The Emperor's New Markov Blankets

Jelle Bruineberg (corresponding author)
Department of Philosophy, Macquarie University, Sydney, Australia
E-mail: jelle.bruineberg@mq.edu.au

Krzysztof Dolega
Institut für Philosophie 2, Ruhr-Universität Bochum, Bochum, Germany
E-mail: krzysztof.dolega@rub.de

Joe Dewhurst
Munich Center for Mathematical Philosophy, Ludwig-Maximilians-Universität München, Germany
E-mail: joseph.e.dewhurst@gmail.com

Manuel Baltieri (corresponding author)
Laboratory for Neural Computation and Adaptation, RIKEN Centre for Brain Science, Wako City, Japan
E-mail: manuel.baltieri@riken.jp

Abstract

Markov blankets have been used to settle disputes central to philosophy of mind and cognition. Their development from a technical concept in Bayesian inference to a central concept within the free-energy principle is analysed. We propose to distinguish between instrumental Pearl blankets and realist Friston blankets. Pearl blankets are substantiated by the empirical literature but can do limited philosophical work. Friston blankets can do philosophical work, but require strong theoretical assumptions. Both are conflated in the current literature on the free-energy principle. Consequently, we propose that distinguishing between an instrumental and a realist research program will help clarify the literature.

1 Introduction

The formal concept of a Markov blanket plays a central role in many recent formulations of the Free Energy Principle (FEP), and to applications of the FEP to the study of life and cognition within the active inference framework. Our aim in this paper is to present an overview of the history and development of this concept in the context of Bayesian inference, and then to argue that its use in the active inference framework has stretched the concept too far beyond its formal origins. This new use of the Markov blanket concept allows the proponents of active inference to draw (we think) unwarranted metaphysical conclusions on the basis of a purely mathematical formalism. We conclude that those wishing to use Markov blankets for these purposes are faced with a dilemma: either they stick to the original innocuous-but-metaphysically-uninteresting formulation; or they bolster it with novel metaphysical premises. However, in the latter case it is the additional premises and

not the mathematical construct itself that carries out most of the theoretical work leading to novel conclusions, undermining any claim that these conclusions simply follow from the original Markov blanket formalism.

The FEP is a mathematical framework, developed by Karl Friston and colleagues (Friston, Kilner, and Harrison 2006; Friston, Daunizeau, et al. 2010; Friston 2010; Friston, FitzGerald, et al. 2017; Friston 2019), which specifies an objective function that self-organizing systems need to minimize in order to ensure adaptive exchanges with their environment. This minimization is made possible through variational inference, a machine learning technique previously developed by (Neal and Hinton 1998). One major appeal of the FEP is that it aims for (and seems to deliver) an unprecedented integration of the life sciences (including psychology, neuroscience, and theoretical biology). The difference between the FEP and earlier inferential theories (e.g., (Gregory 1970)) is that not only perceptual processes, but also other cognitive functions such as learning, attention, and action planning (Friston 2010; Friston, FitzGerald, et al. 2017), can be subsumed under one single principle, the minimization of free energy, through the active inference framework. Furthermore, it is claimed that this principle applies not only to human and other cognitive agents, but also self-organizing systems more generally, offering a unified approach to the life sciences (Friston 2013b; Friston, Levin, et al. 2015).

Another appealing claim made by proponents of the FEP and active inference is that it can be used to settle fundamental metaphysical questions in a formally motivated and mathematically grounded manner. The FEP has been used to (supposedly) resolve debates about the boundaries of the mind (Hohwy 2017; Clark 2017; Kirchhoff and Kiverstein 2019), the boundaries of living systems (Friston 2013b; Kirchhoff 2018; Kirchhoff et al. 2018), the relationship between mind and matter (Friston, Wiese, and Hobson 2020), has been proposed as an ordering principle by which the spatial and temporal scales of mind, life, and society are linked (Ramstead, Badcock, and Friston 2018; Veissière et al. 2020), and has been applied to the Earth's climate system in support of the Gaia hypothesis (Rubin *et al* 2020). In some places, FEP formalisms are explicitly presented as replacing (perhaps outdated) philosophical arguments (Ramstead et al. 2019; Ramstead, Friston, and Hipólito 2020). A complicating factor here is that the core of the FEP rests upon an intertwined web of mathematical constructs borrowed from physics, computer science, computational neuroscience, and machine learning. This web of formalisms is developing at an impressively fast pace and the constructs it describes are often assigned a slightly unconventional meaning whose full implications are not always obvious.[1] While this might ironically explain some of its appeal, as it can seem to the layperson to be steeped in unassailable mathematical justification, it also risks the possibility of 'smuggling in' unwarranted metaphysical assumptions. To its critics the FEP can appear like a moving target, each time introducing new constructs that make the previous criticism inapplicable (see for example the exchange between (Sun and Firestone 2020b; Seth et al. 2020; Van de Cruys, J. Friston, and Clark 2020) and (Sun and Firestone 2020a)). Here we want to focus on just one of these formal constructs, the concept of a Markov blanket that comes originally from the literature on

[1] Some of the newest additions are non-equilibrium steady states (Friston 2019), dual information geometry (Parr, Costa, and Friston 2020), each requiring detailed knowledge of non-equilibrium thermodynamics and differential geometry.

Bayesian inference and graphical modelling, and demonstrate how it is now being used in ways that stretch it far beyond its innocuous formal definition.

We think there are a set of issues arising from the (mis)use of Markov blankets that threaten the possibility of the FEP doing the metaphysical work that its proponents expect it to carry out. In our view the FEP literature consistently fails to clearly distinguish between the 'map' (a representation of reality) and the 'territory' (reality itself).[2] This slippage becomes most apparent in their treatment of the concept of a Markov blanket. In statistics and machine learning, Markov Blankets are a formal property of nodes in a Bayesian network. They designate a set of nodes that essentially *shield* a random variable (or a set of variables) from the rest of the variables (Pearl 1988; Bishop 2006; Murphy 2012). Bayesian networks are typically used as useful abstractions of complex phenomena. By contrast, in the FEP literature Markov blankets are frequently assigned a status as worldly boundaries with a variety of different roles: they belong to the territory. This discrepancy in the use of Markov blankets is indicative of a broader tendency within the FEP literature, in which mathematical abstractions are treated as worldly entities with causal powers. By focusing here on the case of Markov blankets, we hope to give a specific diagnosis of this problem, and a suggested solution, but our analysis does have potentially wider implications for the general use of formal constructs within the FEP literature.

The aim of this paper is twofold. First, we want to explain how it has been possible for such an innocuous technical concept as a Markov blanket to come to be used in order to settle central debates in philosophy of biology and cognition. We will trace the development of Markov Blankets starting from their standard application in Bayesian networks, through the role they play in variational inference, to their use in the literature on active inference. We will argue that in the course of this transition (Friston 2012, 2013b; Friston 2019) a new and largely independent theoretical construct has emerged (Friston, Da Costa, and Parr 2020; Biehl, Pollock, and Kanai 2020; Rosas et al. 2020), one that is more closely aligned with notions of sensorimotor loops and agent-environment boundaries (Tishby and Polani 2011; Ay and Zahedi 2014). For this reason, we propose to distinguish between 'Pearl blankets' to refer to the standard use of Markov blankets and 'Friston blankets'[3] to refer to the new construct. While Pearl blankets are unambiguously part of the map, Friston blankets are best understood as part of the territory. Since these are different formal constructs with different metaphysical implications, the scientific credibility of Pearl blankets should not automatically be extended to Friston blankets.

The second aim of this paper is to use the above distinction between Pearl blankets and Friston blankets in order to critically assess claims resting on the application of Markov blankets to philosophical problems. We find that in many cases map and territory are not clearly differentiated, thereby conflating Friston and Pearl blankets to draw potentially unwarranted conclusions. We suggest that this literature would do well in differentiating between two different research programs, which we call 'inference *with* a model' and

[2] Arguments loosely along these lines have been developed in Andrews (2020) and van Es (2020).

[3] The authors wish to credit Martin Biehl for this name, after first pointing out to some of them the novelties introduced by Friston in his use of Markov blankets.

'inference *within* a model'. These two approaches differ not only in how they interpret Markov Blankets, but also in their overall goals:

'Inference *with* a model' assumes that a system in the world can be usefully described using the tools of Bayesian probability theory, for example in the form of graphical models. Markov blankets might then be utilised in these models as constructs for describing conditional independencies among variables, but both the blankets and the models exist only as tools for the scientist performing inference *with* a model, not as ontological truths about the intrinsic nature of a system 'out there in the world'. Another dominant assumption in the literature is that agents themselves use a generative model of their environment to perform inference. Understood in this way, the explanatory project for cognitive neuroscience is to discover what generative model an agent is using to infer the states of its environment, but here also Markov blankets are understood instrumentally, as properties of an agent's model of the world, not as real properties of the world itself.

'Inference *within* a model', on the other hand, seeks to understand inference as it is physically implemented in a system, and places literal Markov blankets at the boundary between the system and its environment. The 'model' within which these Markov blankets are used is usually understood ontologically: here the map *is* the territory – the system performing inference is itself a model of its environment, and its boundary is demarcated by Markov blankets. This ontological understanding of Markov blankets (unlike the above instrumental understanding) cannot simply be justified by pointing to the mathematical formalisms involved, nor can it be justified by pointing to the previous successes of inference *with* a model and Pearl blankets more generally. The resulting approach is quite far removed from an empirical and naturalistic research program, and might be better seen as a branch of formal metaphysics applied to a scientific framework. We will argue that although this approach might have interesting philosophical consequences, it is dependent upon additional metaphysical assumptions that are not themselves contained withing the Markov blanket construct.

In section 1 we introduce the formal machinery required for variational Bayesian inference, in order to lay the groundwork for our discussion in section 2 of the traditional role played by Markov blankets in probabilistic inference. In section 3 we present the active inference framework and different roles played by Markov blankets within this framework, which we suggest has ended up stretching the original concept beyond its original formal purpose (here we distinguish between the original 'Pearl' blankets and the novel 'Friston' blankets). In section 4 we expand on this suggestion, focusing specifically on the role now played by Markov blankets in distinguishing the sensorimotor boundaries of organisms, which we argue stretches the original notion of a Markov blanket in a philosophically unprincipled manner. Finally, in section 5 we consider some of the theoretical consequence of conflating these two different uses of the Markov blanket concept, and conclude that it would be more usefull and productive to keep the two clearly distinct from one another when discussing active inference and the FEP.

2 Variational Bayesian inference

The last twenty years in cognitive science have been marked by what can be called 'a Bayesian turn', with an emerging number of theories and methodological approaches appealing to or making use of Bayesian methods (Dayan et al. 1995; Knill and Richards 1996; Rao and Ballard 1999; Knill and Pouget 2004; Friston, Kilner, and Harrison 2006; Doya 2007; Clark 2013, 2015; Hohwy 2013). In particular, the application of Bayesian formulations to the study of perception and other processes described as problems of inference has generated a huge literature, highlighting a large interest in Bayesian probability theory for the study of brains and minds. In this section we will review the formal background to variational Bayesian inference to lay the foundations for what will follow.

2.1 Bayes theorem

In statistics, inference is the process by which one can estimate some hidden property, usually the state or a parameter, of a system given some (often uncertain and limited) evidence. For instance, how do we determine if a watermelon is ripe by knocking on it? Or how can a cognitive system estimate a presence of some object on the basis of the state of its receptors alone? From the perspective of Bayesian reasoning (Robert 2007; Berger 2013), one can approach these kinds of inferential problems by applying Bayes theorem to determine the optimal solution. Bayes theorem normally takes the following form: [4]

$$p(x \mid y) = \frac{p(y,x)}{p(y)} = \frac{p(y \mid x)p(x)}{p(y)} \tag{1}$$

This formula is a recipe for calculating the *posterior probability*, $p(x \mid y)$, of a hypothesis/hidden state x given observation y. The probability $p(x)$ captures a priori knowledge about state x (i.e., a *prior probability*), while $p(y \mid x)$ describes the *likelihood* of observing y when x is assumed. The remaining term, $p(y)$, represents the likelihood of observing y independently of the hidden state x and is usually referred to as the *marginal likelihood* or *model evidence*, and plays the role of a normalising factor that ensures that the posterior is expressed on the [0,1] interval and sums up to 1. In other words, the posterior probability $p(x \mid y)$ represents the Bayes optimal combination of prior information represented by $p(x)$ (e.g., what we know about ripe watermelons, before we get to knock on the one in front of us) and a likelihood model $p(y \mid x)$ of how observations are generated in the first place (e.g., how different (ripe or not) watermelons give rise to different sounds, including the observed y), normalised by the knowledge about the observations integrated over all possible hidden variables, $p(y)$ (e.g., how watermelons may sound, regardless of the specific maturation stage).

[4] To simplify the notation, we follow the convention used by standard treatments such as (Blei, Kucukelbir, and McAuliffe 2017), where we denote both variables and their value assignments using lowercase letters (i.e., X=x is assumed) while bold letters are used to denote vectors of variables (e.g., x=[x1,x2,…,xn]T).

Although this scheme offers a powerful tool for probabilistic inference, it is mostly limited to simple, low-dimensional, often discrete or analytically tractable problems. This can be easily seen when we consider the model evidence $p(y)$ as a normalisation term, computed as a marginal likelihood, i.e., a likelihood integrated over all possible hidden variables x:

$$p(y) = \int p(y, x)\, dx. \tag{2}$$

In practice, computing the exact model evidence is rarely feasible. The process is, in fact, often analytically intractable (i.e., no closed-form solution for the posterior) or computationally too expensive (i.e., a large of infinite number of hidden states x) (MacKay 2003; Beal 2003; Bishop 2006). To obviate some of the limitations of exact Bayesian inference schemes, different approximations can be deployed, which rely on either stochastic or deterministic methods. Stochastic approximations of Bayesian inference are based for example on Monte Carlo sampling (e.g., Markov Chain Monte Carlo/particle filtering approaches (Chen 2003; Bishop 2006; Murphy 2012)) and while very effective, they can be computationally expensive and in some cases may not offer the best analogy to describe brains and biological systems in their more natural, dynamic and fast-paced environments.[5] Deterministic approximations are often less precise but can arguably be more easily used as models of biologically plausible implementations. In this context, variational methods (Hinton and Zemel 1994; Jordan et al. 1999; MacKay 2003; Beal 2003; Bishop 2006; Blei, Kucukelbir, and McAuliffe 2017; Zhang et al. 2018) are a popular choice, including for the FEP framework discussed in this paper.

2.2 Variational inference

The main idea behind variational inference is that the problem of inferring the posterior probability of some *latent* or *hidden* variables from a set of observations (i.e., the posterior $p(x \mid y)$) can be transformed into an optimisation problem. Roughly speaking, the method involves stipulating a family Q of probability densities over the latent variables, such that each $q(x) \in Q$ is a possible approximation to the exact posterior. The goal of variational inference then is to find an optimal distribution $q^*(x)$ which is closest to the true posterior. The candidate distribution is often called the recognition or variational density, because the methods used employ variational calculus, i.e., functions $q(x)$ are varied with respect to some partition of the latent variables in order to achieve the best approximation of $p(x \mid y)$. In variational Bayes, the problem is stated using a common measure of dissimilarity between two probability distributions, the Kullback-Leibler or KL divergence (here denoted by D_{KL}):

$$q^*(x) = \text{□□□ □□□}_{q(x)\in Q}\, D_{KL}(q(x) \parallel p(x \mid y)). \tag{3}$$

By using this definition of KL divergence, one can obtain the following equation:

[5] See however (Sanborn and Chater 2016) for a different perspective.

$$D_{KL}(q(x) \parallel p(x \mid y)) = \int q(x) \log \frac{q(x)}{p(y \mid x)}\, dx = E_q[\log(\frac{q(x)}{p(y \mid x)})] \\ = E_q[\log q(x) - \log p(x, y) + \log p(y)], \tag{4}$$

where E_q denotes expectations with respect to the variational density $q(x)$.

The trick of variational Bayes consists in letting go of trying to minimise the KL divergence in equation (3) directly, shifting the objective to optimising a different functional which bounds the model evidence. Since $\log p(y)$ is constant with respect to $q(x)$, the KL divergence from the previous equation can be restated as:

$$D_{KL}(q(x) \parallel p(x \mid y)) = \log p(y) - L(x), \tag{5}$$

where

$$L(x) \equiv -E_q[\log q(x)] + E_q[\log p(x, y)] \tag{6}$$

is usually referred to as the evidence lower bound (ELBO) because it constitutes a lower-bound on the model evidence, such that $L(q(x)) \leq \log p(y)$ for any $q(x)$ (this follows from equation (5) and the fact that the KL divergence is always non-negative (Bishop 2006)). This lower bound is also commonly referred to, by analogy with free energy in statistical physics, as negative *variational free energy* (Hinton and Zemel 1994; Beal 2003; Murphy 2012). One can in fact see negative ELBO $-L(q(x))$ as the difference between an (expected) energy term and a (Shannon) entropy term

$$F(x) \equiv -L(x) = E_q[\log q(x)] - E_q[\log p(x, y)] \\ = -H(q(x)) + E_q[E(x)], \tag{7}$$

with

$$E(x) \equiv -\log p(x, y) \tag{8}$$

as an internal energy term (Murphy 2012). Crucially, minimising the free energy F, or maximising the ELBO, implies a minimisation of the KL divergence in equation (3) while, importantly, leaving the log-model evidence $\log p(y)$ unchanged. This then implies that

$$q^*(x) = \arg\min_{q(x) \in Q} F(x) \tag{9}$$

In the next section we will then look into *how* the problem stated in equation (9) is effectively solved by variational inference.

2.3 The mean-field approximation

One of the most crucial components of variational inference is the choice of a variational family Q. Typically, one proceeds by either introducing a parametrized family $Q(\theta)$ with parameters θ optimised to find an approximate posterior (e.g., θ might be the mean and covariance matrix for a family of Gaussian distributions (Opper and Archambeau 2009)) or by applying a *mean field approximation* commonly adopted in statistical physics (Parisi 1988), which considers M partitions of N hidden variables $x = \{x_1, x_2, \dots, x_N\}$[6] such that

$$q(x) = \prod_{i=1}^{M} q(x_i) \tag{10}$$

Using the calculus of variations (Beal 2003; Bishop 2006; Friston, Trujillo-Barreto, and Daunizeau 2008), one can then show that the optimal $q^*(x)$ (i.e., the one that maximises the ELBO or minimises variational free energy) is also a product of terms $q^*(x_i)$, which are (marginally) independent and can be expressed in the form of:

$$q^*(x_i) = \frac{1}{Z_{x_i}} \exp\left(E_{j\neq i}[\ln p(x,y)]\right) \tag{11}$$

where $Z_{x_i} = \int \exp\left(E_{j\neq i}[\ln p(x,y)]\right) dx$ is a partition function (i.e., a normalising factor) and $E_{j\neq i}$ denotes an expectation with respect to all partitions of variational density $q(x_j)$, $j = \{1, \dots, M\}$, excluding $q(x_i)$, i.e., each partition $q(x_i)$ is averaged with respect to all other partitions $q(x_{j\neq i})$[7]. To simplify some calculations and express this quantity in a more familiar form, most practical applications restate the above equation in terms of logarithms, so that:

$$\ln q^*(x_i) = E_{j\neq i}[\ln p(y,x)] - \ln Z_{x_i}, \tag{12}$$

By resting on a product of *independent* partitions, as expressed by equation (10), the mean-field approximation introduces a strong assumption on the relationships between different hidden variables $\in x$, essentially stating that different partitions of x do not exert a strong influence on each other and can thus assumed to be marginally (or unconditionally) independent. In particular, this means that the interactions of a partition with other partitions are assumed to be mediated only by their mean-field effects, i.e., their interactions correspond to the expected, or average, effects over all other partitions $E_{j\neq i}$ (Jordan et al. 1999; Fox and Roberts 2012), see equation (11). This constitutes a drastic simplification of the effective interactions between variables (Bishop 2006; Zhang et al. 2018), but it is often effective due to the mathematical tractability achieved with the simplified version of the inference problem, and the fact that in some (simplified) cases, mean-field effects can even exactly describe the solutions to some problems (Jordan et al. 1999). It is however crucial to

[6] Notice that M≤N, with M=N corresponding to a *fully* factorised variational density (Zhang et al. 2018), and M<N to the case where different partitions contain more than one element of x (Bishop 2006).

[7] Another way to see this is by rewriting equation (10) as q(x)=q(xi)q(xj≠i) to denote the variational density q(x) in terms of the product of one partition, q(xi), and all the remaining ones, q(xj≠i).

highlight that the mean-field assumption operates only on the variational density $q(x)$, and therefore does not encode the 'real' set of dependencies that may in fact exist among variables $x_i \in x$. As we shall see briefly, when one considers the 'real' set of dependencies in the joint probability $p(x, y)$ utilised to infer x via the posterior $p(x \mid y)$, it is possible to further simplify the inference problem by defining more specifically which partitions $q(x_{j\neq i})$ should be used to build the average $E_{j\neq i}$, i.e., all elements with $j \neq i$? Or only some? If so, which ones?

As it turns out, when there are relations of *marginal* or *unconditional* independence between variables $x_i \in x$ such that, for instance

$$p(x_1, x_2) = p(x_1)p(x_2) \tag{13}$$

these should be taken into account, so that marginally independent variables can be excluded from the set of variables with $j \neq i$ and thus from the computation of the average in equation (11). At the same time, it is interesting to note that another type of relation, i.e., *conditional* independence, can play a similar and in some cases even more impactful role (at least in terms of simplifying an inference problem). Two variables, x_1 and x_2, are said to be conditionally independent given a third one, x_3 if

$$p(x_1, x_2 \mid x_3) = p(x_1 \mid x_3)p(x_2 \mid x_3), \tag{14}$$

This corresponds, intuitively, to the idea that x_3 effectively 'shields' (or d-separates (Pearl 2009)) x_1 from x_2, and x_2 from x_1. As we will see in the next section, where this idea is unpacked in more detail, x_3 can also be said to be a 'Markov blanket' (in the Pearl sense (Pearl 1988)) for x_1 and x_2, such that no information about x_2 can improve estimates of x_1 or vice versa, when x_3 is known. Crucially, this implies that conditionally independent variables will also be excluded from the set $j \neq i$ and thus from the average in equation (11). In practice, this means that the factorised distributions $q(x_{j\neq i})$ left in equation (11) essentially constitute a 'shield' for the the partition $q(x_i)$.

While this description may be quite intuitive for a small number of variables (e.g., 'we will not consider x_2 when we compute the average effects on x_1 given the fact that x_3 'shields' x_1 from x_2'), describing analytically these dependence relations, and especially their effects on problems of variational inference, is not trivial. To overcome this impracticality, one is often inclined to look for alternative representations that can more easily expresses the relationships between variables $x_i \in x$, and the ensuing simplifications for the computation of optimal $q^*(x_j)$, while maintaining a rigorous mathematical formalism. In the next section we will introduce probabilistic graphical models as one way to accomplish this simplification, and then use them to make clear the concept of a Markov blanket as it first appeared in the FEP literature.

3 Markov blankets and probalistic graphical models

A common way to represent probabilistic models and their typical algebraic manipulations comes in the form of probabilistic *graphical* models. Probabilistic graphical models are a family of mathematical representations describing relationships between random variables using diagrams (Pearl 1988; Bishop 2006; Murphy 2012). Random variables are drawn as *nodes* in a graph, with shaded nodes usually representing variables that are *observed*, or empty ones used for variables that are latent, or hidden. The (probabilistic) relationships between such random variables are then expressed using edges connecting the nodes. These connections can be directed, conventionally depicted as arrows, or undirected, in which case simple lines are used. Although these relationships are formally defined in terms of basic manipulations on probability distributions (including the two fundamental operations of marginalization and conditioninalization (Bishop 2006)), graphical models provide some practical advantages in reasoning about these formal properties, presenting a clear and easily interpreted depiction of the relationships between variables.

For the purposes of the present manuscript we will focus on graphs with directed links, which provide the basis for Bayesian networks, and play a crucial role in the context of active inference (Friston, Parr, and Vries 2017). Standard introductions to these models and other types of graphical representations such as Markov random fields and factor graphs can be found in, for instance, (Pearl 1988; Bishop 2006; Murphy 2012).

3.1 Bayesian networks

Formally, a Bayesian network β is defined as:

$$\beta = (G, p), \tag{15}$$

where $G = (v, d)$ is a *directed acyclic graph* (DAG) consisting of a set of variables, vertices or nodes v and edges among them d, and p is a collection of tables containing dependencies between these variables as a set of stochastic matrices, i.e., matrices where all entries p_{ij} are nonnegative real numbers $0 \leq p_{ij} \leq 1$, and each row represents a probability density such that $\sum_j \quad p_{ij} = 1$. The graph G is often represented by an adjacency matrix A, such that $A(r, s) = 1$ for each edge $r \to s$ of nodes v in the graph g, i.e., the matrix A contains ones in positions (r, s) when there is a connection between node r and node s in g, and zeros for all missing connections in the graph. The tables p then contain the specific factorisation of a joint probability distribution over the variables v characterised, for a DAG, by the following equation (Murphy 2012):

$$p(v \mid G) = \prod_{i=1}^{n} \quad p(v_i \mid pa(v_i)), \tag{16}$$

where $pa(v_i)$ is the set of variables v_i depends on. This dependence relation is visually illustrated with connections in the graph G, using arrows originating from variables in the

set $pa(v_i)$ and terminating in v_i. Such relationships between the variables are often described using genealogical terms, with $pa(v_i)$ being the *parents*, or 'ancestors', of their *child*, or 'descendant', node v_i.

A simple Bayesian network can be found in Fig. 1, where variables $\{t, w, y, x, u, z\}$ are variously connected to exemplify different types of dependencies. Algebraically, this model can be expressed as

$$p(t, w, y, x, u, z) = p(z \mid x)p(x \mid w, y)p(u \mid y)p(t). \tag{17}$$

Graphically, the same relations can be represented in a model where the node t, which is completely disconnected from the rest of the network, is unconditionally independent from all other variables. The remaining variables then express the three canonical examples of (in)dependencies among 3-node graphs, constituting the basis for a general notion of d-separation (separation in directed acyclic graphs, or directed separation) provided in (Pearl 1988),

- w and y are marginally independent but only conditionally dependent if x is observed (i.e., when x becomes a shaded node), a case technically known also as *head-to-head* relation,
- w and z are marginally dependent but conditionally independent if x is observed, also known as *head-to-tail*,
- x and u are marginally dependent but conditionally independent if y is observed, also known as *tail-to-tail*.

For the sake of the topics discussed in this manuscript, it is worth stressing that, unlike other kinds of graphical models, e.g., the undirected Markov random fields, Bayesian networks can only be *acyclic*, meaning that no closed path can be followed from an initial node to go back to the same node.[8]

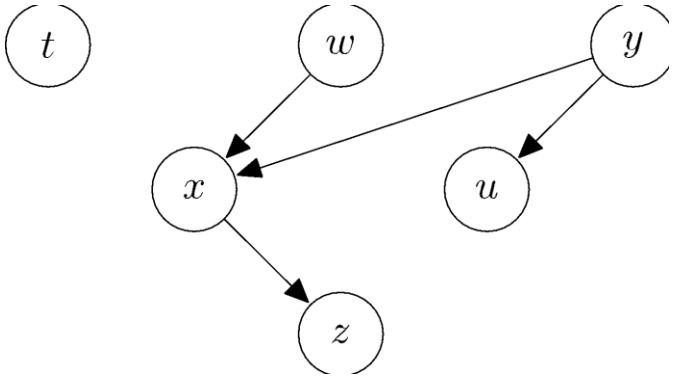

Figure 1: ***An example of a Bayesian network with different dependence relations among nodes.*** *This network represents the statistical relationships of six random variables. Here there is no observed (i.e., shaded) node, assuming that instances from different variables may be measured at different points in time. The edges*

[8] Note, however, that for many applications of cyclical models, the networks can be *unfolded* by indexing the state of each variable in time and forming an acyclic graph that serves as as a 'snapshot' of what is going on in the original model. For the application of this idea to the work on active inference and the FEP introduced in the next section, see (Kwisthout, Bekkering, and Van Rooij 2017).

take the form of arrows that indicate directed dependencies among variables. Marginal, or unconditional independencies can be seen, for instance, between the random variable t and all other variables in the graphs (i.e., no connections with other nodes). Conditional dependence instead appears in between w and z given x, and between x and u given y. On the other hand, w and y are marginally independent (no direct path between the two), but not conditionally given x (e.g., knowing y and measuring x will tell us something about w).

3.2 Markov blankets in Bayesian networks, or Pearl blankets

Bayesian networks play an especially prominent role in the visualisation of marginal and conditional independence relations introduced in the previous section, with the former represented by the lack of direct connections between two nodes, and the latter defined in terms of a set of nodes 'shielding' one variable (or set of variables) from all others. Shielding is usually cashed out using the notion of *d-separation* (Pearl 1988), heuristically defined above in terms of the three fundamental types of graphical connections that can be used to determine conditional independencies in any Bayesian network. Thus, the concept of (d-)separation can be used to describe the minimal set of nodes that renders a particular node conditionally independent of all other nodes in a Bayesian graph, also known as the *Markov blanket* of the node rendered conditionally independent (Pearl 1988) As the concept was first introduced by Judea Pearl, we will refer to Markov blankets in this traditional sense as *Pearl blankets* throughout the rest of the paper, in order to keep them distinct from the *Friston blankets* introduced in section 4.

Pearl blankets are especially relevant when it comes to visualising, understanding, and simplifying networks of considerable size. Thus, while trying not to overcomplicate our current presentation, we introduce a slightly bigger graph in Fig. 2 to showcase the presence of a shielding set of nodes in a graphical and hopefully more intuitive way. This network represents a joint density $p(y, x)$ with 2 observations and 18 hidden states variously connected. For example, we can define the Pearl blanket for node x_{10} (dashed border) as the set of nodes

$$mb(x_{10}) = \{x_4, x_5, x_9, x_{11}, x_{15}, x_{16}\} \tag{18}$$

These nodes are highlighted in Fig. 2 using thick borders, a slight abuse of notation but sufficient for our purposes. Here the Pearl blanket includes the parents of x_{10}, $\{x_4, x_5\}$, its children $\{x_{15}, x_{16}\}$, and the so-called *co-parents* of its children, $\{x_9, x_{11}\}$, i.e., all other nodes that the children of x_{10} depend on. Formally, a Pearl blanket for a set of variables x_i is thus equivalent to

$$mb(x_i) = pa(x_i) \cup ch(x_i) \cup copa(x_i), \tag{19}$$

where $pa(x_i)$ corresponds to the parents of x_i, $ch(x_i)$ to the children and $copa(x_i)$ to the co-parents of x_i respectively.

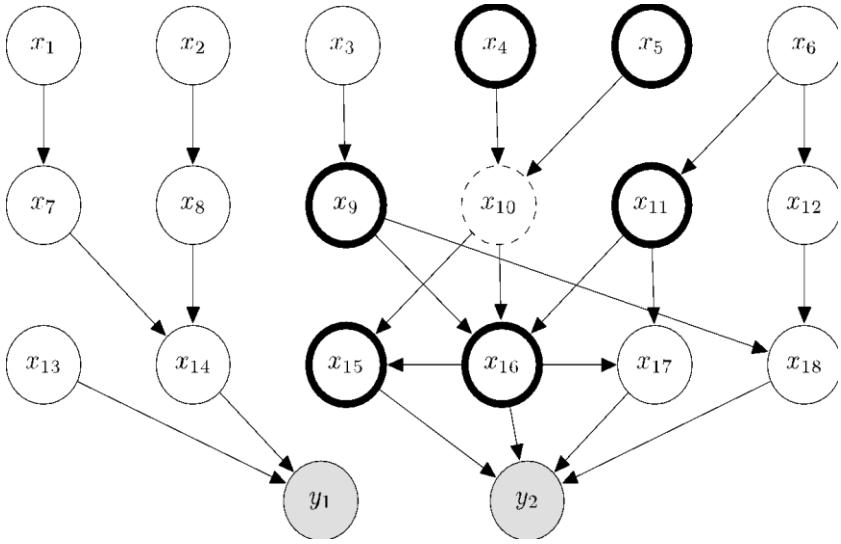

Figure 2: ***A Bayesian network describing the dependence of some observable variables*** y ***on hidden variables*** x ***following*** $p(y,x) = p(y \mid x)p(x)$***.*** *Thick lines are used for nodes constituting the Pearl blanket for a selected node* x_{10}*, depicted here with a dashed border.*

Importantly, the specification of conditional independencies, presented here in terms of Pearl blankets, can be used to simplify the approximation of the posterior when conducting variational inference, as we mentioned briefly at the end of the previous section. In particular, following the mean-field approximation introduced earlier, once the minimal set of conditional dependencies for a node is defined, the problem of finding the optimal variational density $q^*(x)$, given by the product of independent factors $q^*(x_i)$ specified in equation (12), can be simplified (following (Bishop 2006; Fox and Roberts 2012)) to the computation of

$$\ln q^*(x_i) = E_{mb(x_i)}[\ln p(y, x_i, mb(x_i))] - \ln Z_{x_i,mb(x_i)} \tag{20}$$

Here the mean-field effects, or expectations over all other independent factorisations $E_{j\neq i}[\ln p(y,x)]$, are now replaced by expectations with respect to the variables known to form the Pearl blanket of a node $E_{mb(x_i)}[\ln p(y, x_i, mb(x_i))]$, and the normalisation factor Z_{x_i} is replaced by $Z_{x_i,mb(x_i)}$, representing a new partition function calculated without the conditionally independent variables excluded using the Pearl blanket definition.

To see a practical application of this result, let's consider our example network in Fig. 2. For a given factorisation of the variational density $q(x)$ into 18 components $q(x_i)$, with $i = 1, \ldots, 18$ (i.e., one for each random variable forming the network), to compute $q^*(x_{10})$ initially we would have computed, following equation (12), a series of components

$$\begin{aligned} \ln q^*(x_1) &= E_{j\neq 1}[\ln p(y,x)] - \ln Z_{x_1} \ln q^*(x_2) \\ &= E_{j\neq 2}[\ln p(y,x)] - \ln Z_{x_2} \ln q^*(x_3) \\ &= E_{j\neq 3}[\ln p(y,x)] - \ln Z_{x_3} \ln q^*(x_4) \\ &= E_{j\neq 4}[\ln p(y,x)] - \ln Z_{x_4} \ \ldots \ln q^*(x_{10}) \\ &= E_{j\neq 10}[\ln p(y,x)] - \ln Z_{x_{10}} \ \ldots \end{aligned} \tag{21}$$

for all $j \neq i$, where the average of each factor is taken on the remaining 17 factors. However, thanks to the marginal independence between different groups of variables, e.g., between the 'left' and 'right' hand side sub-networks in Fig. 2, which are completely disconnected, and the conditional independence of other variables given their Pearl blanket, one gets

$$\begin{aligned} q^*(x_1) &= E_{mb(x_1)}[p(y,x)] - Z_{x_1,mb(x_1)}, \quad mb(x_1) = \{x_7\} \\ &= E_{mb(x_2)}[p(y,x)] - Z_{x_2,mb(x_2)}, \quad mb(x_2) = \{x_7, x_8\} \\ &= E_{mb(x_3)}[p(y,x)] - Z_{x_3,mb(x_3)}, \quad mb(x_3) = \{x_9\} \\ &= E_{mb(x_4)}[p(y,x)] - Z_{x_4,mb(x_4)}, \quad mb(x_4) \\ &= \{x_5, x_{10}\} \qquad (22) \\ &= E_{mb(x_{10})}[p(y,x)] - Z_{x_{10},mb(x_{10})}, \quad mb(x_{10}) \\ &= \{x_4, x_5, x_9, x_{11}, x_{15}, x_{16}\} \end{aligned} \quad \begin{aligned} &\text{...} \; q^*(x_2) \\ &\text{...} \; q^*(x_3) \\ &\text{...} \; q^*(x_4) \\ &\text{...} \; q^*(x_{10}) \\ &\text{...} \end{aligned}$$

This not only likely improves the inference process by excluding average effects from parts of the network that are completely disconnected from each other, but also further decreases the number of nodes used to calculate expectations to only the ones forming the Pearl blanket for each node. In larger networks the advantages of using Pearl blankets become even more obvious, considering for example graphs with hundreds or thousands of variables where mean-fields averages can now be computed using only a handful of nodes.

Having presented the basics of conducting Bayesian inference using probabilistic graphical models, and the way in which Pearl blankets can be deployed for variational inference, we now turn to the discussion of how the Markov blanket concept is used in the active inference framework, an approach to the study of biological and cognitive systems inspired by the FEP.

4 Markov blankets and active inference

Active inference is a process theory derived from the application of variational inference to the study of biological and cognitive systems (Friston, Daunizeau, et al. 2010; Friston 2013b; Friston, Rigoli, et al. 2015; Friston, FitzGerald, et al. 2017; Friston 2019). The core assumption underlying active inference is that living organisms can be thought of as systems whose fundamental imperative is to minimise free energy (this constitutes the so called 'free energy principle' (Friston 2010; Friston 2019)). Active inference attempts to explain action, perception, and other aspects of cognition under the umbrella of variational (and expected) free energy minimization (Friston, Daunizeau, et al. 2010; Feldman and Friston 2010; Friston, FitzGerald, et al. 2017)). From this perspective, perception can be understood as a process of optimising a variational bound on surprisal, as advocated by standard methods in approximate Bayesian inference applied in the context of perceptual science (see for instance (Dayan et al. 1995; Knill and Richards 1996; Rao and Ballard 1999; Lee and Mumford 2003; Friston 2005)). At the same time, action is conceptualised as a process that allows a system to create its own new observations, while casting motor control as a form of

inference (Attias 2003; Kappen, Gómez, and Opper 2012), with agents changing the world to better meet their expectations. Active inference integrates a more general framework where minimising *expected* free energy[9] accounts for more complex processes of action and policy selection (Friston, Rigoli, et al. 2015; Friston, FitzGerald, et al. 2017; Tschantz, Seth, and Buckley 2020). While a full treatment of active inference remains beyond the scope of this manuscript[10], here we wish to highlight the formal connections between this framework and the use of variational Bayes in standard treatments of approximate probabilistic inference (as described in the previous two sections). More specifically, we can ask what role Pearl blankets might play in active inference.

4.1 Pearl blankets in active inference

First we need to identify some of the formal notation used by active inference, which is related to the variational approaches described previously. Here we use the notation previously adopted in equation (9) to formulate perception and action as variational problems in active inference, specifying perception as the minimization

$$q^*(x) = \arg\min_{q(x)\in Q} F(x,\pi) \tag{23}$$

based on a process that generates an optimal bound on the posterior $p(x \mid y)$ (see equation (3), and characterizing action in terms of policies (i.e., sequences of actions) π where

$$\pi^* = \arg\min_{\pi} G(x,\pi,\tau) \tag{24}$$

This describes action selection as a minimisation of *expected* free energy, $G(x,\pi,\tau)$, based on beliefs about future and unseen observations y, up to a time τ. In doing so, we immediately notice that equation (23) essentially mirrors the previously defined equation (9), with the important caveat that in active inference, sequences of actions (i.e., policies π) are now a part of the free energy F. In a closed loop of action and perception, policies π can effectively modify the state of the world, generating new observations y, something that classical formulations of variational inference in statistics and machine learning do not consider, instead assuming fixed observations or data (MacKay 2003; Beal 2003; Bishop 2006).

Some formulations of active inference, especially the earlier ones (Friston et al. 2007; Friston, Trujillo-Barreto, and Daunizeau 2008; Friston 2008)), have thus explicitly relied on a set of assumptions similar to the ones highlighted in the previous section: a mean-field approximation and the use of Pearl blankets. The latter were seen as an integral part of the

[9] The free energy expected in the future for unknown (i.e., yet to be seen) observations, combining a trade-off between negative instrumental and negative epistemic values.

[10] For some technical treatments and reviews, see e.g., (Bogacz 2017; Buckley et al. 2017; Friston, FitzGerald, et al. 2017; Biehl et al. 2018; Sajid, Ball, and Friston 2019; Da Costa et al. 2020).

standard variational inference toolkit, where they are used to simplify the minimisation of variational free energy (or maximisation of the ELBO) by specifying which variables need to be considered for mean-field averages via appropriate constraints of conditional independence (see Fig. 2). The only noticeable difference in these formulations is in the very definition of the mean-field assumption, here implemented as 'structured', in the sense that variables x are partitioned in three independent sets ($M = 3$): hidden states and inputs, parameters, and hyper-parameters. In this case, the use of Pearl blankets is entirely consistent with existing literature and definitions of conditional independence in graphical models, if not slightly overzealous given the typical focus on a relatively low number of partitions. Indeed, it is not entirely clear what Pearl blankets actually add to this formulation, since it is often claimed that given a partition of variables 'the Markov [= Pearl] blanket contains all [other] subsets, apart from the subset in question' (Friston 2013b, 2008; Friston et al. 2007; Friston, Trujillo-Barreto, and Daunizeau 2008), where "*all* other sets" corresponds to $M = 2$. However, in more recent formulations of active inference the concept has been applied in a slightly different way, as more than just a formal tool.

4.2 From Pearl blankets to Friston blankets

In a number of recent theoretical and philosophical treatments based on ideas from active inference and the FEP, Markov blankets have been assigned a much more prominent role that cannot be explained just in terms of the formal properties of Pearl blankets. In some formulations of active inference, starting with (Friston and Ao 2012; Friston 2013b; Friston, Sengupta, and Auletta 2014), Markov blankets are in fact introduced as a tool to describe a specific form of conditional independence between a dynamical system and its environment, serving as a kind of boundary between organism and world.

As an emblematic example of this transition, we'll focus first on just one paper, Friston's 'Life as we know it" (Friston 2013b), where he presents a proof-of-principle simulation for conditions claimed to be relevant for the origins of life. This paper is often used as an example of how to extend the relevance of Markov blankets beyond the realm of probabilistic inference and into cognitive (neuro)science and philosophy of mind. The paper aims to show how Markov blankets spontaneously form in a (simulated) 'primordial soup'. This simulation consists of a number of particles that are moving through a viscous fluid. The interaction between the particles is governed by Newtonian and electrochemical forces, both only working at short-range. This, in turn, means that one third of the particles is then prevented from exerting any electrochemical force on the others. The result of running the simulation is something resembling a blob of particles (Fig. 3).

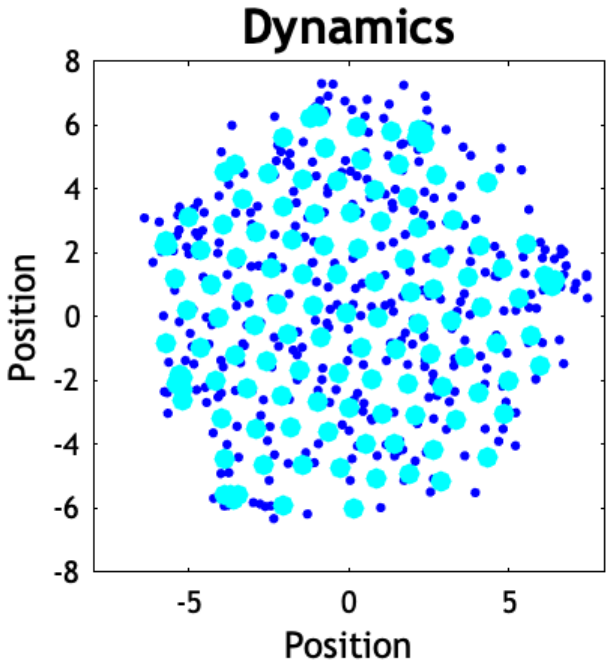

Figure 3: ***The 'primordial soup' simulated in (Friston 2013b).*** *The larger (cyan) dots represent the location of each particle. There are three smaller (blue) dots associated with each particle, representing the electrochemical state of that particle*

Using the model adopted in the simulations (for details please refer to (Friston 2013b)), one can then plot an adjacency matrix A based on the coupling (i.e., dependencies) between different particles at a final (simulation) time T, representing the particles in a 'steady-state' (under the strong assumption that the system has evolved towards and achieved its final steady state at time T, when the simulation is stopped). The adjacency matrix is itself a representation of the electrochemical interactions between particles, but can be interpreted as an abstract depiction of a Bayesian network. A dark square in the adjacency matrix at element r, s indicates that two particles are electrochemically coupled, and hence we could imagine that there is a directed edge from node r to node s (see notation in section 3.1). In this work, the directed edge is drawn if and only if particle r electrochemically affects particle s (Fig. 4). Because of the way the simulation is set up, the network will not be symmetrical (since a third of the randomly selected particles will not electrochemically affect the remaining ones).

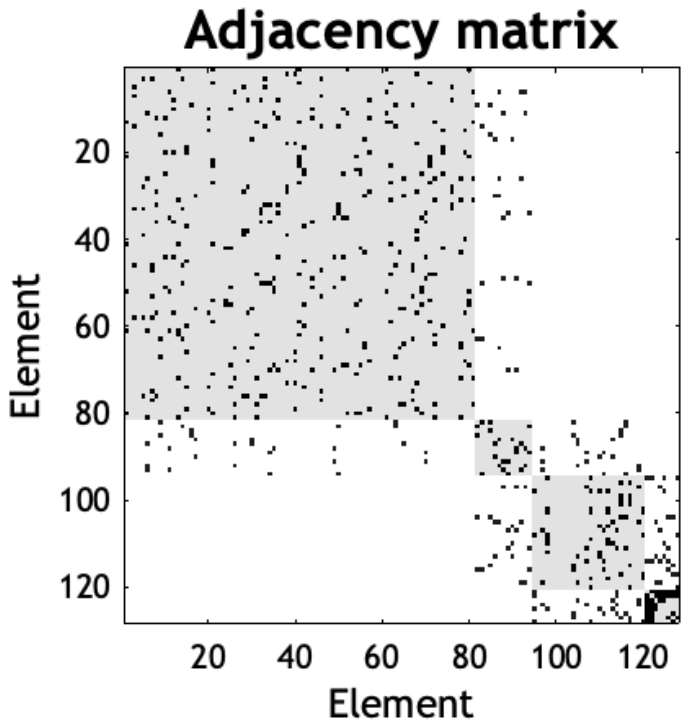

Figure 4: ***The adjacency matrix of the simulated soup at steady-state.*** *Element i, j has value 1 (a dark square) if and only if subsystem i electrochemically affects subsystem j. The four grey squares from top left to bottom right represent the hidden states, the sensory states, the active states and the internal states respectively.*

Spectral graph theory is then used to identify the 8 most densely coupled nodes, which are defined as the 'internal' states. Given these internal states, the Markov blanket is then found through tracing the parents, children and co-parents of children in the network (see Eq. 18 in (Friston 2013b)). As an extra interpretive step, the nodes in this Markov blanket can be further separated into 'sensory' and 'active' states. The 'sensory states' correspond to the parents of the internal states. The 'active states' correspond to the children of the internal states and their co-parents.[11] States that are not internal states and part of the Markov blanket are then called 'external states'. This procedure thus delivers four sets:

- ϕ: external states
- μ: internal states
- a: active states
- s: sensory states

Applied to the primordial soup simulation, each particle can be coloured to indicate which of these sets it has been assigned to (see Fig. 5). Given the dominance of short-range interactions and the density of particles, it should not come as a surprise that the particles that are labeled as active and sensory states form a spatial boundary around the states that are labelled as internal states. Given their placement in the simulated state space, this gives the impression that the active and sensory form a structure similar to a cell membrane.

[11] See section 5.1 for a discussion on the role of co-parents.

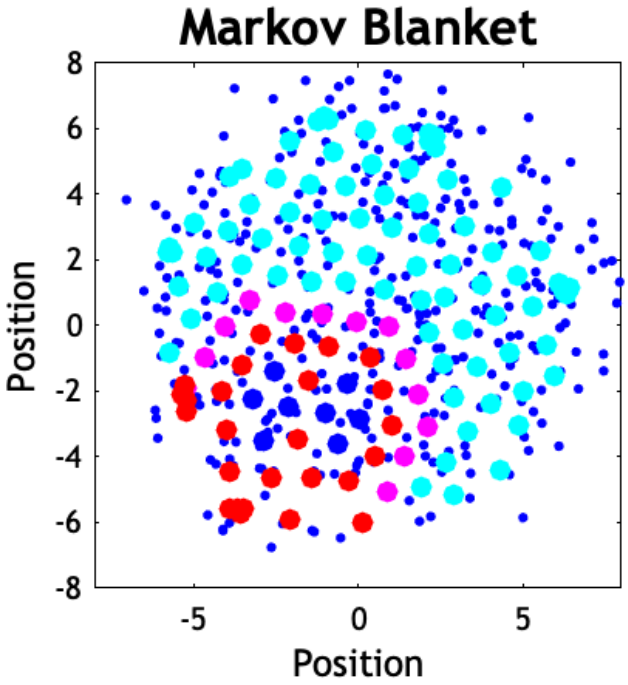

Figure 5: ***The Markov blanket of the simulated soup at steady-state in (Friston 2013b).*** *Figure reproduced using the code provided with (Friston 2013b). Similarly to Fig. 3 particles are indicated by larger dots. Particles which belong to the set of sensory states are in magenta, active states are in red, while internal states are in dark blue. A 'blanket' of active and sensory cells surrounding the internal particles can be seen.*

The 'Markov blanket formalism' advocated by Friston (2013b) and described formally above does most of the work in the FEP literature when it comes to identifying internal, sensory, active, and external states It is important to note that the partitioning of the primordial soup is not done directly, but requires a formal representation of the system. This formalizing step requires a number of additional assumptions that Friston does not provide any justification for. For example, it is unclear why only electrochemical interactions are used to construct the adjacency matrix while other forms of influence included in the simulation (such as Newtonian forces) are ignored. The demarcations made by analysing the adjacency matrix are than used to label the nodes in the original system (as in Fig. 5).

The simulation assumes that by viewing the system through the Markov blanket formalism, plus some additional assumptions about the separation of its states into different sets of variables, it is possible to uncover hidden properties of the target system which, in some sense, 'instantiates' or 'possesses' a Markov blanket. This procedure of attributing to the territory (the dynamical system) what is a property of the map (the Bayesian network) is a clear example of the reification fallacy: treating something abstract as something concrete (without any further justification). At the very least, we think that this way of using the formalism goes beyond the merely epistemic role that Markov blankets (in the Pearl sense) were originally intended to carry out. As we will show, we think that many of FEP's proponents are using the blanket formalism in a much more metaphysically robust sense, one whose details cannot simply be assumed to follow from the formal properties of Markov blankets. Therefore, we propose to distinguish between 'Pearl blankets' to refer to the

standard 'epistemic' use of Markov blankets and 'Friston Blankets' to refer to this new 'metaphysical' construct. While Pearl blankets are unambiguously part of the map (i.e., the graphical model), Friston blankets are best understood as parts of the territory (i.e., the system being studied). We will now look in more detail at some of the philosophical claims about agent-environment boundaries that Friston blankets have been taken to support, following the claims made by Friston in "Life as we know it".

4.3 Friston blankets as agent-environment boundaries

Why and how have Markov blankets been reified to act as parts of the target system, e.g., by delineating its spatiotemporal boundaries, rather than merely formal tools intended for scientific representation and statistical analysis? When did the map become conflated with the territory? Here we aim to answer this question by presenting a series of different treatments inspired by Friston's use of Markov blankets in "Life as we know it". In doing so we can see how what was once an abstract mathematical construct used to describe conditional independence in graphical models came to be seen as a physical entity that somehow *causes* conditional independence. This latter interpretation has potentially interesting philosophical implications, but does not follow straightforwardly from the former mathematical construct. Perhaps surprisingly, many authors in the field are seemingly not aware of this process of reification, and this has led to the conflation of several different kinds of boundaries in the literature: Markov blankets are characterized alternatively as statistical boundaries, causal boundaries, spatial boundaries, epistemic boundaries, and autopoietic boundaries, and each characterisation is treated as somehow equivalent to (and interchangeable with) the others.

For instance, Allen and Friston (2018) write rather uncontroversially:

> The boundary (e.g., between internal and external states of the system) *can be described* as a Markov blanket. The blanket separates external (hidden) from the internal states of an organism, where the blanket per se can be divided into sensory (caused by external) and active (caused by internal) states. (p. 2474, our italics)

It is possible to read this passage in an entirely instrumentalist way. That the boundary 'can be described' using a blanket suggests the system can be *modeled* as having a blanket. This way of applying the Markov blanket is in line with the standard use of the notion introduced by Pearl and explained in the first part of this paper. On the other hand, this instrumentalist reading is put under pressure on the very next page:

> In short, the very existence of a system depends upon conserving its boundary, known technically as a Markov blanket, so that it remains distinguishable from its environment—into which it would otherwise dissipate. The computational 'function' of the organism is here fundamentally and inescapably bound up into the kind of living being the organism is, and the kinds of neighbourhoods it must inhabit. (p. 2475)

where the Markov blanket is exactly *equated with* the physical boundary in the world. Markov blankets here function to distinguish a system from its environment, much in the way a cell membrane does: the loss of a Markov blanket is equated with the loss of systemic integrity. This removes the distinction between the model of the system and the system itself. Map and territory have become indistinguishable, conflating what we are calling Friston blankets with the original Pearl blankets.

Other works seem to maintain a slightly more neutral perspective. Clark (2017), for example, carefully distinguishes between the causal process (the territory) and the Bayesian network (the map):

> Notice that the mere fact that some creature (a simple feed-forward robot, for example) is not engaging in active online prediction error minimization in no way renders the appeal to a Markov blanket unexplanatory with respect to that creature. *The discovery of a Markov blanket indicates the presence of some kind of boundary responsible for those statistical independencies.* The crucial thing to notice, however, is that those boundaries are often both malleable (over time) and multiple (at a given time), as we shall see. (p.4, our italics)

Here the discovery of a Markov blanket, perhaps only in our model of the system, serves to indicate the presence of a physical boundary in the system itself. Clark seems to hold that Markov blankets are discovered within the modelling domain, and that this discovery indicates the presence of something important ("some kind of boundary") in the target domain. While it is perhaps relatively unobjectionable, this move seems to presuppose a tight (and hence non-arbitrary) relation between the model and its target domain of an agent and its environment, with potentially crucial consequences for our understanding of cognitive systems (cf. Clark's previous work on 'cognitive extension' (Clark and Chalmers 1998)).

In a similar fashion, other works enforce the perspective that Markov blankets are a useful indicator to look for when attempting to define the boundaries of a system of interest. Kirchhoff et al. (2018), for example, write that:

> A Markov blanket defines the boundaries of a system (e.g., a cell or a multi-cellular organism) in a statistical sense.

They then go on to say, with much stronger implications, that

> [A] teleological (Bayesian) interpretation of dynamical behaviour in terms of optimization allows us to think about any system that possesses a Markov blanket as some rudimentary (or possibly sophisticated) 'agent' that is optimizing something; namely, the evidence for its own existence.

It is however never made explicit in the rest of their paper how to conceive specifically of a 'boundary in a statistical sense', perhaps indirectly relying on the inflated version of a Markov blanket proposed in (Friston and Ao 2012; Friston 2013b).

Hohwy (2017) also equates the internal states identified by the Markov blanket formalism with the agent:

> The free energy agent maps onto the Markov blanket in the following way. The internal, blanketed states constitute the model. The children of the model are the active states that drive action through prediction error minimization in active inference, and the sensory states are the parents of the model, driving inference. If the system minimizes free energy — or the long-term average prediction error — then the hidden causes beyond the blanket are inferred. (pp. 3-4)

For Hohwy, the Markov blanket is not just a statistical boundary, but also an epistemic one. Because the external states are conditionally independent from the internal states (given the Markov blanket), the agent needs to infer the value of the external states (the 'hidden causes') based upon the information it is receiving 'at' its Markov blanket, i.e., the sensory surface. Hohwy even goes as far as to define the philosophical position of epistemic internalism in terms of a Markov blanket:

> A better answer is provided by the notion of Markov blankets and self-evidencing through approximation to Bayesian inference. Here there is a principled distinction between the internal, known causes as they are inferred by the model and the external, hidden causes on the other side of the Markov blanket. This seems a clear way to define internalism as a view of the mind according to which perceptual and cognitive processing all happen within the internal model, or, equivalently, within the Markov blanket. This is then what non-internalist views must deny.

In other words, Markov blankets 'epistemically seal-off' agents from their environment. In the same paper, Hohwy, like Allen and Friston above, seems to equate an agent's physical boundary with the Markov Blanket:

> Crucially, self-evidencing means we can understand the formation of a well-evidenced model, in terms of the existence of its Markov blanket: if the Markov blanket breaks down, the model is destroyed (there literally ceases to be evidence for its existence), and the agent disappears. (p.4)

Finally, in a similar vein Ramstead, Badcock, and Friston (2018) characterize Markov blankets as at once statistical, epistemic, and systemic boundaries:

> Markov blankets establish a conditional independence between internal and external states that renders the inside open to the outside, but only in a conditional sense (i.e., the internal states only 'see' the external states through the 'veil' of the Markov blanket; [32,42]). [...] With these conditional independencies in place, we now have a well-defined (statistical) separation between the internal and external states of any system. A Markov blanket can be thought of as the surface of a cell, the states of our sensory epithelia, or carefully chosen nodes of the World Wide Web surrounding a particular province.

We can see now how Markov blankets have moved from a rather simple statistical tool used for specifying a particular structure of conditional independence within abstract random variables, to structures in the world that cause conditional independence, that separate an organism from its environment, and that epistemically seal off agents from their environment. These characterizations would sound bizarre to the average computer

scientist, about the only people aware of Markov blankets before 2012-2013, who are familiar only with the original 'Pearl blanket' formulation. In the next section we will consider the novel construct of a 'Friston blanket' in more detail, and highlight a number of additional assumptions that are necessary for Markov blankets to do the kind of philosophical work they have been proposed to do by the authors quoted above.

5 Friston blankets and sensorimotor loops in active inference

The more recent formulations of active inference, starting with (Friston and Ao 2012; Friston 2013b; Friston, Sengupta, and Auletta 2014), have effectively attempted to use Markov blankets as a tool to characterize a specific form of conditional independence in systems that can be understood as being composed of an agent and its environment (given the blanket, c.f. (Hipolito et al. 2020)). In particular, this use of Markov blankets assumes that the networks of interest can be meaningfully partitioned[12] into four distinct classes of variables, mapping to constructs usually stipulated for the purpose of defining sensorimotor loops, i.e., the action-reaction cycles between an organism and its ecological niche (see for instance (Tishby and Polani 2011; Ay and Zahedi 2014; Montúfar, Ghazi-Zahedi, and Ay 2015; Biehl 2017) for explicit connections to the Bayesian networks formalism). These four sets include, as highlighted in section 4.2 (and here repeated as a reminder): ϕ external states, μ internal states, a active states, and s sensory states.

Active inference assumes that the sequences of actions used to update observations form a closed loop with perceptual inference, such that any action taken will have a subsequent effect on perceptual inference, which in turn drives the generation of novel actions. When taken together with the partitioning of the system into internal/external and perceptual/active states, Friston's conceptualization of agent-environment systems under the FEP begins to resemble previous Bayesian treatments of sensorimotor loops. In sensorimotor theory, external states refer to world variables ϕ that generate observations s, sensed by a system whose internal states μ determine actions a that can affect the state of the environment in a causally circular closed loop (see the black arrows in Fig. 6). However, unlike other Bayesian treatments of sensorimotor loops, Friston and colleagues also assume a rather general set of connections (depicted in grey, see for instance (Friston 2013b; Friston 2019)), often eschewed by other authors (Tishby and Polani 2011; Ay and Zahedi 2014; Montúfar, Ghazi-Zahedi, and Ay 2015). These connections include bidirectional effects between sensors and actuators, ways in which sensors may influence external states, and ways in which actuators may influence internal states. Considering all the connections (black and grey alike) in Fig. 6 leads to the emergence of an 'interactional asymmetry' in the agent-environment coupling (Barandiaran, Di Paolo, and Rohde 2009), due to the lack of directed connections from internal states to sensors (meaning peripheral observations are not directly affected by the internal state of a system) and from external states to actuators

[12] At least at (nonequilibrium) steady-state, as this now appears to be one of the new assumptions packaged with the more explicit definition of a Friston blanket (Friston, Da Costa, and Parr 2020; Friston, Fagerholm, et al. 2020).

(meaning that states defined as external cannot directly affect actuators).[13] We will now turn to the novel role played by Markov blankets — or rather Friston blankets — in determining the sensorimotor boundaries of such systems, and highlight some ways in which this role differs from the role typically played by the more traditional Pearl blankets in Bayesian inference.

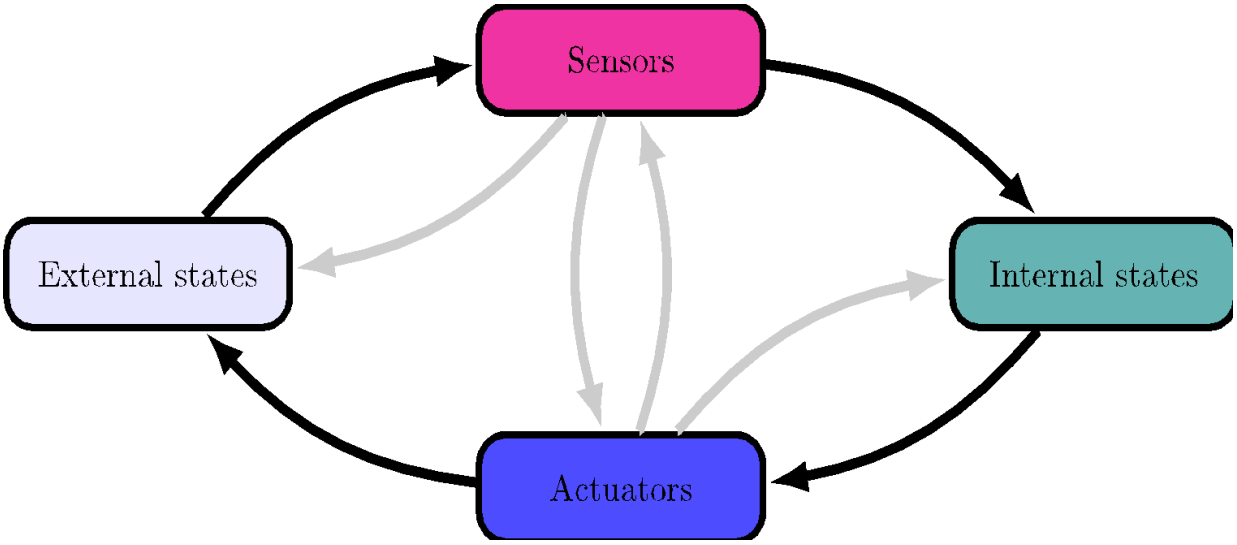

Figure 6: ***A sensorimotor loop.*** *A diagram representing possible dependences between different components of interest: sensors, internal states, actuators and external states. Notice that although this figure uses directed edges to signify causal influence (Ay and Zahedi 2014), it is not strictly a Bayesian graph, as it depicts cyclic sets of circular dependencies (some between pairs of components, and an overall loop including all components).*

5.1 Friston blankets as sensorimotor boundaries

To bring the novel role played by Friston blankets into full view we will first apply the four-way partitioning of random variables proposed by Friston and colleagues to the arbitrary Bayesian network that we introduced in Fig. 2. We hope that this schematic example will demonstrate that the partitioning cannot simply be applied to any graphical model without first making some additional assumptions. For instance, here we will label a node, say x_{10} from Fig. 2, as an 'internal' state (signified by a teal colour as in Fig. 6), which is conditionally separated from all the remaining 'external' variables (in lavender) by a set of nodes constituting its Friston blanket (see Fig. 7a). Following Friston's proposal, the nodes in this Friston blanket can be further separated into 'sensory' (magenta) and 'active' (blue) states, the former corresponding to the parents of the internal state (i.e., node x_{10}) and the latter including its children. Picking a different node, such x_9, to be labelled 'internal' generates a correspondingly different blanket (see Fig. 7b). It is clear here that the Friston blanket does

[13] Notably, the diagram in Fig. 6 is not a Bayesian network, nor is it intended to approximate one. However, even though the sensorimotor loop assumes cycles (due to both the bidirectional connections between different components and the overall circular causality imposed by the very definition of such a loop) it can be mapped to an acyclic directed graph by explicitly representing sets of nodes indexed by an appropriate temporal notation that removes the apparent cyclicality, see for instance (Ay and Zahedi 2014), and (Friston, Parr, and Vries 2017) more specifically for active inference.

not *define* what is inside and what is outside (or at least not without further assumptions), but can rather only be identified once we have already made this choice (by labelling one node as 'internal').

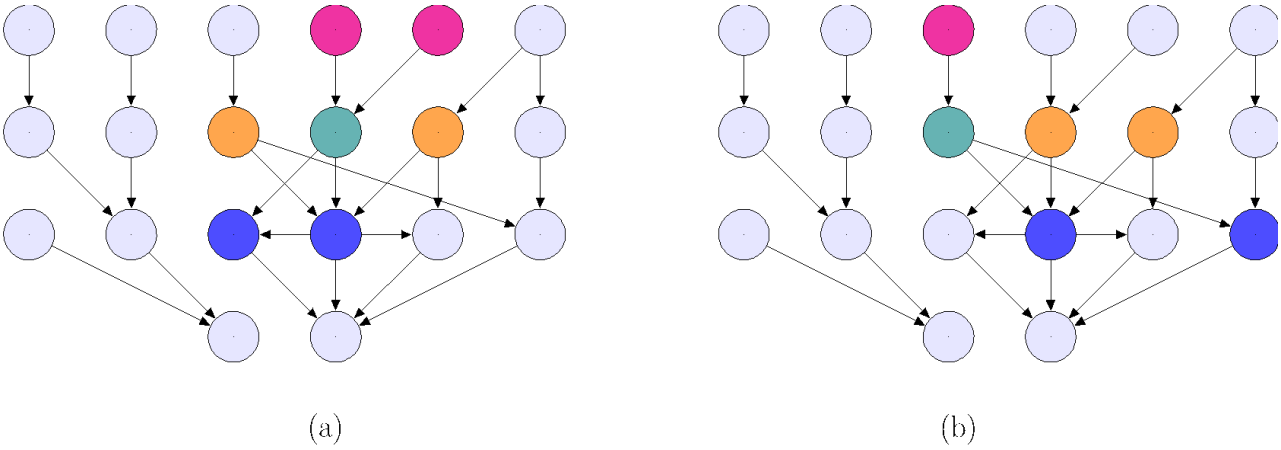

Figure 7: ***The Bayesian network described in Fig. 2, labelled with the Friston blanket notation****. Example sensorimotor blankets for different (internal) states, x_{10} and x_9 respectively, with labels removed for clarity. External states in lavender, sensory states in magenta, internal states in teal (x_{10} in (a) and x_9 in (b)), active states in blue, and putative co-parents in mustard. Notice that the partitions obtained here do not map to the separation of internal states (hidden states x Fig. 2) and world states (observations y Fig. 2). This suggests that sensorimotor blankets are Markov blankets applied under a specific set of assumptions that cannot be traced to standard uses of Markov blankets in variational inference (Jordan et al. 1999; Fox and Roberts 2012).*

The status of co-parents (labelled in mustard) in this model is also somewhat ambiguous, as while they can influence active states, they presumably should not be counted as 'internal' to the Fristonian agent defined by the sensorimotor boundary (Friston 2013b; Friston, Levin, et al. 2015; Hohwy and Michael 2017; Kirchhoff et al. 2018; Ramstead, Friston, and Hipólito 2020; Hipolito et al. 2020). On this account, Friston (2013b) writes that:

> [...] the Markov blanket can itself be partitioned into two sets that are, and *are not, children of external states*. We will refer to these as a surface or sensory states and *active states*, respectively. (p. 2, emphasis added)

This means that co-parents should be seen as sensory or active states depending on whether they are themselves dependent on external (i.e., not belonging to either blanket or internal) states, as suggested explicitly, for instance, in Figure 1. of (Ramstead, Friston, and Hipólito 2020) or Figure 1. of (Hipolito et al. 2020). However, in other works, co-parents are sometimes discussed implicitly, with Friston, Levin, et al. (2015) for example writing:

> External states cause sensory states that influence—but are not influenced by—internal states, *whereas internal states cause active states* that influence—but are not influenced by—external states [...] (p. 3, emphasis added)

or Kirchhoff et al. (2018) observing:

> The partitioning rule governing Markov blankets illustrates that external states—which are 'hidden' behind the Markov blanket—cause sensory states, which influence, but are not themselves influenced by, internal states, *while internal states*

> *cause active states*, which influence, but are not themselves influenced by, external states [7]. (emphasis added)

In other cases, their role is on the other hand largely ignored. Demekas, Parr, and Friston (2020) for example state

> The parents of internal states are the sensory states that mediate the influence of the outside world, and *their children are the active states* that mediate their influence on the outside world. (p. 2, our italics)

Kirchhoff and Kiverstein (2019) similarly write that

> The internal states of an agent can be shown to be formally equivalent to the internal states of the model [...]. *The children of the model can be mapped onto the active states that cause actions* [...]. The parents of the model are in turn a formalisation of the sensory states that influence the dynamics of internal states so as to further guide and inform action. (pp. 5-6, our italics, note that co-parents are not classified or even considered in this paper)

while Palacios et al. (2020) describe active states as

> Active states $f_A: S \times M \times \Omega \rightarrow^{f_A} A \in R^A$ states of action on the world (e.g., exocytosis of signalling molecules) that depend upon sensory and internal states. (Table 1, again there is no mention of coparents, which could also be seen as active states in the classification proposed by (Friston 2013b))

Another example comes from Hohwy and Michael (2017), who remark that

> The key is that active states are the downstream effects of what we have called deeply hidden endogenous states, and these endogenous states are the downstream effects of sensory states.

and

> It follows that the part of the model that is involved in active inference is the self: this part of the model (the active states and their more deeply hidden causes) are the very endogenous causes that can be inferred in perceptual inference, which therefore become part of the self-model that in turn, in a dynamic downstream manner, shape active inference.

In several of these works, co-parents are essentially glossed over in the definition of Friston blankets, perhaps in line with other work on Bayesian networks for sensorimotor loops, where no variables playing this role are usually mentioned (Tishby and Polani 2011; Ay and Zahedi 2014; Montúfar, Ghazi-Zahedi, and Ay 2015; Biehl 2017). However, this omission is not immediately obvious and often not stated explicitly (if active states are 'caused' only by primary internal ones, we can only assume that the authors mean there's no room for co-parents), and most importantly sheds light on an important formal difference between Friston blankets and Pearl blankets.

To bring this difference into full view, consider how the conditions which lead up to and modulate the patellar reflex (or knee-jerk reaction) could be illustrated using a Bayesian graph. This a common example of a mono-synaptic reflex arc in which a movement of the leg can be caused by mechanically stretching the quadriceps leg muscle by striking it with a small hammer. The stretch produces a sensory signal sent directly to motor neurons in the spinal cord which, in turn, produce an efferent signal that triggers a contraction of the quadriceps femoris muscle (or what is observed more familiarly as a jerking leg movement). If we project these conditions onto the left arm of our sample network Fig. 2, we get something like Fig. 8.

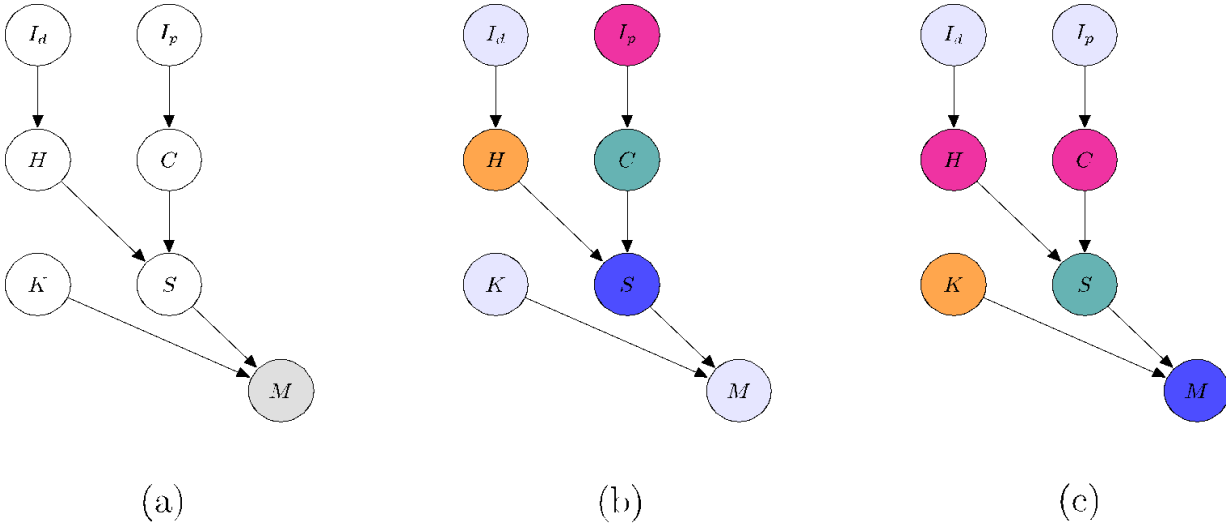

Figure 8: ***Conditions leading up to the knee-jerk reflex.*** *On the left, a Bayesian network where I_d and I_p denote the motor intentions of the doctor and the patient respectively. H indicates the medical intervention with a hammer, while C stands for the cortical motor command sent to the S which denotes the spinal neurons which are directly responsible for causing the kicking movement M. K stands for an independent way of moving the patient's leg, e.g., by someone else kicking it. In the middle and on the right, the same network partitioned using a 'naive' Friston blanket with different choices of internal states, C and S respectively.*

This simple network allows us to illustrate several problems with interpreting the co-parents in Friston blankets. Take S, i.e., the activation of the cortical motor neurons, as the node of interest. As the graph makes clear, the activation of these neurons can be *explained away* by either a strike of a medical hammer into the tendon (H) or a motor command from the central nervous system (C). This reflects the fact that the patellary reflex can also be modulated by the motor intentions of the patient. Under one possible interpretation of Friston blankets, the spinal signal which causes the movement would be an active state, meaning that the motor command C would be interpreted as an internal state of the patient. However, this leads to a puzzle about the way in which we should interpret H, which would then fall into the Friston blanket of C (see Fig. 8b), but stands for an external condition influencing the spinal neurons S. One could object that our example delineates internal states in the wrong way, and that S should be considered an internal state, as in Fig. 8c, while it is the bodily movement M that should be considered as an active state. Notice, however, that this would not help in any way, since there is always some possible external intervention K that could lead to the same kind of bodily movement, and has exactly the same formal properties as any putatively 'internal' cause of the movement. This example points back to

the problem of differentiating between effects produced by an agent (internal states) and those brought about by nodes not constitutive of an agent (co-parents). The state of a node is not simply the joint product of its co-parents, as completely separate causal chains (the doctor's intention vs. the patient's intention) can produce the same outcome (i.e., spinal neuron activation). Hence the partitioning of states into internal/external by means of a Markov blanket does not necessarily equate with the boundary between agent and environment found in sensorimotor loops. If Friston blankets are to serve the role of demarcating this boundary, they will require some additional assumptions that cannot simply be read off the original Markov blanket formalism (i.e., what we have been calling Pearl blankets).

A similar problem arises when we turn to the sense in which some states can be called 'active' or 'sensory', given that they may not uniquely map onto the internal and external states of the system (or the observed and unobserved variables in a graph) that we are interested in studying. This issue can be demonstrated by the fact that under the Friston blanket formulation, the location of a node in the graph layout is not sufficient to identify whether or not it is an 'internal' state in Friston's sense (recall that we had to start by arbitrarily selecting an internal state when formulating the graphs in Fig. 7). If we decide to interpret the active/sensory distinction by exclusively following Friston's use of the sensorimotor analogy, we will soon realise that using a Friston blanket as if it was a Pearl blanket falls short of explaining (and in some cases may be directly inconsistent with) the general identification of active and sensory states in agent-environment coupled systems. In particular, states described as 'sensory' are typically associated with observations made by an agent (Friston explicates their role by talking about 'the agent's sensations' (Friston 2013b)), but the formal definition of a Markov blanket does not guarantee that, given some arbitrary partition like in Fig. 2, these states will correspond with the observable variables in a Bayesian network. In our example neither of the 'sensory' states of the two Markov blankets overlap with the two nodes, y_1 and y_2, that constitute observations in the network. Somewhat paradoxically, a naive use of Friston's understanding of the Markov blanket formalism results in treating these observed variables as external states for the teal-coloured internal nodes used in our example (x_9 and x_{10}). Markov blankets provide the conceptual tools to deal with statistical (in)dependence and causal mediation, but under the Friston blanket application they are employed to account for epistemic mediation. This might seem intuitive for systems that we consider to be sentient or epistemic agents anyway (such as a cell or a larger organism), but becomes wildly implausible when applied more generally. Do the spinal neurons infer the doctor's intentions given the presence of the hammer? This would be a highly unusual kind of agent, and it seems like applying the Markov blanket formalism to this case stretches it far beyond its original purpose.

A further, and perhaps even more substantial, problem is that conditional independence is itself model-relative. One possible objection to the patellar reflex network presented above is that the conditions making up the graph are not fine grained enough, i.e., that the model is too simple. After all, the hammer does not directly intervene on the neurons in the spinal column, but rather on the tendon that causes the contraction of the muscle, which is responsible for the afferent signal that is the true proximal cause of the activation of the spinal motor neurons. However, just as it is difficult (and potentially ill-defined) to identify

the most proximate cause of the knee-jerk, it is difficult to identify *the* most proximate cause and consequence of any internal state. Since the very distinction between sensory and active states (the sensorimotor boundary) and external states (the rest of the world) hangs upon the distinction between 'most proximate cause' and 'causes further removed', the identifiability of such a cause is crucial. This point is well made by (Anderson 2017) who writes on the identifiability of the proximal cause:

> An obvious candidate answer would be that I have access only to the last link in the causal chain; the links prior are increasingly distal. But I do not believe that identifying our access with the cause most proximal to the brain can be made to work, here, because I don't see a way to avoid the path that leads to our access being restricted to the chemicals at the nearest synapse, or the ions at the last gate. There is always a cause even "closer" to the brain than the world next to the retina or fingertip.

A possible solution to this granularity problem would be to give up entirely on the idea of having some 'privileged' sense in which observations can be defined in the first place, i.e., each variable can become an observation, or a measurement, for any other variable (Friston 2019; Palacios et al. 2020). For example, the way in which some states count as 'sensory' could be understood in a way that has little to do with everyday agentic language, but rather is more metaphorical, analogous to the way that for a physicist, electrons might be 'observing' protons and 'acting' by rotating around a nucleus or jumping energy levels. This interpretation would imply that the distinction between sensory/active states should be understood along the lines of a 'causal' interpretation in which some states that are causally influenced by variables outside of the blanket are considered 'reactive', while active states are just those that exert causal influences on the variables external to the blanket. Although Friston and his collaborators have recently come to propose this non-literal interpretation (Friston 2019; Hipolito et al. 2020), it is important to note that interpreting their notion of a Markov blanket in this causal way does not solve all the ambiguities brought on by their refinements. It may even introduce new problems, especially for the study of cognitive agents, and is unlikely to be popular among early adopters of the approach who hoped for a more literal interpretation of Friston blankets as sensorimotor boundaries. If the formalism can be legitimately applied as widely as it is now suggested, it is no longer clear that it will have anything interesting to say about cognition in particular, as the 'literalists' might be hoping for (see (Baltieri, Buckley, and Bruineberg 2020) for a similar discussion).

While we do not want to try and solve all of these issues at this stage, we do think that they point to the importance of recognising that the notion of a Friston blanket as employed in the active inference literature is intended to carry out a very different role from the standard definition of a Pearl blanket used in the formal modelling literature. The open question here is whether Bayesian networks and Markov blankets are really the right kinds of conceptual tools to delineate the sensorimotor boundaries of agents and living organism, or whether there are really two different kinds of project going on here, each of which deserves its own set of formal tools and assumptions.

6 Two (very) different tools for two (very) different projects

So far, we have presented the conceptual journey on which Markov blankets have been taken. They started out as an auxiliary construct in the probabilistic inference literature (Pearl blankets), and have ended up at as a tool to distinguish agents from their environment (Friston blankets). The analysis above already showed the deep differences between Pearl blankets and Friston blankets, both in terms of their more technical assumptions and of the general aims of these two constructs. However, in the literature on the FEP and active inference, the two have not yet really been distinguished. Even in very recent work there is an obvious conflation of Pearl and Friston blankets, using the former to try and define, justify, or explain the latter. For example, see the figures presented in (Kirchhoff et al. 2018; Ramstead, Friston, and Hipólito 2020; Sims 2020) and (Hipolito et al. 2020), where Bayesian networks are used to describe what we would call Friston blankets. However, there are a series of extra assumptions that are necessary to move from Pearl blankets to Friston blankets, and these are rarely (if ever) explicitly stated or argued for. Some of these assumptions were implicitly touched upon in (Friston 2013b), where Friston blankets were defined by looking at the adjacency matrix of a set of particles simulated via a (random) dynamical system, which is *assumed* to be ergodic after it heuristically appeared to have reached a steady-state, and crucially, after arbitrarily *assuming* the number of clusters of particles that ought to be forming the 'internal states'. More recently, after Biehl, Pollock, and Kanai (2020) questioned some of the technical assumptions underlying the use of Markov blankets by Friston, the idea of Friston blankets being understood as a distinct construct has gained traction in the literature.

In a recent paper, Friston blankets are formalised in terms of constraints on sparse coupling of dynamics (or with arbitrary thresholds for non-sparse couplings), and identified via the non-zero components of the Hessian of the non-equilibrium steady state density represented in Langevin form using Ao's decomposition (Friston, Da Costa, and Parr 2020; Friston, Fagerholm, et al. 2020) — taking the construct far away from its Pearl blanket origins. Importantly, these points serve to highlight a pervasive confusion with the use of the Friston blanket construct as adopted so far (i.e., at least up until work such as (Friston, Da Costa, and Parr 2020)). For example, Kirchhoff and Kiverstein (2019) simply assume that the Markov blanket construct can be transposed from the formal to the physical domain, writing:

> The notion of a Markov blanket is taken from the literature on causal Bayesian networks (Pearl 1998). *Transposed to the realm of living systems, the Markov blanket allows* for a statistical partitioning of internal states (e.g., neuronal states) from external states (e.g., environmental states) via a third set of states: active and sensory states. The Markov blanket formalism can be used to define a boundary for living systems that both segregates internal from external states and couples them through active and sensory states. (p. 2, our italics)

Such a transposition is not at all straightforward, and the phrasing 'transposed to the realm of living systems' covers up a great explanatory leap from the merely formal Pearl blanket construct to the metahpysically-laden Friston blanket construct. It remains unclear what additional assumptions are being made in order to support the claim that the Markov blanket formalism can settle philosophically relevant questions. Another example that illustrates the

ambitiousness of the philosophical prospects of the Friston blanket construct is again provided by Kirchhoff and Kiverstein (2019):

> We employ the Markov blanket formalism to propose precise criteria for demarcating the boundaries of the mind that unlike other rival candidates for "marks of the cognitive" avoids begging the question in the extended mind debate. (p.1)

Based on what we have presented above however, the philosophical validity of using Friston blankets to draw sensorimotor boundaries cannot simply be assumed from the formal credibility of the original Pearl blanket construct. In order to evaluate the plausibility of using the Markov blanket formalism as a model of life and mind, we will now take a brief foray into some of the philosophical literature on models and modelling.

6.1 Models and modelling

As a general rule, one should not mistake the map described by a model for the territory it is describing: a model of the sun is not itself hot, a model of an organism is not itself alive, and so on. Scientific models ('maps') are typically understood as representations of some part of the world ('territory') that we can use to better understand something about that part of the world. No model is an entirely precise representation of its target system (if that were true then the model would cease to be any more useful than an exact replica of its target domain, reminiscent of Borge's one-to-one map of the world (Borges 1946)). It always involves some degree of abstraction or approximation, ideally in a way that draws attention to or otherwise clarifies some relevant aspect of what is being modelled. Moreover, a scientific model is always used by some researcher or research community for a specific purpose (Weisberg 2007, 2013), and only has meaning relative to that research context. The form a model takes is dependent on our own epistemic capacities. As we saw in section 3.1, Bayesian networks (and more generally all graphical models in the statistic and machine learning literature) are a visual way of representing probability distributions that makes conditional dependencies easily visible and intuitive. Nothing more, nothing less. The only formal difference between, e.g., equation (17) and Fig. 1, is the mode of presentation, i.e., they are two *maps* of the same *territory*. All of this is relatively uncontroversial, although of course there are many more fine-grained disputes about the exact nature and function of scientific models (Downes 2020).

Zooming back in on Bayesian networks (for example Fig. 2), the hidden states are scientific unobservables. There is a broader debate in philosophy of science about whether it is justified to believe in the reality of scientific unobservables ('realism') or whether they are auxiliary constructs helpful for explaining scientific observables ('instrumentalism'). Without going into this debate in detail, we think that the answer to this question is partially dependent on the modeling tools used. So what about Bayesian networks? In general, Bayesian networks are not tied to a particular level of abstraction: their power lies in the fact that they remain agnostic about the relationship between the random variables they represent. This method works well for complex phenomena in the medical and social

sciences (see for instance (Pourret, Naïm, and Marcot 2008)), where no clear causal pathways are available and multiple levels need to be integrated into one model. In the knee-jerk example introduced above Fig. 8, we drew a direct arrow between the doctor's intention and the hammer hitting the patient's knee, disguising lots of (for our purposes) unnecessary fine-grained detail.

What then decides what makes a *good* Bayesian network? The dominant assumption in the literature is that the best model is one that accounts for the data in the most parsimonious way ((Stephan et al. 2009; Friston, FitzGerald, et al. 2017)). This intuition can be formalised via a process of model comparison, using different criteria, for example the Akaike information criterion (AIC), the Bayesian information criterion (BIC), or, as we presented in Section 1, variational free energy (via the maximisation of model evidence, equivalent to the minimisation of surprisal). In the case of variational free energy, one can then take into account a trade-off between the complexity of a model and the accuracy with which it is able to predict the data, or observations. When minimizing free energy using a range of different models, the one with the lowest free energy is thus taken to be the one that accounts for the data in the most parsimonious way (cf. the Occam factor (MacKay 2003; Bishop 2006; Friston 2010; Daunizeau 2017).

This means that the graphical model itself is already an abstraction, reflecting (at least partially) a choice by the modeller about which observations to include. When used to describe neural behaviour, such models will typically cluster groups of neurons into single nodes, but even if each node represented a single neuron it could still be further decomposed, revealing the internal structure of each neuron, and so on (Friston, Fagerholm, et al. 2020; Hipolito et al. 2020). However, as long as the data does not necessitate such complexity, a model in which groups of neurons are clustered will be selected. In other words, the epistemic aim (even for the models used in the context of active inference) is not to arrive at a complete model of the world, but rather to obtain the most parsimonious model that captures the relevant relations (Baltieri and Buckley 2019).

What does this imply for the philosophical prospects of the Friston blanket construct serving as a sensorimotor boundary? Simply put, where Friston blankets are located in a model depends (at least partially) on modeling choices, i.e., Friston blankets cannot simply be 'detected' in some objective way and then used to determine the boundary of a system. This can be easily seen by the fact that Markov blankets are defined *only* in relation to a set of conditional (in)dependencies, or the equivalent graphical models (in either static systems (Pearl 1988), or dynamic regimes at steady-state (Friston, Da Costa, and Parr 2020)). The choice of a particular graphical model is then usually enforced by Bayesian model selection, which is in turn dependent on the data used (e.g., one cannot hope to model the firing activity of neurons, given as data fMRI recordings that already measure only at the grain of voxels). These considerations point, in our opinion, to a strongly *instrumentalist* understanding of Bayesian networks, and hence of Markov blankets, which would not justify the kinds of strong philosophical conclusions drawn by some from the idea of a Friston blanket (see e.g., (Hohwy 2016; Friston, Wiese, and Hobson 2020)).

There is a second conceptual issue latent in the current discussion. We started our paper with the parallel between perceptual inference and scientific inference. Both use a

previously learned model and a set of observations to infer the causal structure of the unobserved outside world. This parallel puts (model-based) cognitive neuroscience in a rather special place: it makes models of how animals model their environment. A cognitive neuroscientist uses both behavioral and neural data to infer the most likely model that the agent's brain implements. For example, Parr et al. (2019) use both MEG and eye-tracking to disambiguate a number of causal models for active vision. These putative models correspond in a fairly straightforward way to a neural network and make concrete predictions about both neural dynamics as well as oculomotor behavior. By scoring these models based on their accuracy in predicting neural dynamics and oculomotor behavior, weighted by the complexity of those models the most 'likely' causal model is selected (i.e., the one that best explain the data in the most parsimonious way). In other words, the agent implements a causal model of its environment, and the scientist uses *another* causal model to infer which particular model the agent implements.

There is nothing wrong with this doubling up of modeling relations as long as one is conceptually careful: one needs to distinguish between properties of the environment, properties of the agent's model of the environment and properties of the scientist's model of the agent modelling its environment. Considering these different modeling relations provides a new lens to analyse the different between Pearl and Friston blankets: Pearl blankets are boundaries drawn on the scientist's map of the agent-environment system (in the form of a Bayesian network). The question is whether Friston blankets are similarly drawn on the scientist's map or whether they are boundaries in the agent-environment system itself. The former option is rather uncontroversial: it makes Friston blankets (with the exception of the issue of the co-parents of children, and a few more technical constraints highlighted in (Friston, Da Costa, and Parr 2020)) closely akin to Pearl blankets, but unlikely to be of much philosophical interest (at least when it comes to the question of boundaries between agents and environments). The latter option might do interesting philosophical work, but requires a number of substantial metaphysical commitments, and cannot simply be assumed to follow from the previous success and formal validity of Pearl blankets. We will further describe how these two projects could be developed in the next section.

6.2 Inference *with* a model and inference *within* a model

If Pearl blankets play a fundamentally instrumental role in assigning probabilities to different outcomes, based on different modelling choices, spatial and temporal coarse-grainings, etc. can the same be said about Friston blankets? In an ambitious projects started perhaps with (Friston 2013b) and currently under development with recent works such as (Friston 2019; Friston, Wiese, and Hobson 2020), what seems to emerge is a desire to use Friston blankets as the basis for a philosophical distinction between agent and environment, other core constructs in the life and the social sciences (Ramstead, Badcock, and Friston 2018; Veissière et al. 2020), and perhaps even a metaphysical characterization of what it is to be *any* kind of system (Friston 2019). This suggests that Friston blankets are to be understood as something more than a merely instrumental statistical construct, i.e., as an actual *thing* out there in the world that can be identified and theorised about. We would like now to highlight what we take to be one of the fundamental differences between the notions

of Friston and Pearl blankets. In the previous sections we drew attention to the fact that the notion of a Friston blanket rests on the assumption that there is some kind of agent at the centre of the blanket (i.e., the internal states ought to constitute an agent, see e.g., (Friston 2013b; Hohwy 2016; Kirchhoff et al. 2018)), and that this agent is separated by some kind of (statistical) boundary from the rest of its environment (i.e., external states of the graphical model are now equated to a real environment). As we have seen, this talk in terms of agents and environment is not present in, and thus not justified by, the original notion of a Pearl blanket. The distinction between the two blanket constructs (Pearl and Friston) can then be easily identified once we look at *who* appears to be performing the inference and *what* system that inference is performed on, when each kind of blanket is deployed. To do so, however, it is important that we first distinguish among potentially four kinds of (Bayesian) networks that are sometimes used in the literature to describe processes of (approximate) inference:

1. a generative process $p_{GP}(y, x_{GP})$ capturing the actual causal structure of the environment where hidden states x generate observations y.
2. a generative model $p_{GM}(y, x_{GM})$ representing our best (epistemic) understanding x_{GM} of how some given data/observations y are generated from real hidden states x_{GP} for a given environment.
3. a posterior density $p_{GM}(x_{GM} \mid y)$ encoding the inversion scheme on a generative model, i.e., the most likely state of the environment given the observations, using either exact or approximate methods.
4. a variational or recognition density $q(x_{GM})$ (for variational inference schemes) used to determine the best approximation of the (usually uncomputable) posterior $p_{GM}(x_{GM} \mid y)$ given a series of constraints on $q(x)$.

These four kinds of network can then be combined in different ways, leading to two quite different kinds of research program, which we refer to as, respectively, 'inference *with* a model' and 'inference *within* a model'. We will now discuss each in turn.

6.2.1 Inference *with* a model

As mentioned in the section above on models, the reason why model-based inference is used in science is because the causal structure of the world is not directly given to us. Typically, we want to know the state of some aspect of the world x_{GP} while only having access to some observations y. For example, using fMRI we can observe the change in the magnetic field surrounding the head due to blood-oxygen-level dependent (BOLD) contrast, and based on this activity wish to study some cognitive processes. Applying our terminology to this example, we can say that unobservable cognitive processes x_{GP} cause (in an indirect and complex way) observable changes in the magnetic field y. The process (1) $p_{GP}(y, x_{GP})$ by which the environment (including cognitive processes) generates observations is assumed to exist, but is beyond our reach. Our current assumptions about this generative process are represented in the generative model (2). The network drawn in Fig. 2 could be a schematic version of such a model. One can think of node y_2 as the observed magnetic field and think of node x_{10} as representing the cognitive process we want to investigate. The cognitive

process is then intermeshed in a causal web that ultimately generates our observations, and thus allows us to build a model inferring its structure.

It is important to note that in order to to effectively perform inference on this network, we would have to draw a second network that reverses the information flow to represent the calculation of a posterior $p_{GM}(x_{GM} \mid y) = p(y \mid x_{GM})p(x_{GM})/p(y)$ as the most likely explanation of a given set of observations (see for example Figure 8.37 and following figures in (Bishop 2006)). In such a case, we would talk of the posterior $p_{GM}(x_{GM} \mid y)$ (3) as an 'inverse' model, in the sense that it is derived from an inversion of the generative model (2). Finally, as we already explained in the first sections of this manuscript, should this inference problem prove to be too complex for calculating the posterior directly, we could instead use an approximation by introducing a variational or recognition density $q(x_{GM})$. In this case (4), yet another Bayesian network, could be drawn to describe $q(x)$, e.g., using the mean-field assumption as described in section 2.2, so that methods like variational message passing (Bishop 2006) could be used to approximate the posterior $p_{GM}(x_{GM} \mid y)$.

In all of the above cases, Pearl blankets simply capture relations of conditional independence between variables in the model, regardless of the different roles that random variables play in each network. In other words, the above cases say nothing about what any postulated Pearl blankets might correspond to in the external world, especially in scenarios where the generative model and the generative process are not assumed to be identical. In this sense, Pearl blankets are simply a tool for the experimenter, who sits *outside* of some system of interest, trying to peer into that system by performing inference *with* a (perhaps graphical) model. Generative models should here be seen as epistemic tools to formulate predictions about the world that are useful to the experimenter, rather than ontological statements about some objective truth to be found out there in the world. Nobody using Pearl blankets to assist them in performing variational Bayesian inference believes that those blankets correspond to literal boundaries in the systems that they are studying.

As mentioned above, an important motivation for the free-energy principle is the parallel between scientific inference and perceptual inference. Like the scientist, the agent wants to know and control the state of some aspect of the world (1) x_{GP} while only having access to some observations y. The agent can solve this problem via a generative model (2) of its environment. The agent uses (or appears to use) variational inference to obtain a recognition density (4) which approximates the posterior density (3). Like in scientific inference, Pearl blankets might appear in the agent's own model, but again they would be a tool used by the agent, not a literal feature of either the agent or its environment (or indeed, the boundary between the two).

In model-based cognitive neuroscience, the two approaches are stacked together. The explanatory project is to infer what generative model an agent is using to infer the states of its environment. This, to us, seems to be one of the strongest empirical applications of the FEP, as can be seen in the influential work of the likes of (Parr et al. 2019; Adams et al. 2013; Pezzulo, Rigoli, and Friston 2018) and reflects a more general explanatory strategy in cognitive neuroscience (Lee and Mumford 2003). As an instrumental modeling strategy, we take no issue with treating the brain as functioning analogous to a scientist, but as a more realist claim it has a number of problems. Most notably, the FEP denies the distinction

between scientist and model: an agent does not *have* a model of its environment that it uses using to perform inference, but rather an agent *is* a model of its environment (Friston 2013a; Bruineberg, Kiverstein, and Rietveld 2018; Friston 2019; Baltieri and Buckley 2019). There is no separate entity that uses a generative model to perform inference, instead the agent performs (or appears to perform) inference, and it is at once scientist *and* model. It might be that for this reason some theorists have turned away from 'inference with a model' towards a different (and perhaps even more ambitious) explanatory project, which we will call 'inference within a model'.

6.2.2 Inference *within* a model.

The 'primordial soup simulation' that we presented in 4.2 presents a very different research direction for the FEP and active inference framework. This simulation starts out with a soup of coupled particles and aims to show how a distinction between 'agent' and 'environment' emerges naturally as the dynamics of the system reach equilibrium. Here we will use the example of Fig. 7, where the lavender nodes represent external states, coupled to a set of internal nodes which, under the Friston blanket interpretation, is claimed to imply that the system represented by these internal nodes is an *agent* performing inference on the external states (Friston 2013b; Friston 2019). Effectively, this Bayesian network integrates and characterises different processes that are usually (graphically) represented separately: (1) a generative process $p_{GP}(y, x_{GP})$ in the form of external states producing sensory states, and (4) a variational density $q(x)$[14] or (3) an exact posterior (in the first part of (Friston 2019)) encoded in a set of internal states that 'use' sensory information to 'produce' actions affecting the environment in the future. Because of the presence of both these blocks within the same network, one gets to describe both data generation (the evolution of the generative process $p_{GP}(y, x_{GP})$ as a stochastic process), and the (active) inversion scheme (based on the approximate posterior $q_{GM}(x_{GM})$ and the recognition dynamics that optimise its sufficient statistics (Friston 2013b; Friston 2019)) in a single graphical model, where the system performing inference is postulated given a set of initial assumptions (weakly mixing random dynamical systems, sparse dynamical coupling, and the ensuing Friston blanket, nonequilibrium steady-state) and explicitly drawn within a graphical Bayesian network (Friston 2019; Friston, Da Costa, and Parr 2020).

Here the constructed Friston blanket is not one that describes which variables are probabilistically shielded from which other variables, for example in the computation of mean-field averages for an internal node, but is rather posited as the statistical boundary that conditionally separates an assumed or postulated set of internal states (via sparsity constraints or arbitrary thresholding) from another set of states postulated or assumed to be external to the 'agent'. These two sets of states are then interpreted as an agent-environment coupled system given an arbitrary 'context-dependent' partition that determines, for example, how the Friston blanket for a whole cell characterises something qualitatively different from the Friston blanket for half of the same cell. Thus, the main

[14] Under a mean-field formulation (Bishop 2006), a variational Gaussian approximation (Opper and Archambeau 2009; Friston 2013b) or a mix of the two (Friston, Trujillo-Barreto, and Daunizeau 2008))

feature of Friston blankets, and what differentiates them from Pearl blankets, is that they are not simply used by a scientist to *perform* inference but rather *explain* inference itself, by explaining the existence of a system distinct from its environment (i.e., an agent). In other words, the explanatory project here is to start with a Bayesian network or a random dynamical system and ask questions like: 'under what conditions does one part of the model (the agent) come to infer the states of another part of the model (the environment)'.

It should be clear now that the philosophical bounty here is potentially large. The project of 'inference within a model' is to define in mathematical terms *what it is* to be a system (Ramstead, Badcock, and Friston 2018), *where* to draw the boundary between agent and environment (Kirchhoff and Kiverstein 2019), *what it is* to be a sentient and conscious being (Friston, Wiese, and Hobson 2020), and *what is required* for an agent to have a representation (Ramstead, Friston, and Hipólito 2020), all hotly contested philosophical questions. Clearly, researchers working with the FEP tradition want to draw metaphysical conclusions out of the Friston blanket construct. But metaphysical consequences require metaphysical premises, and cannot simply be read off the formal model itself (i.e., from previous work on Pearl blankets). One obvious consideration here is that 'inference within a model' is performed on an idealized mathematical structure, either a random dynamical systems or a Bayesian network, not the physical world itself. The question is then whether the mathematical structures posited by the FEP are merely a map of self-organizing systems (in which case the non-metaphysical Pearl blanket construct), or are themselves the territory. In the latter case the FEP framework might constitute something like an 'information ontology', perhaps an appealing picture for some but certainly not something that comes without any further metaphysical commitments. Menary and Gillett (2020) suggest something like this when they write "Our point here in drawing these connections is to highlight the strong Platonist and Pythagorean metaphysical attitudes that are implicit in Ramstead and colleagues' formal ontology approach" (p. 24). Such an approach could be valid and interesting, but it would certainly not be metaphysically innocent!

Another route to go would be to see the 'primordial soup' simulation as a mathematical formalization demonstrating the emergence of a sensorimotor boundary (Friston blanket) in a highly idealized domain. In a similar vein, concepts in theoretical biology have been formalized in the idealized domain of the Game of Life (Beer 2004, 2014, 2020). This might be an interesting way of modelling emergent processes in complex systems, but it would not support any metaphysical claims about Friston blankets. We will not pursue this idea any further here, but offer it as a more modest, perhaps 'instrumental' interpretation that some proponents of the FEP and active inference might be inclined to adopt in order to avoid any stronger commitments.

Perhaps the clearest expression of the metaphysical commitments implied by the use of Friston blankets is provided by Ramstead et al. (2019), who write:

> The claims we are making about the boundaries of cognitive systems are ontological. We are using a mathematical formalism to answer questions that are traditionally those of the discipline of ontology, but crucially, we are not deciding any of the ontological questions in an a priori manner. The Markov blankets are a result of the system's dynamics. In a sense, we are letting the biological systems

> carve out their own boundaries in applying this formalism. Hence, we are endorsing a dynamic and self-organising ontology of systemic boundaries. (p. 3)

where the claim seems to be that the answers to these ontological questions can be simply assumed by 'doing the maths' and then checking where the Markov blanket lies. If by Markov blanket they mean here the traditional *Pearl* blanket, then something like this might be possible, but it will not have the desired ontological consequences. If, however, they mean *Friston* blanket (and we assume that they do), then the ontological consequences might follow, but not without further metaphysical premises. This is this is the dilemma faced by the proponent of Friston blankets within the active inference framework. Later in the same paper, they write:

> By placing our Markov blanket around Homo sapiens, we necessarily encapsulate all of the dynamic, lower-level processes responsible for producing every phenotype, while imposing a clear upper limit on the complex adaptive system under scrutiny. Although the human Markov blanket is nested within the broader dynamics of other global Markov blankets that extend out into the universe, these lie beyond the limits of the system that this ecobiopsychosocial framework endeavours to explain. (p. 13)

Here the Markov blankets are 'placed' instead of being a 'result of the system's dynamics', and can indeed be placed at a multitude of different points, resulting in a nested hierarchy of blankets from the smallest cell out into the widest reaches of the universe. The picture is one of 'Markov blankets all the way down', but if this is the case then the boundaries demarcated by such blankets can no longer do any interesting work. If blankets can be 'placed' so as to cut any of the joints of any modeled system, then they are effectively nothing more than a purely instrumental construct, useful perhaps for studying these systems, but not to be understood as anything 'real' out there in the world. To be clear, we think that Markov blankets understood in this instrumental Pearl blanket sense are extremely *valuable* tools, we just don't think that they license any of the metaphysical claims made by those using Friston blankets to demarcate the boundaries of systems.

In sum, the main difference between the concepts of Pearl blankets and Friston blankets is that while the former describes a property of statistical models that can be used for different purposes ((1) to (4) above), the latter is a particular interpretation of that property for the purpose of studying agent-environment systems (broadly construed, but based on drawing networks where a generative model (2) precisely equates with the generative process (1) and a second entity, statistically separated by a Friston blanket, is drawn that performs exact (3) or approximate (4) inference *within* the model). Any interpretation of a classical Pearl blanket beyond the statistical one depends on the researchers' goals, interests, and metaphysical assumptions. In the case of Friston blankets, the interpretation is already in some important ways fixed, because the very concept of a Friston blanket depends on the assumption that the system of interest is an agent who is in some way bounded from its external milieu, and performs activities that can be conceptualised as inferences about the state of that milieu. Classical Pearl blankets are formal tools that are used to make inferences about some system, using a model of that system, while Friston blankets assign a sensorimotor interpretation of the model and assume that the system of interest is itself

performing inferences. Conflating these two notions, or assuming that the latter follows uncontroversially from the former, can too easily lead one to draw some unwarranted philosophical conclusions, and for this reason we encourage caution and conceptual hygiene when using Markov blankets of either kind.

7 Conclusion

The free energy principle and active inference framework have recently gained traction in the fields of neuroscience and biology due to their ambitious claims regarding a definition of a unifying principle that ought to characterise living and cognitive systems, and their functions and behaviours (Friston 2010; Friston, FitzGerald, et al. 2017; Hesp et al. 2019; Friston 2019; Kuchling et al. 2020). Under the umbrella term of predictive processing, they have also gained popularity in philosophy of mind and cognitive science, where they appear to play the role of a new thinking tool that could settle centuries-long disputes in the most disparate areas of mind and life (Clark 2013, 2015, 2020; Hohwy 2013; Friston, Wiese, and Hobson 2020). At the same time, different parts of the FEP and its associated process theories, in the form of prediction error minimization, hierarchical predictive coding or active inference, have raised some important, and in some cases yet-to-be-answered, scientific and philosophical questions. Some of them have to do with the capacity of the framework to account for traditional folk psychological distinctions between belief and desire (see e.g., (Klein 2018; Yon, Heyes, and Press 2020)), although its defenders have argued that it can either account for desire in a novel way (Wilkinson et al. 2019), or that it is a mistake to expect neuroscientific theories to account for folk psychological constructs at all (Dewhurst 2017). Another, very common, kind of critique is that the framework either does not, or cannot even in principle, enjoy any empirical support, and should at best be considered a theoretical redescription of our existing data (see e.g., (Colombo, Elkin, and Hartmann 2018; Liwtin and Miłkowski 2020; Cao 2020)). Yet another kind of critique argues that there is no significant connection between the (a priori) FEP formalism on the one hand, and the (empirical) process theories it is intended to support on the other ((Colombo and Wright 2018; Williams 2020), or that it presents a false equivocation between probability and adaptive value (Colombo 2020). Finally, Andrews (2020) and van Es (2020) have recently argued against a realist interpretation of the mathematical models described by free energy principle, which are claimed to be better interpreted instrumentally. Along the same lines, Baltieri, Buckley, and Bruineberg (2020) provided a worked-out example of this instrumentalist view, where an engine coupled to a Watt (centrifugal) governor is shown to perform active inference as an example of 'pan-(active-)inferentionalism', asking what can possibly be gained by thinking of the behaviour of a coupled engine-mechanical governor system in terms of perception-action loops under the banner of free energy minimisation.

These last three works come closest, at least in spirit, to the topics discussed in this paper, which have to do with a disconnect between the formal properties of Markov blankets and the way these are deployed in the arguments used to support the metaphysical claims made by free energy principle. More specifically, in this paper we have addressed some possible concerns regarding the use of Markov blankets (Pearl 1988) within the free energy principle.

After having been initially adopted in the context of (variational) inference problems, as a tool to simplify the calculations of approximate posteriors by taking advantage of relations of conditional independence (Bishop 2006; Murphy 2012), in the context of the free energy principle these tools have ultimately been used to perform work that claims to clarify boundaries of the mind (Hohwy 2017; Clark 2017; Kirchhoff and Kiverstein 2019), of living (Friston 2013b; Kirchhoff 2018; Kirchhoff et al. 2018) and even social systems (Ramstead, Badcock, and Friston 2018; Veissière et al. 2020). What is interesting here is that mere (statistical) divisions made within a Bayesian network representing relations of independence of a generative model define what it is to be a system. In other words, the Bayesian network takes precedence over the physical world that it is supposed to model. In some passages it even appears that the world is taken *to be* a Bayesian network, with the Markov blankets defining what it is to be a 'thing' (Friston 2013b; Kirchhoff et al. 2018; Friston 2019; Hipolito et al. 2020). This then brought us to some possible issues, namely the question of whether Bayesian networks are merely an instrumental modelling tool for the free energy principle framework and consequent adoptions in cognitive science and philosophy of mind, or whether the framework presupposes some kind of more fundamental Bayesian graphical ontology.

As we mentioned in section 6, all of this points towards a fundamental dilemma for anyone wanting to use Markov blankets to make substantial philosophical claims about biological and cognitive systems, which is what we take proponents of free energy principle to be wanting to do. On the one hand, Markov blankets can be used instrumentally in their original Pearl blanket guise, as a formal mathematical construct for inference on a generative model, for example in the form of a Bayesian network. This usage is philosophically innocent, but cannot, without further assumptions that need to be stated explicitly, justify the kinds of conclusions that it is sometimes used for in the literature based on the FEP (Hohwy 2017; Kirchhoff et al. 2018; Kirchhoff and Kiverstein 2019). On the other hand, Markov blankets can be used in a more realist fashion, which we have called Friston blankets, as an ontological construct demarcating actual boundaries 'out there in the world', such that inferential processes also become also something to be see seen 'out there in the (physical) world' (Friston 2019). This is surely a more exciting application of the Markov blanket formalism, but it cannot be simply or innocently 'read off the mathematics' of the more standard usage advocated in statistics and machine learning (Pearl 1988), and requires some additional technical (Friston 2019; Biehl, Pollock, and Kanai 2020; Friston, Da Costa, and Parr 2020) and philosophical (Ramstead, Badcock, and Friston 2018; Friston, Wiese, and Hobson 2020; Hipolito et al. 2020) assumptions, that may in the end be doing all of the interesting work themselves.

The difference between inference *with* and inference *within* a model, here roughly corresponding to the use of Pearl and Friston blankets, shows why the potential payoff of the latter construct is much larger than the former. In inference *with* a model, the graphical model is an *epistemic tool* for a scientist to perform inference. In inference *within* a model the scientist disappears from the scene, becoming a mere spectator of the unravelling inference show before their eyes. Here the (Friston) blanket specifies the anatomy of inference: it is a formalization of what it is to be a cognising living system, and defines the boundary between this system and its environment.

Ultimately, the considerations presented in this paper leaves the free energy theorist with a dilemma. One can accept a rather innocent conception of Markov blankets that merely licenses an instrumental interpretation, one under which any system can be treated 'as-if' it had a blanket, and which is admittedly scientifically useful but that has not lead to much more philosophically interesting conclusions so far; or, one can import a number of stronger metaphysical assumptions about the mathematical structure of reality to support a realist reading where the blanket becomes a literal boundary between agent and world. At any rate, such a strong realist reading cannot be justified by 'just following from the mathematics', but needs to be independently argued for, and such an argument has not yet been offered.

Acknowledgements

The authors would like to thank Mel Andrews, Martin Biehl, Daniel Dennett, Richard Menary, Fernando Rosas, Filippo Torresan, Nina Poth and other members of Tobias Schlicht's research group for insightful discussions and timely feedback on previous versions of the manuscript. MB is a JSPS International Research Fellow supported by a JSPS Grant-in-Aid for Scientific Research (No. 19F19809). KD's work is funded by the Volkswagen Stiftung grant no. 87 105.

References:

Adams, Rick A, Klaas Stephan, Harriet Brown, Christopher Frith, and Karl J Friston. 2013. "The Computational Anatomy of Psychosis." *Frontiers in Psychiatry* 4: 47.

Allen, Micah, and Karl J Friston. 2018. "From Cognitivism to Autopoiesis: Towards a Computational Framework for the Embodied Mind." *Synthese* 195 (6): 2459–82.

Anderson, Michael L. 2017. "Of Bayes and Bullets: An Embodied, Situated, Targeting-Based Account of Predictive Processing." In *In Philosophy and Predictive Processing: 3*, edited by Wanja Wiese and Thomas K Metzinger, 60–73. Frankfurt am Main, Germany: MIND Group.

Attias, Hagai. 2003. "Planning by Probabilistic Inference." In *AISTATS*. Citeseer.

Andrews, Mel. 2020. *The Math is not the Territory: Navigating the Free Energy Principle.* [Preprint] URL: http://philsci-archive.pitt.edu/id/eprint/18315 (accessed 2020-11-30).

Ay, Nihat, and Keyan Zahedi. 2014. "On the Causal Structure of the Sensorimotor Loop." In *Guided Self-Organization: Inception*, 261–94. Springer.

Baltieri, Manuel, and Christopher L. Buckley. 2019. "Generative Models as Parsimonious Descriptions of Sensorimotor Loops." *Behavioral and Brain Sciences* 42: e218.

Baltieri, Manuel, Christopher L Buckley, and Jelle Bruineberg. 2020. "Predictions in the Eye of the Beholder: An Active Inference Account of Watt Governors." *arXiv Preprint arXiv:2006.11495*.

Barandiaran, Xabier E, Ezequiel Alejandro Di Paolo, and Marieke Rohde. 2009. "Defining Agency: Individuality, Normativity, Asymmetry, and Spatio-Temporality in Action." *Adaptive Behavior* 17 (5): 367–86.

Beal, Matthew J. 2003. *Variational Algorithms for Approximate Bayesian Inference*. University of London London.

Beer, Randall D. 2004. "Autopoiesis and Cognition in the Game of Life." *Artificial Life* 10 (3): 309–26.

———. 2014. "The Cognitive Domain of a Glider in the Game of Life." *Artificial Life* 20 (2): 183–206.

———. 2020. "An Investigation into the Origin of Autopoiesis." *Artificial Life* 26 (1): 5–22.

Berger, James O. 2013. *Statistical Decision Theory and Bayesian Analysis*. Springer Science & Business Media.

Biehl, Martin. 2017. "Formal Approaches to a Definition of Agents." *arXiv Preprint arXiv:1704.02716*.

Biehl, Martin, Christian Guckelsberger, Christoph Salge, Simón C. Smith, and Daniel Polani. 2018. "Expanding the Active Inference Landscape: More Intrinsic Motivations in the Perception-Action Loop." *Frontiers in Neurorobotics* 12: 45.

Biehl, Martin, Felix A Pollock, and Ryota Kanai. 2020. "A Technical Critique of the Free Energy Principle as Presented in" Life as We Know It" and Related Works." *arXiv*, arXiv–2001.

Bishop, Christopher M. 2006. *Pattern Recognition and Machine Learning*. Springer-Verlag New York.

Blei, David M, Alp Kucukelbir, and Jon D McAuliffe. 2017. "Variational Inference: A Review for Statisticians." *Journal of the American Statistical Association* 112 (518): 859–77.

Bogacz, Rafal. 2017. "A Tutorial on the Free-Energy Framework for Modelling Perception and Learning." *Journal of Mathematical Psychology* 76: 198–211.

Borges, JL. 1946. “On Exactitude in Science (a. Hurley, Trans.).” *Collected Fictions. New York: Viking Penguin*.
Bruineberg, Jelle, Julian Kiverstein, and Erik Rietveld. 2018. “The Anticipating Brain Is Not a Scientist: The Free-Energy Principle from an Ecological-Enactive Perspective.” *Synthese* 195 (6): 2417–44.
Buckley, Christopher L, Chang Sub Kim, Simon McGregor, and Anil K Seth. 2017. “The Free Energy Principle for Action and Perception: A Mathematical Review.” *Journal of Mathematical Psychology* 14: 55–79.
Cao, Rosa. 2020. “New Labels for Old Ideas: Predictive Processing and the Interpretation of Neural Signals.” *Review of Philosophy and Psychology* 11 (3): 517–46.
Chen, Zhe. 2003. “Bayesian Filtering: From Kalman Filters to Particle Filters, and Beyond.” *Statistics* 182 (1): 1–69.
Clark, Andy. 2013. “Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science.” *Behavioral and Brain Sciences* 36 (03): 181–204.
———. 2015. *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press.
———. 2017. “How to Knit Your Own Markov Blanket.” In *In Philosophy and Predictive Processing: 3*, edited by Thomas K Metzinger and Wanja Wiese. Open MIND. Frankfurt am Main: MIND Group.
———. 2020. “Beyond Desire? Agency, Choice, and the Predictive Mind.” *Australasian Journal of Philosophy* 98 (1): 1–15.
Clark, Andy, and David Chalmers. 1998. “The Extended Mind.” *Analysis* 58 (1): 7–19.
Colombo, Matteo. 2020. “Maladaptive social norms, cultural progress, and the free-energy principle.” *Behavioral and Brain Sciences* 43: e100.
Colombo, Matteo, Lee Elkin, and Stephan Hartmann. 2018. “Being Realist about Bayes, and the Predictive Processing Theory of Mind.” *The British Journal for the Philosophy of Science*, August.
Colombo, Matteo, and Cory Wright. 2018. “First Principles in the Life Sciences: The Free-Energy Principle, Organicism, and Mechanism.” *Synthese*.
Da Costa, Lancelot, Thomas Parr, Noor Sajid, Sebastijan Veselic, Victorita Neacsu, and Karl J Friston. 2020. “Active Inference on Discrete State-Spaces: A Synthesis.” *arXiv Preprint arXiv:2001.07203*.
Daunizeau, Jean. 2017. “The Variational Laplace Approach to Approximate Bayesian Inference.” *arXiv Preprint arXiv:1703.02089*.
Dayan, Peter, Geoffrey E Hinton, Radford M Neal, and Richard S Zemel. 1995. “The Helmholtz Machine.” *Neural Computation* 7 (5): 889–904.
Demekas, Daphne, Thomas Parr, and Karl J Friston. 2020. “An Investigation of the Free Energy Principle for Emotion Recognition.” *Frontiers in Computational Neuroscience* 14.
Dewhurst, Joe. 2017. “Folk Psychology and the Bayesian Brain.” In *Philosophy and Predictive Processing*. Frankfurt am Main: MIND Group.
Downes, Stephen M. 2020. *Models and Modeling in the Sciences a Philosophical Introduction*. Routledge.
Doya, Kenji. 2007. *Bayesian Brain: Probabilistic Approaches to Neural Coding*. MIT press.
Es, Thomas van. 2020. “Living Models or Life Modelled? On the Use of Models in the Free Energy Principle.” *Adaptive Behavior*.

Feldman, Harriet, and Karl J Friston. 2010. “Attention, Uncertainty, and Free-Energy.” *Frontiers in Human Neuroscience* 4: 215.
Fox, Charles W, and Stephen J Roberts. 2012. “A Tutorial on Variational Bayesian Inference.” *Artificial Intelligence Review* 38 (2): 85–95.
Friston, Karl, Lancelot Da Costa, and Thomas Parr. 2020. “Some Interesting Observations on the Free Energy Principle.” *arXiv Preprint arXiv:2002.04501*.
Friston, Karl J. 2005. “A theory of cortical responses.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 360 (1456): 815–36.
———. 2008. “Hierarchical models in the brain.” *PLoS Computational Biology* 4 (11).
———. 2010. “The free-energy principle: a unified brain theory?” *Nature Reviews. Neuroscience* 11 (2): 127–38.
———. 2012. “A Free Energy Principle for Biological Systems.” *Entropy* 14 (11): 2100–2121.
———. 2013a. “Active Inference and Free Energy.” *Behavioral and Brain Sciences* 36 (03): 212–13.
———. 2013b. “Life as We Know It.” *Journal of the Royal Society Interface* 10 (86): 20130475.
———. 2019. “A Free Energy Principle for a Particular Physics.” *arXiv Preprint arXiv:1906.10184*.
Friston, Karl J, and Ping Ao. 2012. “Free Energy, Value, and Attractors.” *Computational and Mathematical Methods in Medicine* 2012.
Friston, Karl J, Jean Daunizeau, James Kilner, and Stefan J. Kiebel. 2010. “Action and behavior: A free-energy formulation.” *Biological Cybernetics* 102 (3): 227–60.
Friston, Karl J, Erik D Fagerholm, Tahereh S Zarghami, Thomas Parr, Inês Hipólito, Loïc Magrou, and Adeel Razi. 2020. “Parcels and Particles: Markov Blankets in the Brain.” *arXiv Preprint arXiv:2007.09704*.
Friston, Karl J, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, and Giovanni Pezzulo. 2017. “Active Inference: A Process Theory.” *Neural Computation* 29 (1): 1–49.
Friston, Karl J, James Kilner, and Lee Harrison. 2006. “A Free Energy Principle for the Brain.” *Journal of Physiology-Paris* 100 (1): 70–87.
Friston, Karl J, Michael Levin, Biswa Sengupta, and Giovanni Pezzulo. 2015. “Knowing One’s Place: A Free-Energy Approach to Pattern Regulation.” *Journal of the Royal Society Interface* 12 (105): 20141383.
Friston, Karl J, Jérémie Mattout, Nelson Trujillo-Barreto, John Ashburner, and Will Penny. 2007. “Variational Free Energy and the Laplace Approximation.” *Neuroimage* 34 (1): 220–34.
Friston, Karl J, Thomas Parr, and Bert de Vries. 2017. “The Graphical Brain: Belief Propagation and Active Inference.” *Network Neuroscience* 1 (4): 381–414.
Friston, Karl J, Francesco Rigoli, Dimitri Ognibene, Christoph Mathys, Thomas Fitzgerald, and Giovanni Pezzulo. 2015. “Active inference and epistemic value.” *Cognitive Neuroscience*, 1–28.
Friston, Karl J, N. Trujillo-Barreto, and J. Daunizeau. 2008. “DEM: A variational treatment of dynamic systems.” *NeuroImage* 41 (3): 849–85.
Friston, Karl J, Wanja Wiese, and J Allan Hobson. 2020. “Sentience and the Origins of Consciousness: From Cartesian Duality to Markovian Monism.” *Entropy* 22 (5): 516.
Friston, Karl, Biswa Sengupta, and Gennaro Auletta. 2014. “Cognitive Dynamics: From Attractors to Active Inference.” *Proceedings of the IEEE* 102 (4): 427–45.
Gregory, Richard. 1970. *The Intelligent Eye*. Weidenfeld; Nicolson.

Hesp, Casper, Maxwell Ramstead, Axel Constant, Paul Badcock, Michael D Kirchhoff, and Karl J Friston. 2019. "A Multi-Scale View of the Emergent Complexity of Life: A Free-Energy Proposal." In *Evolution, Development, and Complexity: Multiscale Models in Complex Adaptive Systems. 1st Ed.*, Ch. 7. Springer Proceedings in Complexity.
Hinton, Geoffrey E, and Richard S Zemel. 1994. "Autoencoders, Minimum Description Length and Helmholtz Free Energy." In *Advances in Neural Information Processing Systems*, 3–10.
Hipolito, Ines, Maxwell Ramstead, Laura Convertino, Anjali Bhat, Karl Friston, and Thomas Parr. 2020. "Markov Blankets in the Brain." *arXiv Preprint arXiv:2006.02741*.
Hohwy, Jakob. 2013. *The Predictive Mind*. OUP Oxford.
———. 2016. "The Self-Evidencing Brain." *Noûs* 50 (2): 259–85.
———. 2017. "How to Entrain Your Evil Demon." In *In Philosophy and Predictive Processing: 2*, edited by Thomas K Metzinger and Wanja Wiese. Open MIND. Frankfurt am Main: MIND Group.
Hohwy, Jakob, and John Michael. 2017. "Why Should Any Body Have a Self?" In *The Subject's Matter: Self-Consciousness and the Body*, 363. MIT Press.
Jordan, Michael I, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. 1999. "An Introduction to Variational Methods for Graphical Models." *Machine Learning* 37 (2): 183–233.
Kappen, Hilbert J, Vicenç Gómez, and Manfred Opper. 2012. "Optimal Control as a Graphical Model Inference Problem." *Machine Learning* 87 (2): 159–82.
Kirchhoff, Michael D. 2018. "Autopoiesis, Free Energy, and the Life–Mind Continuity Thesis." *Synthese* 195 (6): 2519–40.
Kirchhoff, Michael D, and Julian Kiverstein. 2019. "How to Determine the Boundaries of the Mind: A Markov Blanket Proposal." *Synthese*, 1–20.
Kirchhoff, Michael, Thomas Parr, Ensor Palacios, Karl J Friston, and Julian Kiverstein. 2018. "The Markov Blankets of Life: Autonomy, Active Inference and the Free Energy Principle." *Journal of the Royal Society Interface* 15 (138): 20170792.
Klein, Colin. 2018. "What Do Predictive Coders Want?" *Synthese* 195 (6): 2541–57.
Knill, David C, and Alexandre Pouget. 2004. "The Bayesian Brain: The Role of Uncertainty in Neural Coding and Computation." *Trends in Neurosciences* 27 (12): 712–19.
Knill, David C, and Whitman Richards. 1996. *Perception as Bayesian Inference*. Cambridge University Press.
Kuchling, Franz, Karl J Friston, Georgi Georgiev, and Michael Levin. 2020. "Morphogenesis as Bayesian Inference: A Variational Approach to Pattern Formation and Control in Complex Biological Systems." *Phys Life Rev* 33: 88–108.
Kwisthout, Johan, Harold Bekkering, and Iris Van Rooij. 2017. "To Be Precise, the Details Don't Matter: On Predictive Processing, Precision, and Level of Detail of Predictions." *Brain and Cognition* 112: 84–91.
Lee, Tai Sing, and David Mumford. 2003. "Hierarchical Bayesian Inference in the Visual Cortex." *JOSA A* 20 (7): 1434–48.
Liwtin, Piotr, and Marcin Miłkowski. 2020. "Unification by Fiat: Arrested Development of Predictive Processing." *Cognitive Science* 44.
MacKay, David JC. 2003. *Information Theory, Inference and Learning Algorithms*. Cambridge university press.

Menary, Richard, and Alexander J. Gillett. 2020. “Are Markov Blankets Real and Does It Matter?” In *The Philosophy and Science of Predictive Processing*, edited by Dina Mendonça, Manuel Curado, and Steven S. Gouveia. Bloomsbury Academic.
Montúfar, Guido, Keyan Ghazi-Zahedi, and Nihat Ay. 2015. “A Theory of Cheap Control in Embodied Systems.” *PLoS Comput Biol* 11 (9): e1004427.
Murphy, Kevin P. 2012. *Machine Learning: A Probabilistic Perspective*. MIT press.
Neal, Radford M, and Geoffrey E Hinton. 1998. “A View of the Em Algorithm That Justifies Incremental, Sparse, and Other Variants.” In *Learning in Graphical Models*, 355–68. Springer.
Opper, Manfred, and Cédric Archambeau. 2009. “The Variational Gaussian Approximation Revisited.” *Neural Computation* 21 (3): 786–92.
Palacios, Ensor Rafael, Adeel Razi, Thomas Parr, Michael Kirchhoff, and Karl Friston. 2020. “On Markov Blankets and Hierarchical Self-Organisation.” *Journal of Theoretical Biology* 486: 110089.
Parisi, Giorgio. 1988. *Statistical Field Theory*. Addison-Wesley.
Parr, Thomas, Lancelot Da Costa, and Karl Friston. 2020. “Markov Blankets, Information Geometry and Stochastic Thermodynamics.” *Philosophical Transactions of the Royal Society A* 378 (2164): 20190159.
Parr, Thomas, M Berk Mirza, Hayriye Cagnan, and Karl J Friston. 2019. “Dynamic Causal Modelling of Active Vision.” *Journal of Neuroscience* 39 (32): 6265–75.
Pearl, Judea. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.
———. 2009. “Causal Inference in Statistics: An Overview.” *Statistics Surveys* 3: 96–146.
Pezzulo, Giovanni, Francesco Rigoli, and Karl J Friston. 2018. “Hierarchical Active Inference: A Theory of Motivated Control.” *Trends in Cognitive Sciences* 22 (4): 294–306.
Pourret, Olivier, Patrick Naïm, and Bruce Marcot. 2008. *Bayesian Networks: A Practical Guide to Applications*. John Wiley & Sons.
Ramstead, Maxwell James Désormeau, Paul Benjamin Badcock, and Karl John Friston. 2018. “Answering Schrödinger’s Question: A Free-Energy Formulation.” *Physics of Life Reviews* 24: 1–16.
Ramstead, Maxwell JD, Karl J Friston, and Inês Hipólito. 2020. “Is the Free-Energy Principle a Formal Theory of Semantics? From Variational Density Dynamics to Neural and Phenotypic Representations.” *Entropy* 22 (8): 889.
Ramstead, Maxwell JD, Michael D Kirchhoff, Axel Constant, and Karl J Friston. 2019. “Multiscale Integration: Beyond Internalism and Externalism.” *Synthese*.
Rao, Rajesh PN, and Dana H Ballard. 1999. “Predictive Coding in the Visual Cortex: A Functional Interpretation of Some Extra-Classical Receptive-Field Effects.” *Nature Neuroscience* 2 (1): 79–87.
Robert, Christian. 2007. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer Science & Business Media.
Rosas, Fernando E, Pedro AM Mediano, Martin Biehl, Shamil Chandaria, and Daniel Polani. 2020. “Causal Blankets: Theory and Algorithmic Framework.” *arXiv Preprint arXiv:2008.12568*.
Rubin, Sergio, Thomas Parr, Lancelot De Costa, and Karl Friston. 2020. “Future climates: Markov blankets and active inference in the biosphere.” *Journal of the Royal Society Interface* 17: 20200503.

Sajid, Noor, Philip J Ball, and Karl J Friston. 2019. “Active Inference: Demystified and Compared.” *arXiv Preprint arXiv:1909.10863*.
Sanborn, Adam N, and Nick Chater. 2016. “Bayesian Brains Without Probabilities.” *Trends in Cognitive Sciences* 20 (12): 883–93.
Seth, Anil, Beren Millidge, Christopher L Buckley, and Alexander Tschantz. 2020. “Curious Inferences: Reply to Sun and Firestone on the Dark Room Problem.” *Trends in Cognitive Sciences* 24 (9): 681–83.
Sims, Matthew. 2020. “How to Count Biological Minds: Symbiosis, the Free Energy Principle, and Reciprocal Multiscale Integration.” *Synthese*, 1–23.
Stephan, Klaas Enno, Will D Penny, Jean Daunizeau, Rosalyn J Moran, and Karl J Friston. 2009. “Bayesian Model Selection for Group Studies.” *Neuromiage* 46 (4): 1004–17.
Sun, Zekun, and Chaz Firestone. 2020a. “Optimism and Pessimism in the Predictive Brain.” *Trends Cogn. Sci.* 24: 683–85.
———. 2020b. “The Dark Room Problem.” *Trends Cogn. Sci.* 24: 346–48.
Tishby, Naftali, and Daniel Polani. 2011. “Information Theory of Decisions and Actions.” In *Perception-Action Cycle*, 601–36. Springer.
Tschantz, Alexander, Anil K Seth, and Christopher L Buckley. 2020. “Learning Action-Oriented Models Through Active Inference.” *PLOS Computational Biology* 16 (4): e1007805.
Van de Cruys, Sander, Karl J. Friston, and Andy Clark. 2020. “Controlled Optimism: Reply to Sun and Firestone on the Dark Room Problem.” *Trends Cogn. Sci.* 24 (9): 680–81.
Veissière, Samuel PL, Axel Constant, Maxwell JD Ramstead, Karl J Friston, and Laurence J Kirmayer. 2020. “Thinking Through Other Minds: A Variational Approach to Cognition and Culture.” *Behavioral and Brain Sciences* 43.
Weisberg, Michael. 2013. *Simulation and Similarity: Using Models to Understand the World*. Oxford University Press.
Weisberg, Michael. 2007. “Who Is a Modeler?” *The British Journal for the Philosophy of Science* 58 (2): 207–33.
Wilkinson, Sam, George Deane, Kathryn Nave, and Andy Clark. 2019. “Getting Warmer: Predictive Processing and the Nature of Emotion.” In *The Value of Emotions for Knowledge*, 101–19. Palgrave Macmillan.
Williams, Daniel. 2020. “Is the Brain an Organ for Prediction Error Minimization?” *A Preprint*.
Yon, Daniel, Cecilia Heyes, and Clare Press. 2020. “Beliefs and Desires in the Predictive Brain.” *Nature Communications* 11: 4404.
Zhang, Cheng, Judith Bütepage, Hedvig Kjellström, and Stephan Mandt. 2018. “Advances in Variational Inference.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41 (8): 2008–26.