needs to attribute each term to the proper component of a system for the Bayesian inference to work. These "proper components" will be sensory receptors in some cases, parts of the cytoskeleton of a cell in other cases, ocean-driven global temperature changes, or the beam connecting two coupled pendulums in still other cases. The details of the partition are underdetermined by FEP because the only important consideration is the *partition itself* that grants the structure for Bayesian inference. In this context, it is not surprising that, for instance, FEP accommodates computational realism (Wiese & Friston 2021). Converting any system into a form of computational system—e.g., a kind of variational autoencoder—is both the main outcome and the main driving force of the principle and, more concretely, of the Markov blanket formalism.

In proposing the notion of a Markov blanket trick, we find inspiration in the *reparameterization trick* commonly used in variational autoencoders (Kingma & Welling 2019, pp. 22 & ff.). Put simply, in a network with random and deterministic nodes, the reparameterization trick is a variable change that allows for ELBO to be differentiated in the random nodes. This differentiation would be impossible without the trick consisting in transforming a given random node $\mathbf{z}$ into a function $q(\boldsymbol{\varphi}, \boldsymbol{x}, \boldsymbol{\varepsilon})$ that effectively *partitions* the deterministic and random components of $\mathbf{z}$ in different inputs. The reparameterization trick then allows for a straightforward form of learning based on backpropagation in a variational autoencoder. It should be clear why this trick inspires our characterization of the Markov blanket trick: the reparameterization trick is just a *tool* that implies a kind of partition for a specific algorithm to work. The trick does not say anything specific about the network in which it is applied nor does it determine any of its essential features. It is just a way to make an algorithm work. The same applies in the case of the Markov blanket formalism under FEP: it is not a principled account of the boundary between *things* and their environments, but a tool that permits the kind of statistical structure needed for the variational algorithms of Bayesian inference. This is the actual reason why Markov blankets are postulated in FEP and active inference, as shown by the actual historical development of the principle and its process theory.

4. Active Inference

The notion of FEP as a principle for a theory of every *thing* seems diluted after the analysis of the Markov blanket trick. A better way to understand FEP is as a modeling framework (see Andrews 2020; van Es 2020) that permits us to understand *some* properties of *some* systems in terms of Bayesian inference. In this sense, FEP is not different from the computational metaphor (Milkowski 2013) or from the dynamical hypothesis (van Gelder 1998). Which one of these frameworks is best to model biological and cognitive sciences is an empirical question we will not address here. However, FEP is said to entail *active inference* as a process theory. If FEP holds, so the story goes, the Bayesian mechanics instantiated by biological or cognitive systems can be described in terms of active inference, which can be understood as a specific model within the reinforcement learning framework. In the case of cognitive systems, active inference is said to account for many phenomena related to perception, action, decision making, planning, etc. (Da Costa et al. 2020). We cannot provide an exhaustive review of all of these phenomena here, but we will show that, to date, active inference models do not explain perception and action but presuppose them. To do so, we review a simple active inference model to illustrate our point (4.1) and then analyze how perception and action (4.2) are understood in it.

4.1 Active Inference

The main idea underlying active inference is that cognitive systems maintain their sensations within a viable range by minimizing two specific quantities through a process of (variational) Bayesian inference: variational free energy, $\boldsymbol{F}$, and expected free energy, $\boldsymbol{G}$. Cognitive systems do so by perceiving their environment and by acting upon it. In the context of active inference, *perception* is understood as improving the beliefs (i.e., the internal models) of the environment instantiated by the system and *action* is understood as moving around to gather the sensations predicted by those beliefs. The joint minimization of $\boldsymbol{F}$ and $\boldsymbol{G}$ ensures the system has correct beliefs about its environment and that its courses of action, known as *policies*, are the appropriate ones to maintain its sensations within satisfactory boundaries.

The (variational) Bayesian inference entailed by active inference requires a generative model. A simple generative model[20] for active inference takes the form:

$$p(\tilde{u}, \tilde{v}, \pi, \beta) = p(\pi)P(\beta) \prod_{t=1}^{T} p\,(u_t|v_t)p(v_t|v_{t-1}, \pi) \tag{19}$$

$$p(v_t) = Cat(A)$$

$$p(v_t, \pi) = Cat\left(B\big(a = \pi(t)\big)\right)$$

$$p(v_0) = Cat(C)$$

$$p(\pi) \approx \sigma\big(-G(\pi)\big)$$

Although the formalization seems complicated, the generative model is just the joint probability p of the vector observable variables, $\boldsymbol{\tilde{u}}$, the vector hidden variables, $\boldsymbol{\tilde{v}}$, the policies, $\boldsymbol{\pi}$, and some parameters, β, written as $p(\boldsymbol{\tilde{u}}, \boldsymbol{\tilde{v}}, \boldsymbol{\pi}, \beta)$. The first line of equation (19) says that, in a chunk T of time, the joint

[20] Adapted from Friston et al. (2017) with the notation used in section 3.5 (Dayan & Abbott 2001). Other examples of generative models, with many more details, may be found in Da Costa et al. (2020) and Hesp et al. (2021).

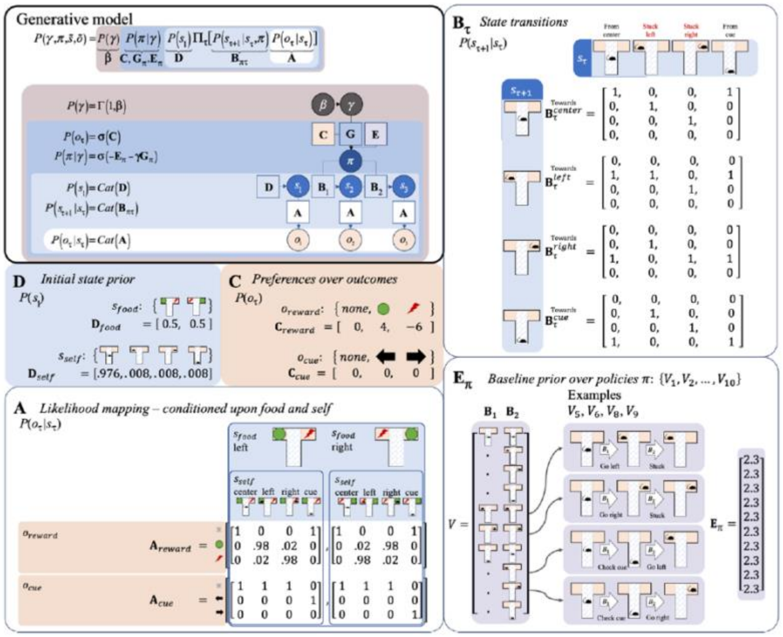
probability is equal to the probabilities of the policies and the parameters, $p(\boldsymbol{\pi})$ and $p(\boldsymbol{\beta})$, multiplied by the product of the probability of the observable variable, $\boldsymbol{u}_t$, given the hidden variable, $\boldsymbol{v}_t$, and the joint probability of the change of the hidden variable at every time step, $\boldsymbol{v}_t |\ \boldsymbol{v}_{t-1}$, when a particular policy $\boldsymbol{\pi}$ is selected. In cognitive systems, for instance, this generative model would describe all the possible relationships between different courses of action ($\boldsymbol{\pi}$), how they affect the relation of the system with its environment while acting on it ($p(\boldsymbol{v}_t |\ \boldsymbol{v}_{t-1}, \boldsymbol{\pi})$), and the likelihood of different sensations (i.e., observable variables) given the environmental states ($p(\boldsymbol{u}_t |\ \boldsymbol{v}_t)$).

For this generative model to work in T, some *prior* knowledge is needed. This is reflected in the following lines of equation (19). These lines say that each of the probabilities p of the generative model are gathered from the categorical matrices $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$ that are parametrized by β. Matrix $\boldsymbol{A}$ contains the values of the probabilities of all possible pairs $\boldsymbol{u}_t$ x $\boldsymbol{v}_t$ and is called the *likelihood mapping* matrix. Matrix $\boldsymbol{B}$ specifies how hidden states $\boldsymbol{v}_t$ evolve from one point in time to the next one given the current action $\boldsymbol{a}$ that pertains to a policy $\boldsymbol{\pi}$, and is called the *state transition* matrix. Finally, the $\boldsymbol{C}$ matrix specifies the *prior expectations* about hidden states $\boldsymbol{v}_t$ and (see figure 3 for concrete examples of these matrices). The three matrices are fundamental components of the generative model that harness the inference that makes the recognition density instantiated by the system, $q(\boldsymbol{x})$, where $\boldsymbol{x} = \{\boldsymbol{v}_{1:T}, \boldsymbol{\pi}, \beta\}$, approximate the true posterior, $p(\boldsymbol{x}|\ \boldsymbol{u}_{1:T})$. This approximation is carried out by minimizing variational free energy, $\boldsymbol{F}$, and expected free energy, $\boldsymbol{G}$:

$$F = D_{KL}(q(x)||p(\tilde{u})) - ln\, p(\tilde{u}) \tag{20}$$

$$G = D_{KL}(q(u_t|\pi)||p(u_t)) - E(v_t)$$

$\boldsymbol{F}$ is just the variational free energy, as discussed above. $\boldsymbol{G}$ relates the probability of the observable variable $\boldsymbol{u}_t$ given a policy $\boldsymbol{\pi}$ with the expected value of the entropy, H, of the observation $\boldsymbol{u}_t$ given the model of the environmental states $\boldsymbol{v}_t$. Also, it appears as well in equation (19) as proportional to negative $p(\boldsymbol{\pi})$, ensuring the selection of the policy (i.e., the course of action) that minimizes expected free-energy. Otherwise, $\boldsymbol{F}$ ensures the recognition density, $q(\boldsymbol{x})$, is as a good model of the true posterior, $p(\boldsymbol{x}|\ \boldsymbol{u}_{1:T})$, as possible. Therefore, equations (19) and (20) formalize the way active inference accounts for action and perception in terms of (variational) Bayesian inference under a generative model by minimizing two forms of free energy.


Generative model
$P(\gamma,\pi,\tilde{s},\tilde{o}) = P(\gamma)\,P(\pi|\gamma)\,P(s_1)\Pi_\tau[P(s_{\tau+1}|s_\tau,\pi)\,P(o_\tau|s_\tau)]$
β
C, G$_\pi$, E$_\pi$
D
B$_{\pi\tau}$
A
$P(\gamma)=\Gamma(1,\beta)$
$P(o_\tau)=\sigma(\mathbf{C})$
$P(\pi|\gamma)=\sigma(-\mathbf{E}_\pi-\gamma\mathbf{G}_\pi)$
$P(s_1)=Cat(\mathbf{D})$
$P(s_{\tau+1}|s_\tau)=Cat(\mathbf{B}_{\pi\tau})$
$P(o_\tau|s_\tau)=Cat(\mathbf{A})$
β
γ
C
G
E
π
D
B$_1$
B$_2$
A
s_1
s_2
s_3
o_1
o_2
o_3
D Initial state prior
$P(s_1)$
s_{food}: { }
$\mathbf{D}_{food}$ = [0.5, 0.5]
s_{self}: { }
$\mathbf{D}_{self}$ = [.976, .008, .008, .008]
C Preferences over outcomes
$P(o_\tau)$
o_{reward}: {none, ● }
$\mathbf{C}_{reward}$ = [0, 4, −6]
o_{cue}: {none, ⬅ ➡}
$\mathbf{C}_{cue}$ = [0, 0, 0]
A Likelihood mapping – conditioned upon food and self
$P(o_\tau|s_\tau)$
s_{food} left
s_{food} right
s_{self}
center left right cue
o_{reward}
$\mathbf{A}_{reward}$ =
o_{cue}
$\mathbf{A}_{cue}$ =
$\mathbf{B}_\tau$ State transitions
$P(s_{\tau+1}|s_\tau)$
From center
Stuck left
Stuck right
From cue
s_τ
$s_{\tau+1}$
Towards $\mathbf{B}_\tau^{center}$ =
Towards $\mathbf{B}_\tau^{left}$ =
Towards $\mathbf{B}_\tau^{right}$ =
Towards $\mathbf{B}_\tau^{cue}$ =
$\mathbf{E}_\pi$ Baseline prior over policies π: {$V_1, V_2, \ldots, V_{10}$}
Examples
V_5, V_6, V_8, V_9
B$_1$ B$_2$
V =
Go left
Stuck
Go right
Stuck
Check cue
Go left
Check cue
Go right
$\mathbf{E}_\pi$ =

$$\mathbf{A}_{reward} = \begin{bmatrix}1&0&0&1\\0&.98&.02&0\\0&.02&.98&0\end{bmatrix},\begin{bmatrix}1&0&0&1\\0&.02&.98&0\\0&.98&.02&0\end{bmatrix}$$

$$\mathbf{A}_{cue} = \begin{bmatrix}1&1&1&0\\0&0&0&1\\0&0&0&0\end{bmatrix},\begin{bmatrix}1&1&1&0\\0&0&0&0\\0&0&0&1\end{bmatrix}$$

$$\mathbf{B}_\tau^{center} = \begin{bmatrix}1,&0,&0,&1\\0,&1,&0,&0\\0,&0,&1,&0\\0,&0,&0,&0\end{bmatrix}$$

$$\mathbf{B}_\tau^{left} = \begin{bmatrix}0,&0,&0,&0\\1,&1,&0,&1\\0,&0,&1,&0\\0,&0,&0,&0\end{bmatrix}$$

$$\mathbf{B}_\tau^{right} = \begin{bmatrix}0,&0,&0,&0\\0,&1,&0,&0\\1,&0,&1,&1\\0,&0,&0,&0\end{bmatrix}$$

$$\mathbf{B}_\tau^{cue} = \begin{bmatrix}0,&0,&0,&0\\0,&1,&0,&0\\0,&0,&1,&0\\1,&0,&0,&1\end{bmatrix}$$

$$\mathbf{E}_\pi = \begin{bmatrix}2.3\\2.3\\2.3\\2.3\\2.3\\2.3\\2.3\\2.3\\2.3\\2.3\end{bmatrix}$$

Figure 3. Example of a generative model for a T-maze setup in which a rodent receives a cue to learn where a reward is in the opposite leg of the T-maze. In the example, matrices **A** and **B** correspond to the *likelihood mapping* and *state transition*, respectively, and matrix **D** corresponds to the *prior expectations*. There are two other matrices, **C** and **E,** that correspond to preferences over outcomes and prior over policies respectively. For a detailed description of the figure and the generative model, see Hesp et al. (2021). (Figure from Hesp et al. 2021, p. 413; used under Creative Common Attribution License CC BY.)

4.2 Neither Perception nor Action: A Funny Kind of Generative Model

Our succinct overview above of the kind of generative models associated with active inference is sufficient to grasp two fundamental ideas. First, generative models are defined as the joint probability of all possible relationships among all relevant variables of the system. And second, generative models are used to harness inferences claimed to give rise to virtually any cognitive activity, including perception and action. In figure 3, for instance, a generative model is used in an inference in which a rodent perceives a cue in its environment and then selects a policy to move towards the position of a reward. This inference, that also includes some affective factors, is understood as underlying the cognitive activity of *planning* (Hesp et al. 2021, p. 403). The problem is that both in this specific example and in the general case, active inference presupposes perception and action without explaining them. And, as was the case with the Markov blanket trick above, what appears at first glance to be a theory-neutral set of assumptions about the phenomenon to be explained in fact hides a set of substantive and theory-laden claims about the phenomenon itself. The form and motivation of these hidden assumptions is the same in both cases. In the case of the Markov blanket trick, the hidden assumption was that things already have a boundary in an unproblematic sense, and because of this it is possible to model anything in terms of FEP and Bayesian inference (but in fact it is only possible to model any *thing*—which is a not well defined notion within the framework—in terms of FEP and Bayesian inference). In the case of active inference, the assumption is that perception and action are themselves a kind inferential process in which the animal updates its prior knowledge of the external situation based on incoming sense data, and because of this it is possible to model perception and action in terms of Bayesian inference too (but in fact this is only possible to the extent that perception and action *actually are* inferential processes and to the extent that the animal *actually does* have the appropriate prior knowledge, which are precisely the things being presupposed).

In active inference, perception requires a generative model to approximate the recognition density, $q(\boldsymbol{x})$, to the true posterior, $p(\boldsymbol{x}|\, \boldsymbol{u}_{1:T})$. This generative model plays the role of *prior knowledge* in the perceptual inference and is assumed to be the *right one* in most if not all the simulations in the literature (e.g., Buckley et al. 2017, Friston et al. 2017, Hesp et al. 2021). The generative model is assumed to provide the right statistical relationships between hidden and observable variables and to grant the right inferences from the latter to the former. In this sense, the generative model is fundamental to bridge the gap between a sensation and its environmental causes, which is arguably the central problem of perception (Friston 2002, 2003, 2005; Marr 1982/2010; Stone 2012; see Raja 2020 for a dissenting view). Therefore, the right generative model makes perception possible. But where does the model and its prior knowledge come from if not from perception? How is it possible to have knowledge of the relationship between sensations and their environmental causes *prior* to perception? There are two different kinds of answers to these questions in FEP literature.

The first kind of answer is inherited from predictive processing models and is based on the idea that the prior knowledge of the generative model comes from the perceptual process itself. Either the system lets sensations (i.e., observable variables) determine the generative model, therefore having a "funny kind of prior" (Hinton & van Camp 1994, p. 8), or the system holds a hierarchical structure in which the activity at each level of the hierarchy provides the prior for the level below (a.k.a., empirical Bayes; Friston 2002, 2003, 2005). Both solutions are explicit ways to avoid having to postulate innate prior knowledge in the system. However, the general intractability and irreversibility of the relationship between sensations and their environmental causes make the funny-kind-of-prior solution highly implausible. Additionally, empirical Bayes, although formally plausible, lacks any empirical support (Buckley et al. 2017, p. 75). Moreover, the concrete implementations of generative models point to further issues. In figure 3, it is clear that the generative model is the product of a rationalization of the task carried out by the experimenters: the different actions of the simulated rodent are classified in several possibilities and, after that, the generative model is described. But even the metric—the length of columns and rows—of the relevant matrices $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$ depends on this rationalization! This knowledge is accessible from the epistemological position of the experimenter that rationalizes the simulation, but it is very difficult to justify it as accessible from the epistemological position of the rodent. More importantly, perception is already *assumed* to work in the simulation: the rodent is assumed to perceive the cue properly without the need for going from whatever sensation it gets to the perception of the cue. Namely, even in a situation in which the central problem of perception is assumed to be solved, the generative model seems to provide an unjustifiable amount of prior knowledge.

The most recent answers to the origin of prior knowledge in generative models do not rely on perceptual processes anymore, but on the NESS of random dynamical systems (e.g., Friston 2019; Ramstead et al. 2018). The underlying idea is that, assuming that having a NESS makes random dynamical systems exhibit specific dynamical flows and that generative models are statistical descriptions of those flows, the NESS itself determines the generative model and the prior knowledge entailed by it. The generative model of a system *x* is determined and given by the very dynamics of *x* (e.g., Ramstead et al. 2018). This explanation, however, has some shortcomings both in the general case and in specific simulations. First, the NESS and all the relevant variables (hidden, observable, etc.) seem to be predicated of the whole organism-environment system (*x*). In this context, the Markov blanket trick gives rise to a partition of *x* that makes the notion of "hidden variable" meaningless. Whatever the hidden variables are, they are not hidden in *x*. Thus, the prior knowledge provided by a generative model is such just because the gap to bridge between a sensation and its causes is already bridged in *x*. The knowledge is already in *x* and the partition between variables is just an artifact of active inference. Second, it is difficult to see how active inference provides a description of a system *x* that is not already in the description of the NESS in this context. If active inference needs a statistical description of the NESS to run the inference, and the NESS already encompasses all the variables needed to understand the dynamics of the system, what is active inference providing other than a mathematical artifact for Bayesian inference? If otherwise the NESS is assumed to be present but not known, how can we even start knowing anything about the generative model that is not fully arbitrary (Luccio 2019)?

In particular simulations, these problems are connected with the psychologists' fallacy (James 1890): the idea that the description of the experimental situation as understood by the experimenter is an accurate description of how the participant solves the experimental task. In the example illustrated by figure 3, for instance, this connects again with building up a generative model from the epistemological position of the experimenters. If the experimenters know the NESS, the role of active inference is just a matter of reframing current knowledge in a Bayesian fashion. If experimenters do not know the NESS, the construction of the generative model is more about them than about the system in the simulation. In either case, the prior knowledge in the generative model is equally difficult to characterize.

A parallel issue affects action. Active inference deals with action selection in terms of the selection of policies π, but it does not deal with the central problem of action: motor control (Bernstein 1967; Meijer 2001; Raja 2020). For instance, in the simulation provided by Hesp et al. (2021), which is based on Friston et al. (2017), it is assumed that rodents are able to develop any policy π without providing details on how they do it. The ability of controlling the movements that constitute particular actions that then constitute particular policies is not explained or accounted for but *assumed* in the framework. In this sense, although active inference has been explicitly claimed to solve the central problem of action (e.g., Hipólito et al. 2020), the framework actually presupposes its solution by, for instance, assuming some form of internal inverse or forward model for the control of every policy (Friston et al. 2017). Thus, the issue regarding action is parallel to the issue regarding perception: it is not just that the generative model assumes some prior knowledge to bridge the gap between sensations and their environmental causes (i.e., perception) but also assumes knowledge about how to execute the selected behavioral policies (i.e., action).

This discussion shows that active inference does not account for perception and action, but both of them are presupposed in the form of a generative model that provides the prior knowledge needed for the Bayesian inference to work. This is an important issue as the kind of inferential architecture in which active inference was developed—predictive processing—was presented precisely as an alternative to other computational or connectionist models which required this kind of knowledge (see Friston 2002). In consonance with other criticisms (e.g., Luccio 2019), we think active inference and similar architectures, even when able to account for *some* cognitive capacities, are unable to account for the basic features of perception and action. And this inability is due to a necessity to postulate a generative model that provides an amount of prior knowledge that the framework has no resources to justify. This is likely the reason why the notion of generative model is still unclear in the relevant literature. The reader can find anything from a *sui generis reductio ad absurdum* of generative models where the consequence is assumed as a premise in the derivation (Parr & Friston 2019) to a situation in which the generative model is described as *just* a mathematical artifact and as a control system in the same paper (Ramstead et al. 2020b). And, more generally, the reader finds a struggle between realist and instrumentalist interpretations of generative models (van Es 2000).

This need for generative models makes active inference unable to account for perception and action and this issue makes generative models poorly described in the field. More generally, active inference fails to be the all-encompassing process theory entailed by FEP that its proponents want it to be.

5. Conclusions: The Unfulfilled Promise

In this paper we have provided a thorough analysis of FEP and active inference. Our conclusions are clear: FEP is not the general principle it claims to be, and active inference is not the all-encompassing process theory it is supposed to be. Our claim against the status of FEP as a principle is not about wording. We are perfectly fine with it being called the free energy *principle*. However, we think principles are such only to the extent that they provide a general account of some system or phenomenon (or both) from some general, evident assumptions. And we think FEP fails to do this.

We have shown that the use of the Markov blanket formalism does not follow from general, evident assumptions regarding physical, biological, or cognitive systems. On the contrary, it is a trick that allows for a very specific kind of modeling, viz. variational Bayesian inference. In this sense, FEP is a framework for modeling and active inference is its process theory. We have shown, however, that even as a modeling tool, active inference is not able to provide proper justification for the generative models it needs to work and falls prey of the same shortcomings similar inferential frameworks face.

We do not think the arguments provided here can or must be seen as final regarding the utility of FEP and active inference in the context of the biological and the cognitive sciences. After all, we do not think Hegelian arguments ever work. Nevertheless, we think we have provided enough arguments to show that the promise with regard to the ability of FEP and active inference to solve all the issues in the sciences of the mind or to provide a theory of everything are so far unfulfilled. There are both theoretical and empirical questions that must be addressed before we can consider FEP and active inference as serious contenders in the field.

References

Anderson, M. L., and Chemero, A. (2019). The world well gained. In M. Colombo, E. Irvine, and M. Stapleton (Eds.), *Andy Clark and His Critics* (pp. 161-173). London: Oxford University Press.

Andrews, M. (2020). The math is not the territory: Navigating the free energy principle. [Preprint: http://philsci-archive.pitt.edu/id/eprint/18315]

Baggs, E., Raja, V., and Anderson, M. L. (2020). Extended skill learning. *Frontiers in Psychology*, 11, 1956. https://doi.org/10.3389/fpsyg.2020.01956

Bernstein, N. A. (1967). *The Coordination and Regulation of Movements*. New York: Pergamon Press.

Biehl, M., Pollock, F. A., and Kanai, R. (2021) A technical critique of some parts of the Free Energy Principle. *Entropy*, 23, 293. https://doi.org/10.3390/e23030293

Bruineberg, J., Dolega, K., Dewhurst, J. and Baltieri, M. (2020). The emperor's new Markov blanket. http://philsci-archive.pitt.edu/18467/

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

Chemero, A. (2003). An outline of a theory of affordances. *Ecological Psychology*, 15, 181–195.

Chemero, A. (2009). *Radical Embodied Cognitive Science*. Cambridge: MIT Press.

Clark, A. (2015). *Surfing Uncertainty*. London: Oxford University Press.

Csiszár, I., and, Tusnády, G. (1983). Information geometry and alternating minimization procedures. *Statistics & Decisions*, 1, 205-237.

Da Costa, L., Parr, T. Sajid, N., Veselic, S., Neacsu, V., and Friston, K. (2020). Active inference on discrete state-spaces: A synthesis. *Journal of Mathematical Psychology*, 99, 102447. https://doi.org/10.1016/j.jmp.2020.102447.

Dayan, P., and Abbott, L. F. (2001). *Theoretical Neuroscience*. Cambridge, MA: The MIT Press.

Dayan, P., Hinton, G. E., Neal, R. M., and Zemel, R. S. (1995). The Helmholtz machine. *Neural Computation*, 7, 889-904.

Di Paolo, E. A., Buhrmann, T., and Barandiaran, X. E. (2017). *Sensorimotor Life: An Enactive Proposal*. Oxford, UK: Oxford University Press.

Dougherty, J. P. (1994). Foundations of non-equilibrium statistical mechanics. *Philosophical Transactions of the Royal Society A*, 346, 259–305.

Fajen, B. R., & Warren, W. H. (2003). Behavioral dynamics of steering, obstacle avoidance, and route selection. *Journal of Experimental Psychology: Human Perception and Performance*, 29, 343–362.

Feynman, R. (1972). *Statistical Mecha*nics. Reading, MA: Benjamin.

Fox, C. W., and Roberts, S. J. (2012). A tutorial on variational Bayesian inference. *Artificial Intelligence Review*, 28, 85-95.

Friston, K. (2002). Functional integration and inference in the brain. *Progress in Neurobiology*, 68, 113-143.

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

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B*, 360, 815-836.

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

Friston, K. (2013). Life as we know it. *Journal of the Royal Society: Interface*, 10, 20130475. https://dx.doi.org/10.1098/rsif.2013.0475.

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

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

Friston, K., FitzGerald, T., Rigoli, F., Schawartenbeck, P., and Pezzulo, G. (2017). Active inference: A process theory. *Neural Computation*, 29, 1-49.

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

Friston, K., Fagerhim, E. D., Zarghami, T. S., Parr, T., Hipólito, I., Magrou, L., and Razi, A. (2021). Parcels and particles: Markov blankets in the brain. *Network Neuroscience.* Advance publication. https://doi.org/10.1162/netn_a_00175

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Boston: Houghton Miffin.

Goodfellow, I., Y. Bengio, and A. Courville (2016). *Deep Learning*. Cambridge, MA: The MIT press.

Haken, H., Kelso, J. A. S., and Bunz, H. (1985). A theoretical model of phase transitions in human hand movements. *Biological Cybernetics*, 51, 347–356.

Hathaway, R. J. (1986). Another interpretation of the EM algorithm for mixture distributions. *Statistics & Probability Letters*, 4, 53-56.

Heras-Escribano M. (2019). *The Philosophy of Affordances*. Cham, Switzerland: Palgrave Macmillan.

Hesp, C., Ramstead, M. J. D., Constant, A., Badcock, P., Kirchhoff, M. D., and Friston K. (2019). A multi-scale view of the emergent complexity of life: A free-energy proposal. In G. Georgiev, J. Smart, C. Flores Martinez, and M. Price (Eds.), *Evolution, Development and Complexity* (pp. 195-227). Cham, Switzerland: Springer.

Hesp, C., Smith, R., Parr, T., Allen, M., Friston, K. J., and Ramstead, M. J. D. (2021). Deeply felt affect: The emergence of valence in deep active inference. *Neural Computation*, 33(2), 398–446. https://doi-org.proxy1.lib.uwo.ca/10.1162/neco_a_01341

Hinton, G. E., and Sejnowski, T. J. (1983). Optimal perceptual inference. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 1983, 448–453.

Hinton, G. E., and van Camp, D. (1993). Keeping the neural networks simple by minimizing the description length of the weights. *COLT '93: Proceedings of the Sixth Annual Conference on Computational Learning Theory*, 5-13. https://doi.org/10.1145/168304.168306.

Hinton, G. E., and Zemel, R. S. (1993). Autoencoders, minimum description length, and Helmholtz free energy. NIPS' 93: *Proceedings of the 6th International Conference on Neural Information Processing Systems*, 3-10.

Hipólito, I., Baltieri, M., Friston, K., and Ramstead, M. J. D. (2020) Embodied skillful performance: Where the action is. [Preprint: http://philsci-archive.pitt.edu/18121/].

Holt, E. B. (1915). *The Freudian Wish and its Place in Ethics*. New York: H. Holt & Company.

Hohwy, J. (2013). *The Predictive Mind*. London: Oxford University Press.

James, W. (1890). *The Principles of Psychology*. New York: Holt.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *The Physical Review*, 106(4), 620-630.

Kelso, J. A. S. (1995). *Dynamic Patterns*. Cambridge, MA: MIT Press.

Kingma, D. P., and Welling, M. (2014). Auto-encoding variational Bayes. *arXiv*, 1312.6114. https://arxiv.org/abs/1312.6114.

Kingma, D. P., and Welling, M. (2019). An introduction to variational autoencoders. *arXiv*: 1906.02691. https://arxiv.org/abs/1906.02691.

Kirchhoff, M. D., Parr, T., Palacios, E., Friston, K., and Kiverstein, J. (2018). The Markov blankets of life: autonomy, active inference and the free energy principle. *Journal of the Royal Society: Interface*, 15, 20170792. http://dx.doi.org/10.1098/rsif.2017.0792

Kleidon, A., and Lorenz, R. D. (2005). *Non-Equilibrium Thermodynamics and the Production of Entropy*. Berlin, Germany: Springer-Verlag.

Luccio, R. (2019). Limits of the application of Bayesian modeling to perception. *Perception*, 48(10). 901-917.

Marr, D. (1982/2010). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Cambridge, MA: The MIT Press.

Maturana, H., and Varela, F. J. (1980). *Autopoiesis and Cognition: The Realization of the Living*. Dordrecht, Netherlands: D. Reidel Publishing.

Meijer, O. G. (2001). Making things happen: An introduction to the history of movement science. In M. L. Latash & V. M. Zatsiorsky (Eds.), *Classics in Movement Science* (pp. 1–57). Champaign: Human Kinetics.

Milkowski, M. (2013). *Explaining the Computational Mind*. Cambridge, MA: MIT Press.

Neal R. M., and Hinton G. E. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. In M. I. Jordan (Ed.), *Learning in Graphical Models* (pp. 355-368). Dordrecht, Netherlands: Springer.

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

Parr, T., and Friston, K. (2019). Generalised free energy and active inference. *Biological Cybernetics*, 113, 495-513.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann Publishers.

Rao, R. P. N., and Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1), 79-87.

Raja, V. (2019). J. J. Gibson's most radical idea: The development of a new law-based psychology. *Theory & Psychology*, 29(6), 789-806.

Raja, V. (2020). Embodiment and cognitive neuroscience: The forgotten tales. *Phenomenology and the Cognitive Sciences*. https://doi.org/10.1007/s11097-020-09711-0

Ramstead, M. J. D., Badcock, P. B., and Friston, K. (2018). Answering Schrödinger's question: A free-energy formulation. *Physics of Life Reviews*, 24, 1-16.

Ramstead, M. J. D., Kirchhoff, M. D., Constant, A., and Friston, K. (2019). Multiscale integration: beyond internalism and externalism. *Synthese*. https://doi.org/10.1007/s11229-019-02115-x

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

Ramstead, M. J. D., Kirchhoff, M. D., and Friston, K. (2020b). A tale of two densities: Active inference is enactive inference. *Adaptive Behavior*, 28(4), 225-239.

Ramstead, M. J. D., Veissière, S. P. L., and Kirmayer, L. J. (2016). Cultural affordances: Scaffolding local worlds through shared intentionality and regimes of attention. Frontiers in Psychology, 7: 1090. https://doi.org/10.3389/fpsyg.2016.01090

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

Rubin, S., Parr, T., Da Costa, L., and Friston, K. (2020). Future climates: Markov blankets and active inference in the biosphere. *Journal of the Royal Society: Interface*, 17, 20200503. http://dx.doi.org/10.1098/rsif.2020.0503.

Stone, J. V. (2012). *Vision and Brain: How We Perceive the World*. Cambridge, MA: The MIT Press.

van Es, T. (2020). Living models or life modelled? On the use of models in the free energy principle. *Adaptive Behavior*. https://doi.org/10.1177/1059712320918678.

van Gelder, T. (1998). The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences*, 21, 615–665.

Varela, F. J., Thompson, E., and Rosch, E. (1991). *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, MA: MIT Press.

Veissière, S., Constant, A., Ramstead, M. J. D., Friston, K., and Kirmayer, L. (2020). Thinking through other minds: A variational approach to cognition and culture. *Behavioral and Brain Sciences*, 43, E90. https://doi.org/10.1017/S0140525X19001213.

Walsh, K. S., McGovern, D. P., Clark, A., and O'Connell, R. G. (2020). Evaluating the neurophysiological evidence for predictive processing as a model of perception. *Annals of the New York Academy of Sciences*. https://doi.org/10.1111/nyas.14321.

Warren, W. H. (1984). Perceiving affordances: Visual guidance of stair climbing. *Journal of Experimental Psychology: Human Perception and Performance*, 10(5), 683–703.

Warren, W. H. (2006). The dynamics of perception and action. *Psychological Review*, 113(2), 358–389.

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

Wolfram, S. (2002). *A New Kind of Science*. Champaign, IL: Wolfram Media.

Wolfram, S. (2020). *A Project to Find the Fundamental Theory of Physics*. Champaign, IL: Wolfram Media.