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

Wordcounts:

Front page: 84

Short abstract: 104

Long abstract: 265

Keywords: 11

Main text: 16,200

References: 3142

Endnotes: 1352

Total words: 21,158

Short 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 a distinction 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 the two (and their associated research programs) will help clarify the literature.

Long Abstract

The free energy principle, an influential framework in computational neuroscience and theoretical neurobiology, starts from the assumption that living systems ensure adaptive exchanges with their environment by minimizing the objective function of variational free energy. Following this premise, it claims to deliver a promising integration of the life sciences. In recent work, Markov Blankets, one of the central constructs of the free energy principle, have been applied to resolve debates central to philosophy (such as demarcating the boundaries of the mind). The aim of this paper is twofold. First, we trace the development of Markov blankets starting from their standard application in Bayesian networks, via variational inference, to their use in the literature on active inference. We then identify a persistent confusion in the literature between the formal use of Markov blankets as an epistemic tool for Bayesian inference, and their novel metaphysical use in the free energy framework to demarcate the physical boundary between an agent and its environment. Consequently, we propose to distinguish between 'Pearl blankets' to refer to the original epistemic use of Markov blankets and 'Friston blankets' to refer to the new metaphysical construct. Second, we use this distinction to critically assess claims resting on the application of Markov blankets to philosophical problems. We suggest that this literature would do well in differentiating between two different research programs: 'inference with a model' and 'inference within a model'. Only the latter is capable of doing metaphysical work with Markov blankets, but requires additional philosophical premises and cannot be justified by an appeal to the success of the mathematical framework alone.

Keywords: active inference – Bayesian inference – free-energy principle – Markov blankets – scientific realism

1. Introduction

The last twenty years in cognitive science have been marked by what may be called a 'Bayesian turn'. An increasing number of theories and methodological approaches either appeal to, or make use of, Bayesian methods (prominent examples include Oaksford and Chater, 2001; Körding and Wolpert, 2004; Knill and Pouget, 2004; Griffiths and Tennenbaum, 2006; Tenenbaum et al. 2011; Clark 2013). The Bayesian turn pertains to both scientific methods for studying the mind, as well as to hypotheses about the mind's 'method' for making sense of the world. In particular, the application of Bayesian formulations to the study of perception and other inference problems has generated a large literature, highlighting a growing interest in Bayesian probability theory for the study of brains and minds.

Probably the most ambitious and all-encompassing version of the 'Bayesian turn' in cognitive science is the free energy principle (FEP). The FEP is a mathematical framework, developed by Karl Friston and colleagues (Friston, Kilner, and Harrison 2006; Friston et al. 2010; Friston 2010; Friston et al. 2017a; Friston 2019), which specifies an objective function that any self-organizing system needs to minimize in order to ensure adaptive exchanges with its environment. 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 1980, Grossberg 1980, Rao and Ballard 1999, Lee and Mumford 2003) is that not only perceptual processes, but also other cognitive functions such as learning, attention, and action planning can be subsumed under one single principle: the minimization of free energy through the process of active inference (Friston 2010; Friston et al. 2017). 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 2013; Friston et al. 2015a).

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, often using the Markov blanket construct that is the main focus of this paper. Via the use of Markov blankets, the FEP has been used to (supposedly) resolve debates about:

- the boundaries of the mind (Hohwy, 2017; Clark, 2017; Kirchhoff and Kiverstein 2021),
- the boundaries of living systems (Friston, 2013; Kirchhoff et al., 2018, van Es and Kirchhoff, 2021),
- the life-mind continuity thesis (Kirchhoff, 2018; Wiese and Friston, 2021, Kirchhoff and van Es, 2021)
- the relationship between mind and matter (Friston, Wiese, and Hobson 2020; Kiefer, 2020),

while also offering (apparently) new insights on:

- the (trans)formation and survival of social and societal organisations (Boik, 2021; Fox, 2021; Khezri, 2021),
- climate systems and planetary-scale self-organisation and autopoiesis (Rubin et al., 2020),
- the notions of 'self' and 'individual', with studies on the sense of agency and on body ownership (Hafner et al., 2020), (in utero) co-embodiment (Ciaunica et al., 2021), pain experience (Kiverstein, Kirchhoff, and Thacker, 2021) and symbiosis (Sims, 2020),
- multi-level theories of sex and gender (Fausto-Sterling, 2021), and
- ordering principles by which the spatial and temporal scales of mind, life, and society are linked (Hesp et al., 2018; Ramstead, Badcock, and Friston 2018; Veissière et al. 2020) and possibly evolve (Poirier et al., 2021).

The formalisms deployed by the FEP (as outlined in Section 3 and 4 of this paper) are sometimes explicitly presented as replacing older (and supposedly outdated) philosophical arguments (Ramstead et al. 2019; Ramstead, Friston, and Hipólito 2020), suggesting that they might be intended to serve as a

mathematical alternative to metaphysical principles. 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 theoretical constructs it describes are often assigned a slightly unconventional meaning whose full implications are not always obvious. While this might explain some of its appeal, as it can seem to be steeped in unassailable mathematical justification, it also risks the possibility of 'smuggling in' unwarranted metaphysical assumptions. Each new iteration of the theory also introduces novel formal constructs that can make previous criticisms inapplicable, or least require their reformulation (see for example the exchange between Sun and Firestone 2020a; Seth et al. 2020; Van de Cruys, Friston, and Clark 2020, as well as Sun and Firestone 2020b).

In this paper we want to focus on just one of the more stable formal constructs utilized by the FEP, namely the concept of a Markov blanket. Markov blankets originate in the literature on Bayesian inference and graphical modeling, where they designate a set of random variables that essentially 'shield' another random variable (or set of variables) from the rest of the variables in the system (Pearl 1988; Bishop 2006; Murphy 2012). By identifying which variables are (conditionally) independent from each other, they help represent the relationships between variables in graphical models, which serve as useful and compact graphical abstractions for studying complex phenomena. By contrast, in the FEP literature Markov blankets are now frequently assigned an ontological role in which they either represent, or are literally identified with, worldly boundaries. 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. By focusing here on the case of Markov blankets, we hope to give a specific diagnosis of this problem, and then a suggested solution, but our analysis does also have potentially wider implications for the general use of formal constructs in the FEP literature, which we think are often described in a way that is crucially ambiguous between a literalist, a realist, and an instrumentalist reading (see Andrews 2020 and van Es 2021 for broader reviews of these kinds of issues in the FEP literature).

In order to give a comprehensive picture of where the field is now, we need to first go back to basics and explain some fundamental concepts. We will therefore start our paper by tracing the development of Markov blankets in Section 2, beginning with their standard application in graphical models (focusing on Bayesian networks) and probabilistic reasoning, and including some of the formal machinery required for variational Bayesian inference. In Section 3 we present the active inference framework and the different roles played by Markov blankets within this framework, which we suggest has ended up stretching the original concept beyond its initial formal purpose (here we distinguish between the original 'Pearl' blankets and the novel 'Friston' blankets). In Section 4 we focus specifically on the role played by Friston blankets in distinguishing the sensorimotor boundaries of organisms, which we argue stretches the original notion of a Markov blanket in a potentially philosophically unprincipled manner. In Section 5 we discuss some conceptual issues to do with Friston blankets, and in Section 6 we suggest that it would be both more accurate and theoretically productive to keep Pearl blankets and Friston blankets clearly distinct from one another when discussing active inference and the free energy principle. This would avoid conceptual confusion and also disambiguate two distinct theoretical projects that might each be valuable in their own right.

2. Probabilistic reasoning and Bayesian networks

The concept of a Markov blanket was first introduced by Judea Pearl (1988) in the context of his work on probabilistic reasoning and graphical models. In this section we will introduce the formal background that is required in order to understand the role played by Markov blankets in this literature. This will provide the necessary foundation for Sections 3 and 4, where we will discuss the ways in which Markov blankets have been used (and potentially misused) within the FEP literature.

2.1 Probabilistic reasoning

Probabilistic reasoning is an approach to formal decision-making under uncertain conditions. This approach is typically introduced as a middle ground between heuristics-based systems that are fast but will face many exceptions, and rules-based systems that will be accurate but slow and hard to put into practice. The probabilistic reasoning framework is a way to summarize relevant exceptions, providing

a middle ground between speed and accuracy. The first step in this approach is to classify variables in order to distinguish between observables and unobservables. Inference is then the process by which one can estimate an unobservable given some observables. For instance, how is it that we are able to determine if a watermelon is ripe by knocking on it? On the basis of observing the sound (resonant or dull), we are able to infer the unobserved state of the watermelon (ripe or not). When formalizing such kinds of everyday inference problems, we need to answer three interrelated questions:

1.) How do we adequately summarize our previous experience?

2.) How do we use previous experience to infer what is going on in the present?

3.) How do we update the summary in the light of new experience?

In Section 2.2 we will address Bayesian networks, a specific way of answering question 1. In Section 2.3 we will address variational inference, a specific way of addressing question 2. Question 3 is addressed by appealing to Bayes theorem. Bayes theorem normally takes the following form:

$$p(x|y) = \frac{p(y,x)}{p(y)} = \frac{p(y|x)p(x)}{p(y)} \quad (1)$$

This formula is a recipe for calculating the *posterior probability*, $p(x|y)$, of an unobserved set of states $x \in X$ given observations $y \in Y$. The probability $p(x)$ captures prior knowledge about states x (i.e. a *prior probability*), while $p(y|x)$ describes the *likelihood* of observing y for a given x. The remaining term, $p(y)$, represents the probability 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 sums up to 1. In other words, the posterior probability $p(x|y)$ represents the 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 $p(y|x)$ of how observations are generated in the first place (e.g., how watermelons give rise to different sounds at specific maturation stages, including the observed sound y), normalised by the

knowledge about the observations integrated over all possible hidden variables, $p(y)$ (e.g., how watermelons may typically sound, regardless of the specific maturation stage).

What holds for everyday reasoning problems holds for cognition and science as well: how can a cognitive system estimate the presence of some object on the basis of the state of its receptors alone? How can a neuroscientist estimate brain activity on the basis of magnetic fields measured in an fMRI scanner? Both of these kinds of questions can be formalized using Bayes' theorem (see e.g. Gregory 1980, Penny et al. 2011, Friston, Harrison, and Penny 2003).

Although this scheme offers a powerful tool for probabilistic inference, it is mostly limited to simple, low-dimensional, and often discrete or otherwise analytically tractable problems. For example, computing the exact model evidence is rarely feasible, because the computation is often analytically intractable or computationally too expensive (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. 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. We will discuss those in Section 2.3, but first we will introduce the Bayesian network approach developed by Pearl.

2.2 Bayesian networks

Pearl (1988) developed a mathematical language to formulate summaries of previous experience in computer learning systems. That mathematical language constitutes the focus of this paper, due to the ease with which it can be used to demonstrate the use (and misuse) of Markov blankets using probabilistic graphical models. Probabilistic graphical models capture the dependencies between random variables using a visual language that renders the study of certain probabilistic interactions across var-

iables, traditionally defined with analytical methods, more intuitive and easy to track.[i] Random variables are drawn as *nodes* in a graph, with shaded nodes usually representing variables that are *observed* and empty nodes used for variables that are *unobserved* (latent or hidden variables). The (probabilistic) relationships between such random variables are then expressed using edges (lines) connecting the nodes. For present purposes we will focus on acyclic graphs with directed edges, which provide the basis for graphical models, and play a crucial role in the context of active inference (Friston, Parr, and de Vries 2017). Relationships between the variables are often described using genealogical terms, with $pa(a)$ being the *parents* (or 'ancestors') of their *child* (or 'descendant') node a. In Figure 1 below, m is the co-parent (with e) of a and the child of c and b, while c and b are co-parents of m. Although the dependencies are formally defined in terms of basic manipulations on probability distributions, graphical models provide some practical advantages in reasoning about these formal properties, presenting a clear and easily interpretable depiction of the relationships between variables.

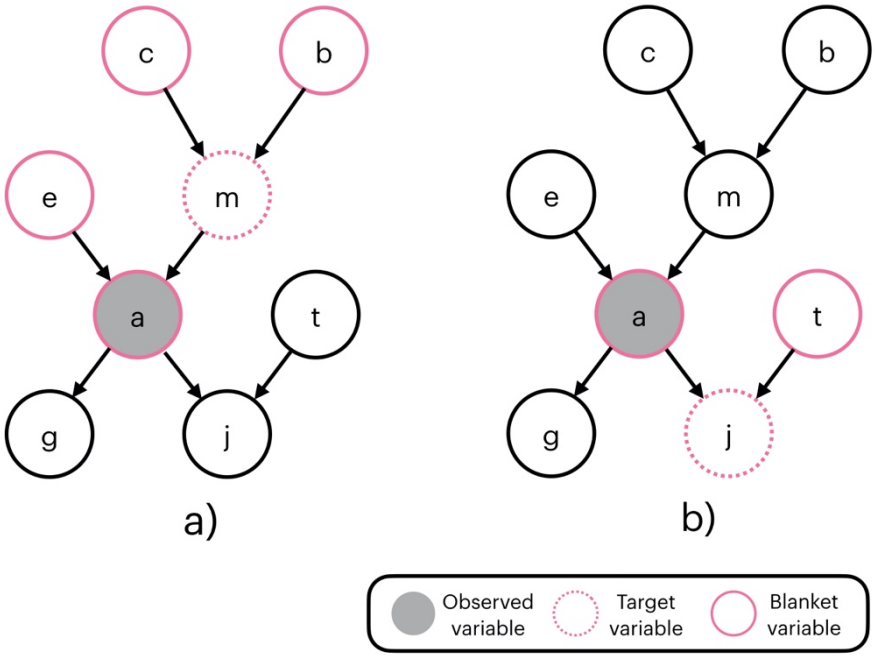

Figure 1: The 'alarm' network with examples of Markov Blankets for two different variables. The target variables are indicated with a dashed pink circle, while the variables that are part of the Markov blanket are indicated with a solid pink circle.

Let us introduce a simple textbook example that will help familiarise us with some of the nuances of Bayesian graphs. The illustration we will consider is a slight modification of a common textbook example, the 'alarm' network (Pearl, 1988). Imagine that you have an alarm system (a) in your house

and it is sensitive to motion, so that it will go off whenever it detects any movement (m). In some cases the movement can be caused by a burglar (b), but it could also be caused by your neighbour's cat (c). The alarm is also sensitive (for independent reasons) to power surges in the electrical grid, and can sometimes be triggered by changes in the supply of electricity (e). Of course, having an alarm is not much help when you're away, so you asked two of your neighbours - Gloria (g) and John (j) - to call you if they hear the alarm. Unfortunately, John suffers from severe tinnitus (t) and has been known to call you even though the alarm wasn't on. This example can be formalized both algebraically and visually.

Algebraically, this example can be expressed by the following joint probability of all the included variables:

$$p(a, b, c, e, g, j, t, m) = p(g|a)p(j|a, t)p(a|e, m)p(e)p(m|c, b)p(c)p(b). \quad (2)$$

This joint probability is not especially easy to interpret. The graph in Figure 1 models the dependencies among the variables in this scenario in a more easily interpretable manner, where directed edges indicate probabilistic relationships between nodes (variables).

The alarm network allows us to illustrate a number of canonical examples of statistical (in)dependencies between nodes, known also as d-separation (Pearl 1988):

- e and m are marginally independent but only conditionally dependent if a is observed (i.e., when a becomes a shaded node), a case technically known also as head-to-head relation. This can be made intuitive in the following way: in general surges in electricity e and other forms of movement m are not related to one another. Once you know that the alarm went off, then knowing that there was no surge implies that some other factor was responsible for the activation (and vice versa).

- c and a are marginally dependent but conditionally independent if m is observed, also known as head-to-tail. Once you know that there was movement, knowing that the cat caused the movement will not make a difference in your estimation for whether the alarm went off.
- g and j are marginally dependent but conditionally independent if a is observed, also known as tail-to-tail. In general, Gloria calling will make it likely that John will call as well. But once you know the alarm went off, Gloria calling will not change the probability of John calling.

Bayesian networks like the one above play an especially prominent role in exemplifying marginal and conditional independence relations. Marginal independence is represented by the lack of a directed path between two nodes. Conditional independence is defined in terms of a node 'shielding' one variable (or set of variables) from another node. This notion of 'shielding' can be made more explicit by introducing the idea of a Markov blanket, which will be the central focus of this paper.

A Markov blanket designates the minimal[ii] set of nodes with respect to which a particular node (or set of nodes) is *conditionally independent* of all other nodes in a Bayesian graph[iii], i.e. it *shields* that node from all other nodes. Formally, a Markov blanket for a set of variables x_i is thus equivalent to:

$$mb(x_i) = pa(x_i) \cup ch(x_i) \cup copa(x_i), \tag{3}$$

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.

To make the notion of a Markov blanket clearer, we have drawn the blankets of different nodes in the alarm network. Figure 1a shows the Markov blanket for node a or $mb(a)$. It is composed of a's parents (e and m), its children (g and j) and its children's other parents (t in the case of j). The $mb(j)$ shown in Figure 1b, on the other hand, is composed of just two nodes (a and t), which means that the state of variable j is independent of other nodes in the network, hence:

$$mb(j) = \{a, t\} \text{ and } mb(a) = \{e, m, g, t\}. \quad (4)$$

What this means intuitively is that given the Markov blanket of a node, any other change in the network will not make a direct difference to one's estimation of the random variable. If you could know John's state of tinnitus and the state of the alarm, you can calculate the probability that he will be calling. The rest of the state of the network does not make a difference for this calculation. In other words, a node's Markov blanket captures exactly all nodes that are relevant to infer the state of that node. As we will illustrate in the next section, the conditional independence of any variable from the nodes outside its Markov blanket is one of the key factors that makes probabilistic graphs useful for inference.

2.3 Variational inference

We mentioned before that exact Bayesian inference will in many cases not be feasible. There are a number of techniques available in the literature to perform approximate inference. The version of approximate inference that we will focus on in this paper is called variational inference, and here Markov blankets play an important role in identifying which variables are actually relevant to any given inference problem.

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 can be transformed into an optimization 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 is then 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 $p(x|y)$. This measure of

closeness is formalized by the Kullback-Leibler divergence, a common measure of dissimilarity between two probability distributions (here denoted by D_{KL}):

$$q^*(x) = argmin_{q(x)\in Q} D_{KL}\big(q(x) \parallel p(x|y)\big). \quad (5)$$

Equation 5 reads: the optimal distribution is the one that minimizes the dissimilarity between the variational density and the exact posterior. This can be shown to be bounded (above) by the minimisation of a quantity that is called variational free energy (see Murphy 2012 and Bishop 2006):

$$q^*(x) = argmin_{q(x)\in Q} \int q(x)\, ln \frac{q(x)}{p(y,x)} dx$$

$$= argmin_{q(x)\in Q} F(x) \quad (6)$$

One of the most crucial components of variational inference is the choice of a family Q. If the chosen Q is too complicated, then the inference will remain unfeasible, but if it is too simple then the optimal distribution might be too far removed from the exact posterior. Popular choices for Q include a treatment in terms of conjugate priors (Bishop, 2006), a mean-field approximation (Parisi, 1988), the variational Gaussian approximation (Opper and Archambeau, 2009) and the Laplace method (MacKay, 2003).

It is however crucial to highlight that such methods operate only on the family Q of the variational density $q(x)$. This means that they do not necessarily encode dependencies capturing constraints among variables $x_i \in x$ derived from knowledge of the underlying system to be modelled (e.g., its physics). These further constraints are instead captured in the joint probability $p(y, x)$, used to infer x via the posterior $p(x|y)$, of which $q(x)$ is an approximation (see equation 6). It is here that the concepts of marginal and conditional independence show up again. Inferential processes can in fact be simplified by orders of magnitude if we consider that each variable will only exert some (direct) influence on a number of (other) variables that is usually quite limited.

In the mean-field approach for example, mean-field effects (i.e., averages) for a particular partition (i.e., a subset) of variables are constructed only using its Markov blanket (Jordan et al. 1999). This means that such partition need only be optimized with respect to its blanket states, hence the idea of 'shielding', intended to highlight how only a relatively small number of variables need actually be considered in most problems of inference (Bishop, 2006; Murphy, 2012). In more concrete terms, and using our previous example of the alarm-network, to infer the most likely cause that set off the alarm one need not consider burglary (b) directly, as the effects of this variable are already captured by motion (m). Likewise, when trying to infer if John (j) will have to call us, we need only consider if the alarm was actually set off, regardless of whether it was because of some electricity supply problem (e) or some motion detected by the alarm (m), or whether John's tinnitus (t) is the true cause of John's call. Through an iterative procedure in which each (subset of) node(s) is optimized given its Markov blanket, the process will settle on the best estimate of the posterior distribution given the simplifying assumptions that were made for a particular model. As we can see by now, Markov blankets are a relatively technical construct traditionally applied to problems of inference.

2.4 Bayesian model selection

One of Pearl's main innovations when it comes to Bayesian networks was the idea that dependencies between different variables of the original system could be discovered by manipulating (i.e. 'intervening on') a chosen variable and seeing which other variables are affected. This idea has proven to be immensely useful when trying to infer the organization of some system with an unknown structure, i.e., for *structure learning*, or *structure discovery*. Historically, however, other distinct approaches have also been adopted to tackle this problem. For example, structure learning can be utilized either with or without the causal assumptions advocated by Pearl and others (see Vowels et al. 2021 for a recent review). In this family of methods, the class of score-based approaches (Vowels et al. 2021) is of particular interest to this paper given its tight relations to the FEP and the use of Markov blankets.

In score-based approaches, to discover the values and relations between variables one simply constructs multiple (classes of) models of the system under investigation and compares them to determine which one of them makes the most accurate predictions about the observable data.

This process of pitting models against each other is often referred to as (possibly Bayesian) model selection (Stephan et al. 2009; Penny et al. 2011). Importantly, while this process optimizes for how well different models fit the data, it also keeps track of the tradeoff between model accuracy and model complexity. For example, it is clear that the alarm network we discussed before could have been more complex: either Gloria's or John's telephone batteries might play a role in whether they phone you or not, perhaps there are other ways in which the alarm might be triggered, and so on. However, the inclusion of such information in the network would have further complicated the graph without necessarily making it more accurate as a modelling tool (at least relative to our purposes).

What then decides the level of complexity that a good Bayesian model should have? Is it one that captures all the possibly relevant facts that might make a difference, or is it the simplest one that still makes a good enough prediction? The dominant assumption in the literature is that there is a tradeoff between making a model fit the data as closely as possible and that model's ability to predict new data points. In other words, the best model is one that accounts for the available data in the most parsimonious way (Stephan et al. 2009; Penny et al. 2011; Friston, Parr, and de Vries 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 variational free energy (via the maximisation of model evidence, equivalent to the minimisation of surprisal), but there is a general agreement that Bayeisan methods offer a quantification of Ockham's Razor (Jefferys and Berger, 1991). 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 discussed by MacKay 2003; Bishop 2006; Friston 2010; Daunizeau 2017).

It is therefore important to note that the basic epistemic aim (even for the models used in the context of active inference) is not to arrive at a *complete* model of the system under investigation, but rather to obtain the most parsimonious model that accurately captures the relevant relations (Stephan et al. 2010; Baltieri and Buckley, 2019). This complexity/accuracy trade-off is important to prevent overfitting the model to the available data.

Of course, which facts are relevant depends on the questions we ask: if we are interested in how an alarm can be sensitive to both motion and changes in electric current, the model drawn in Figure 1 might not be very helpful, but it would do just fine for the purpose of estimating (i.e. inferring) the probability that your house is really being robbed when your tinnitus-struck neighbour calls you to report a ringing noise. There is therefore a sense in which model selection is influenced by pragmatic considerations. By choosing the data worth considering for their analysis, the scientist chooses their level of analysis, and by choosing which dimensions in model space are relevant to answer their question, the scientist chooses what models (or families of models) to consider (Stephan et al. 2010; Penny et al. 2011). The same phenomenon can be analysed using different sources of data. For example, in a study of decision making one can include only behavioral data, or add neural measurements as well. The choice of relevant dimensions in models space is often influenced by previous empirical evidence, meaning that relevant factors and model spaces themselves should be updated as new evidence becomes available. Clearly these considerations are not unique to (Bayesian) model selection. Furthermore, they don't negate any of its merits, but rather simply highlight the requirement for pragmatic constraints in solving difficult problems with infinitely large model spaces, especially in realistic situations and away from hypothetical ideal observer scenarios.

2.5 Taking stock

We have introduced a number of concepts and constructs that jointly form a toolkit for Bayesian inference: Bayesian networks can provide problem-specific summaries of the available data that predict the probability of future observations. Variational inference provides an elegant method to replace an

intractable inference problem with a tractable optimization problem. Variational methods of the kind we have described in this section have been employed across the sciences. In this scientific context, Markov blankets are an auxiliary technical concept that demarcate what additional nodes are relevant for estimating the state of a specific target node.

This technical concept of a Markov blanket has undergone a significant transformation in the literature on the FEP. In order to distinguish this original Markov blanket concept from the one that we will draw out of the FEP literature in Section 4, we will, with apologies to Judea Pearl, refer to instances of the original concept as 'Pearl blankets' throughout the rest of the paper. The novel Markov blanket concept introduced in Section 4, on the other hand, we will refer to as a 'Friston blanket'.[iv]

3. Pearl blankets in the active inference framework

The specific application of the free energy principle that we will focus on here is the active inference framework. In active inference, the concepts of variational inference are applied to living systems. The thought is that living systems are in the same position as data scientists. They 'observe' the activity at their sensory receptors and need to infer the state of the world. However, the framework goes even further and postulates that living systems need to also act on the world so as to stay within viable bounds, as merely inferring the states of the environment cannot guarantee survival (this idea is illustrated in Figure 2). In this section we will introduce the way that Pearl blankets are used for modelling purposes in the active inference literature and highlight one initial conceptual issue with this use.

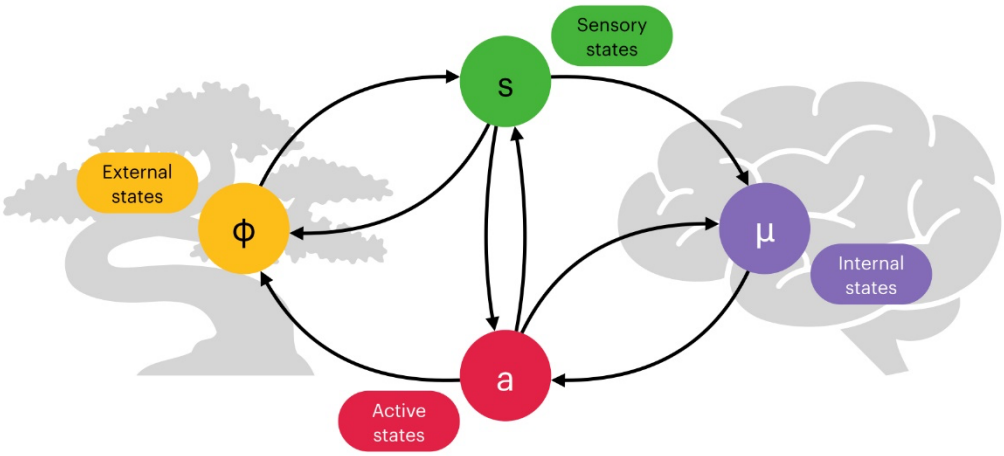

Figure 2: *The Markov blanket as a sensorimotor loop* (adapted from Friston, 2012). A diagram representing possible dependences between different components of interest: sensory states (green), internal states (violet), active states (red), and external states (yellow). Notice that although this figure uses arrows to signify directed influences, the diagram is not a Bayesian network as it depicts different sets of circular dependences (between pairs of components, and an overall loop including all nodes).

3.1 Modelling active inference with Pearl blankets

Active inference is a process theory derived from the application of variational inference to the study of biological and cognitive systems (Friston et al. 2010; Friston 2013; Friston, et al. 2015b; Friston 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'). Active inference attempts to explain action, perception, and other aspects of cognition under the umbrella of variational (and expected) free energy minimisation (Friston et al. 2010; Feldman and Friston 2010; Friston 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 accounts for more complex processes of action and policy selection (Friston et al. 2015b; Friston et al. 2017; Tschantz, Seth, and Buckley 2020). Expected free energy is 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. A full treatment of active inference remains beyond the scope of this manuscript (for some technical treatments and reviews, see e.g. Bogacz 2017; Buckley et al. 2017; Da Costa et al. 2020; Friston, Parr, and de Vries 2017; Biehl et al. 2018; Sajid et al. 2021), but we wish to highlight the formal connection between this framework and the use of variational Bayes in standard treatments of approximate probabilistic inference (as described in the previous section). Acknowledging this relationship is crucial if we want to understand the role Pearl blankets might play in active inference.

To understand the role played by Pearl blankets in active inference, we first need to identify some of the formal notation used by active inference, which is related to the variational approaches described in the previous section. Here we use the notation previously adopted in equation (6), while also introducing a second, distinct, set of hidden random variables: action policies $\pi \in \Pi$, sequences of control states $u \in U$ up to a given time horizon τ with $0 \leq \tau \leq T$, i.e. $\pi = [u_1, u_2, \dots, u_\tau]$. This will allow us to formulate perception and action as variational problems in active inference. Perception is the minimization (at each time step t)[v] of the following equation:

$$q^*(x, \pi) = argmin_{q(x,\pi) \in Q} F(x, \pi) \tag{7}$$

In other words: at each time step t, select the variational density that minimizes free-energy. Action is then characterised (at each time step t) in terms of control states u where:

$$u^* = argmax_{u \in U} \sum_{\pi \in \Pi, \pi_t = u} q\,(\pi) \tag{8}$$

and with the (approximate) prior on a policy π, $q(\pi)$, defined as

$$q(\pi) = \sigma\big(-G(\pi, \tau)\big). \tag{9}$$

This describes action selection as a minimisation of what is called expected free energy, $G(\pi, \tau)$, based on beliefs about future and unseen observations y, up to a time horizon $\tau \leq T$. In other words, at each time step t, select the policy π that you expect will minimize free-energy a number of time steps τ into the future (for a more detailed treatment, see one of the latest formulations found in, e.g., Da Costa et al. 2020, Sajid et al. 2021).

In doing so, we can notice that equation (7) essentially mirrors the previously defined equation (6), with the important caveat that in active inference sequences of control states (i.e., policies π) are now a part of the free energy F (this is conceptually similar to other formulations of control as inference, such as Attias 2003 and Kappen, Gómez, and Opper 2012)[vi]. 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 explicitly relied on a set of assumptions similar to the ones mentioned in the previous section: a mean-field approximation and the use of Pearl blankets to shield nodes. As mentioned in Section 2.3 (see also Jordan et al. 1999), Pearl blankets can be used to simplify the minimisation of variational free energy by specifying which variables need to be considered for mean-field averages via appropriate constraints of conditional independence. Works

such as Friston et al. (2007), Friston, Trujillo-Barreto, and Daunizeau (2008), and Friston (2008), however, make use of a 'structured' mean-field assumption[vii], where variables are partitioned in three independent sets: 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, albeit slightly unnecessary given the 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 (out of three) "the Markov [= Pearl] blanket contains all [other] subsets, apart from the subset in question" (Friston 2013, 2008; Friston et al. 2007; Friston, Trujillo-Barreto, and Daunizeau 2008), where "all [other] subsets" corresponds to the remaining two. As we will see shortly, the concept has gained a new life in more recent formulations of active inference, where it is applied in a substantially different way and as more than just a formal tool.

3.2 Models of models

There is an initial conceptual issue that arirses from 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 latent structure of unobserved features of the world. This parallel puts cognitive neuroscience in a rather special place: as making *models* of how animals *model* their environment. An important strategy in model-based cognitive neuroscience is to use different sources of data (such as behavioral and neural data) to infer the most likely model that the agent's brain might be implementing. For example, Parr et al. (2019) investigate the generative models that underlie active vision. They use both MEG and eye-tracking to disambiguate a number of potential generateve 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. The most likely model is selected (i.e. the one that best explains the data in the most parsimonious way) by scoring each model based on its accuracy in predicting neural dynamics and oculomotor behavior and weighing the scores by that model's complexity. We can identify two separate 'models' in this scenario: one is a computational Matlab model used by scientists for the purpose of causal dynamical inference, while the other is the target system's own model of its environment.

Thus, the scientist uses their Matlab model to infer which particular model the target system might implement.

While not wholly uncontroversial (as we will see in later sections), this kind of doubling up of modeling relations is widespread in neuroscience and remains relatively innocuous, so long as one is conceptually careful. What we mean by this is that one needs to not only 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, but one should also be transparent about one's commitment to the existence of the features represented on different levels of these modelling relations. Paying closer attention to said modeling relations provides a useful lens for analysing the difference between Pearl and Friston blankets: Pearl blankets can be used to identify probabilistic (in)dependencies between the variables in either the scientist's model of the agent-environment system, or the system's own model of the environment (in both cases these relations can be represented using a Bayesian network), while Friston blankets are posited as demarcating real boundaries in the agent-environment system itself (as we will see in the next section). The use of Pearl blankets in active inference, as described in this section, is rather uncontroversial. It is, however, unlikely to be of much philosophical interest, as Pearl blankets exist inside of models and cannot by themselves settle questions about the boundaries between agents and their environments.

4. Friston blankets as organism-environment boundaries

In a number of recent theoretical and philosophical works based on the free energy principle, Markov blankets have been assigned a role that cannot play under the standard definition of Pearl blankets presented in the previous section. In some formulations of active inference, starting with Friston and Ao (2012), Friston (2013), and Friston, Sengupta, and Auletta (2014), Markov blankets are in fact introduced to directly describe a specific form of conditional independence *within* a dynamical system, serving as a boundary between organism and world. In other words, they are considered to be proper parts of the target system and not merely parts of the scientist's model used to map that system. Just as some parts of a cartographical map are considered to represent features of the real world (such as

mountains and rivers) and others are not (such as contour lines), Markov blankets were originally just a statistical tool used to analyse models (akin to contour lines), but in the FEP literature are now often assumed to correspond to some real boundary in the world (akin to mountains and rivers). In order to distinguish this novel use of Markov blankets from the Pearl blankets discussed in the previous section, we will now call Markov blankets, understood in this new Fristonian sense, 'Friston blankets'.

4.1 Life as we know it?

Friston's "Life as we know it" (2013), which presents a proof-of-principle simulation for conditions claimed to be relevant for the origins of life, is one of the milestone publications in the FEP literature and has played a central role in the transition between the two uses of Markov blankets. 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 (some examples are listed in the introduction). Friston's paper aims to show how Markov blankets spontaneously form in a (simulated) 'primordial soup' and how these Markov (or 'Friston') blankets constitute an autopoietic boundary.

In the simulation itself, a number of particles are modeled as moving through a viscous fluid. The interaction between the particles is governed by Newtonian and electrochemical forces, both only working at short-range. By design, 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 (Figure 3). We will go through this simulation in some detail, because it is the archetype for the reification of the Markov blanket construct that we find throughout the active inference literature.

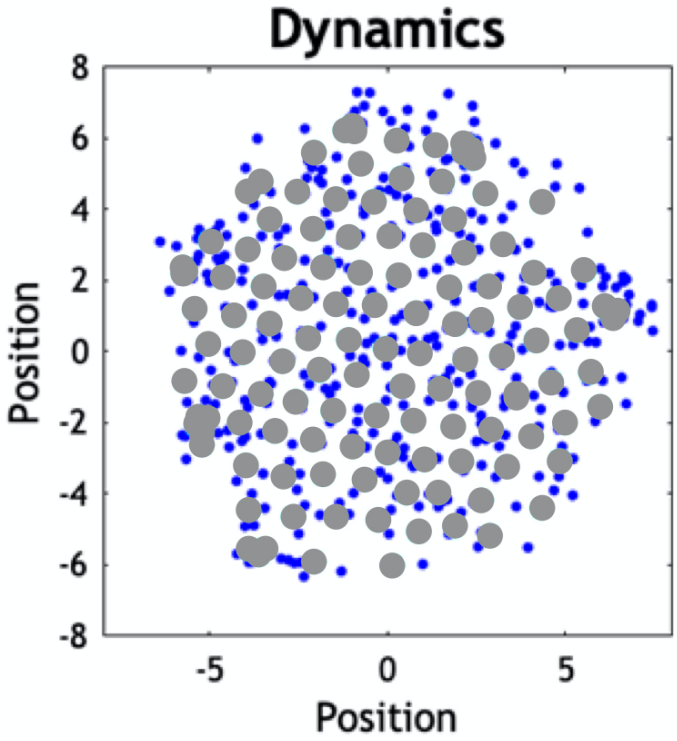

Figure 3: *The 'primordial soup'* (adapted from Friston 2013 using the code provided). The larger (grey) dots represent the location of each particle, which are assumed to be observed by the modellers. There are three smaller (blue) dots associated with each particle, representing the electrochemical state of that particle

Using the model adopted in the simulations (for details please refer to Friston 2013), 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 steady state at time T, when the simulation is stopped – a condition that remains unclear in the original study). The adjacency matrix is itself a representation of the electrochemical interactions between particles, and it is claimed that it can be interpreted as an abstract depiction of a Bayesian network (we would like to note, however, that this claim itself rests on additional assumptions that are not made explicit by Friston). 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. In this work, the directed edge is drawn if and only if particle r electrochemically affects particle s (Figure 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).

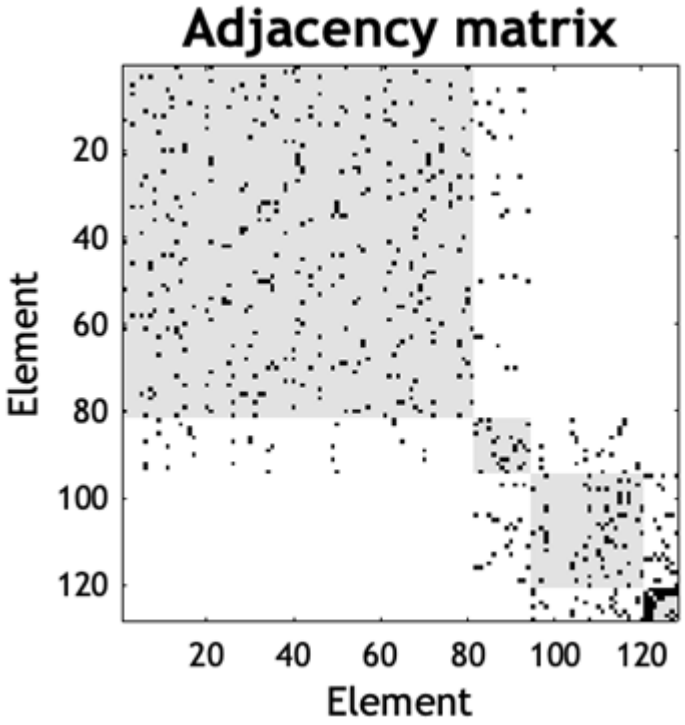

Figure 4: *The adjacency matrix of the simulated soup at steady-state* (from Friston 2013). Element i, j has value 1 (a dark square) if and only if subsystem i electrochemically affects subsystem *j*. The four grey squares from top left to bottom right represent the hidden states, the sensory states, the active states and the internal states respectively.

Spectral graph theory is then used to identify the 8 most densely coupled nodes, which are stipulated to be the 'internal' states.[viii] Given these internal states, the Markov blanket is then found through tracing the parents, children and co-parents of children in the network (see equation 18 in Friston 2013). States that are not internal states and are *not* part of the Markov blanket are then called 'external states'.

At this point of the analysis of the simulation, Friston introduces another interpretive step, proposing that the variables in this Markov blanket can be further separated into 'sensory' and 'active' states. The sensory states are those states of the Markov blanket whose parents are external states, while the active states are all other states of the Markov blanket (typically, but not always, active states will have children who are external states).

This procedure thus consists of first identifying the internal states and the states in their Markov blanket, classifying all other states as external, and then determining whether the states of the Markov blanket are sensory or active states (see Figure 5). This delivers four sets of states:

- μ: internal states: stipulated beforehand (Friston 2013 uses spectral graph theory to choose eight)
- ϕ: external states: all states not part of μ or its Markov blanket
- s: sensory states: states of the Markov blanket of μ whose parents are external states
- a: active states: the remaining states of the Markov blanket of μ

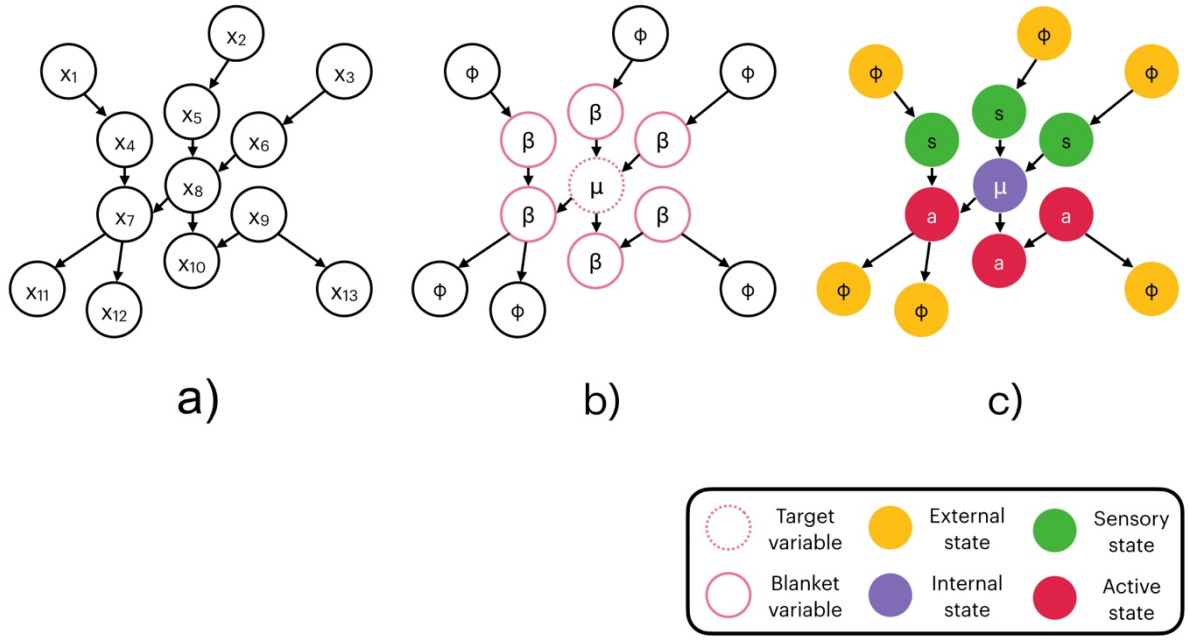

Figure 5: *The Friston blanket.* The three diagrams representing the stages of identifying a Friston blanket described in section 4.1. A system of interest is represented in the form a directed graph (a). Next the variable of interest is identified and a Markov blanket of shielding variables β is delineated separating the internal variable μ from the external ones denoted by ϕ (b). Finally, the variables within the blanket are identified as sensory s or active a depending on their relations with the external states (c).[ix]

Applied to the primordial soup simulation, each particle can be coloured to indicate which of these sets it has been assigned to (see Figure 6). 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, one has the impression that the active and sensory states form a structure similar to a cell membrane.

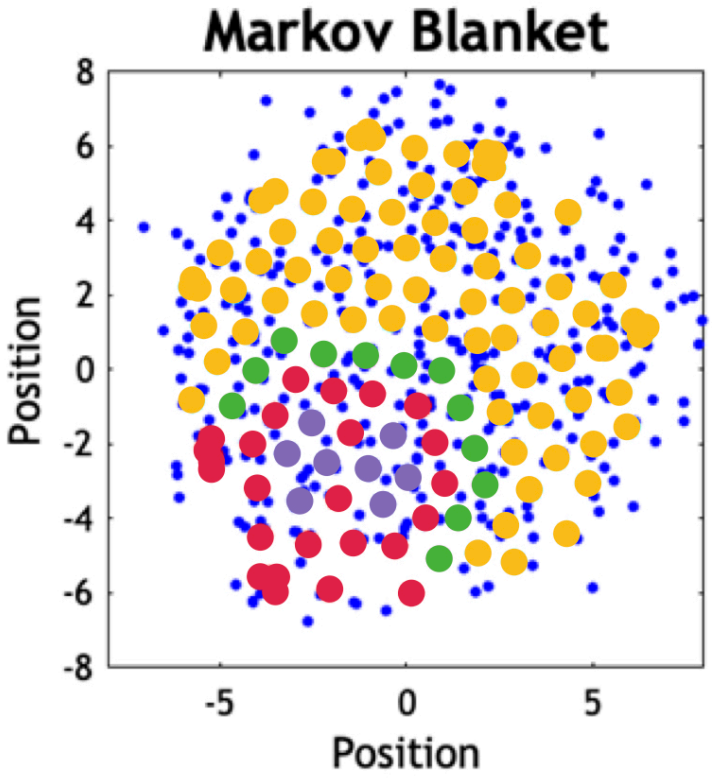

Figure 6: *The Markov blanket of the simulated soup at steady-state in* (adapted from Friston 2013 using the code provided). Similarly to Figure 3, particles are indicated by larger dots. Particles which belong to the set of sensory states are in green, active states are in red. Internal states are violet, while external states are marked in yellow. A 'blanket' of active and sensory cells surrounding the internal particles can be seen.

The 'Markov blanket formalism' advocated by Friston (2013) and described formally above does most of the work in the active inference literature when it comes to identifying internal, sensory, active, and external states. This formalizing step requires a number of non-arbitrary assumptions, some of which are now included in Friston et al. (2021a, b), but were not present in the original "Life as we know it" paper, and thus have been ignored in much of the subsequent literature. 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. If different thresholds were used to determine whether two nodes are connected, the adjacency matrix would look very different. The demarcations made by analysing the adjacency matrix are then used to label the nodes in the original system (as in Figure 6 above).

4.2 Friston blankets

The primordial soup simulation is claimed to provide a formal model for the emergence of agent-environment systems. We need to make a distinction between three different constructs: the 'real' primordial soup (i.e. the target system), a model of the primordial soup (i.e. an idealized representation of the soup) and the adjacency matrix (i.e. a further abstraction of the idealized model). A Friston blanket, according to the treatment in Friston (2013), can be identified using the adjacency matrix once a set of nodes of interest has been identified.[x] A first interpretative step is taken when labeling the nodes of the idealized model as internal, external, active and sensory states (i.e. as part of the Friston blanket). A further, and more problematic step is taken when extending the interpretation to the target system. The idea now is that, using the Markov blanket formalism, it is possible to uncover hidden properties of the target system which, in some sense, "instantiates" (Friston, 2013, p. 2) or "possesses" (*ibid*. p. 1) a Markov blanket. This procedure of attributing a property of the map (the Bayesian network) to the territory (the simulated soup, and by implication, the real primordial soup itself) is problematic because it reifies abstract features of the map (cf. Andrews 2020). A further implication of this step is that Markov blankets, which were initially introduced by Pearl as a formal property of directed, acyclical graphs, are now seen as real parts of systems explicitly modelled using non-directed connections between variables. This surprising shift has gone mostly unnoticed in the literature, even though no formal justification is provided.

There is ample evidence in the literature of this shift from model to target, which we might call a 'reification fallacy'. For instance, Allen and Friston (2018) begin 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)

It is possible to read this passage in an entirely instrumentalist way. That the boundary 'can be described' using a blanket merely suggests that the system can be modeled as having a blanket (see for instance Friston (2013); Palacios et al. (2020)). Without considering the further assumptions explained in Biehl et al. (2021) and Friston et al. (2021a), this notion of a Markov blanket is in line with the standard use of the notion introduced by Pearl and explained in the first part of this paper. However, Allen and Friston undermine this innocent instrumentalist reading 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).

In this passage a Markov blanket is taken to be either equivalent to, or identical with, a physical boundary in the world.[xi] 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 function is far removed from the initial auxiliary role played by Markov blankets in variational inference, where notions of temporal dynamics and system integrity do not come up. Instead, Markov blankets serve here as a real boundary between organism and world, i.e. what we are calling a 'Friston blanket'.

Many proponents of active inference now use the Markov blanket formalism in a much more metaphysically robust sense, one that does not simply follow from the formal details. Whereas the Pearl blankets discussed in the previous section are unambiguously part of the map (e.g. the graphical model), Friston blankets are best understood as parts of the territory (e.g. 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.

4.3 Ambiguous 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 being used as 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" (2013). In doing so we can see how what was once an abstract mathematical construct defined by conditional independences in graphical models (a Pearl blanket) came to be seen as an entity that somehow causes (or 'induces', or 'renders') conditional independence (a Friston blanket).[xii] This latter interpretation has potentially interesting philosophical implications, but does not follow directly from the former mathematical construct. Perhaps surprisingly, many authors in the field are seemingly not aware of this process of reification, leading to the conflation of several different kinds of boundaries in the literature: Markov blankets are characterized alternatively as statistical boundaries, spatial boundaries, ontological boundaries, or autopoietic boundaries, and each characterisation is treated as somehow equivalent to (and interchangeable with) the others.

Some authors are admittedly more careful, for example Clark (2017) makes sure to distinguish between the physical 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)

Here the discovery of a Markov blanket, perhaps only in our model of the system, serves to indicate the presence of "some kind of boundary" in the system itself. Clark holds that Markov blankets are discovered inside the modelling domain (what we call Pearl blankets), and that this discovery indicates the presence of something important ('some kind of boundary') in the target domain (perhaps a Friston blanket). While 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' in e.g., Clark and Chalmers, 1998).

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

> A Markov blanket defines the boundaries of a system (e.g., a cell or a multi-cellular organism) in a statistical sense. (p.1)

They also assume that this statement implies something much stronger, namely 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. (p.2)

However, the authors never explicate exactly how to conceive of a 'boundary in a statistical sense', perhaps indirectly relying on the inflated version of a Markov blanket proposed in Friston and Ao (2012) and Friston (2013).

Hohwy (2017) also equates the internal states identified by a 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)

Furthermore, Hohwy assumes that 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. (p.7)

In other words, Markov blankets 'epistemically seal-off' agents from their environment. In the same paper, Hohwy, like Allen and Friston above, equates 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. (p.4)

All of the above examples show how Markov blankets have moved from a rather simple statistical tool used for specifying a particular structure of conditional independence within a set of abstract random variables, to a specification of structures in the world that are said to 'cause' conditional independence, separate an organism from its environment, or epistemically seal off agents from their environment.[xiii] These characterizations would sound bizarre to the average computer scientist and statistician familiar only with the original Pearl blanket formulation (perhaps the only people commonly aware of Markov blankets before 2012 or 2013). 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. Conceptual issues with Friston blankets

So far, we have provided some initial analysis of both Pearl blankets and Friston blankets, demonstrating that they are used to answer different kinds of scientific and philosophical questions. Since these are different formal constructs with different metaphysical implications, the scientific credibility of Pearl blankets should not automatically be extended to Friston blankets. In this section, we focus on two conceptual issues with Friston blankets. These conceptual issues illustrate the kinds of problems that arise when using conditional independence as a tool to settle the kinds of philosophical questions that we saw Friston blankets being applied to in the previous section.

To bring these conceptual issues into full view, let us introduce a second toy example. 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 is 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 a simple Bayesian network, we get something like Fig. 7.

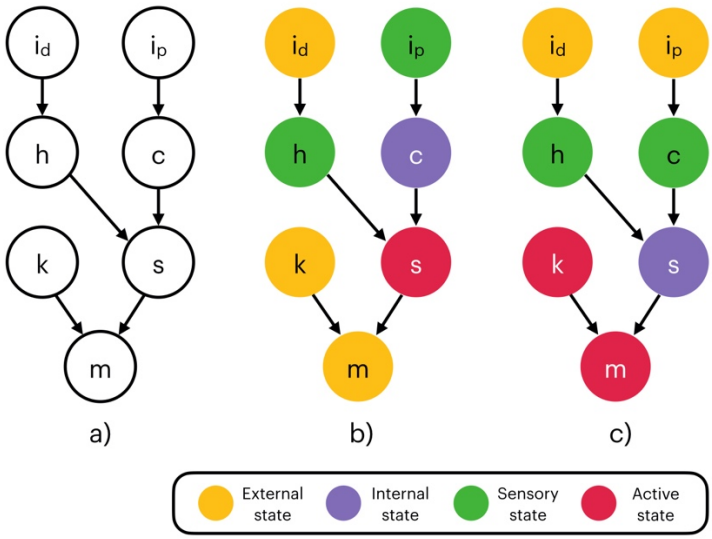

Figure 7: *Conditions leading up to the knee-jerk reflex.* On the left, a Bayesian network where i_d and i_p denote the motor intentions of the doctor and the patient respectively. Node s denotes the spinal neurons that are directly responsible for causing the kicking movement m Node h indicates a medical intervention with a hammer, while c stands for a motor command sent to s from the central nervous system. Finally, node k stands for a third way of moving the patient's leg, e.g., by someone else kicking it to move it mechanically. The middle (b) and the right figures (c) with the colored-in nodes show two different ways of partitioning the same network using a 'naive' Friston blanket with different choices of internal states, c and s respectively.

5.1 Counterintuitive sensorimotor boundaries

This simple network allows us to illustrate some problems with using Friston blankets to demarcate agents and their (sensorimotor) boundaries. The first problem concerns which role to attribute to 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).[xiv] This reflects the fact that the contraction of muscles isolated in the patellar reflex could also be the result of the patient's motor intentions. If we interpret the motor command c as an internal state of the patient, the spinal signal which causes the movement would be an active state. However, this leads to a puzzle about the way in which we should interpret h. Clearly, h is a co-parent of c and hence lies on its Friston blanket. According to the partition system used by (Friston, 2013, 2019; Friston et al. 2021b), h should fall into the Friston blanket of c as a sensory state (see Fig. 7b). But regardless of whether one assigns a sensory or active status to h, its inclusion in the Friston blanket of c is problematic. From a sensorimotor perspective[xv] (see Barandarian et al., 2009; Tishby et al., 2011), h is an environmental variable external to the organism. As such, the medical hammer h should not be identified as part of an active agent, or even attributed a rather generous role as part of its sensory interface with the world.

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. 7c, while the bodily movement m and the external kick *k* should be considered, in the language of Friston blankets, as active states. Notice, however, that this would not help in any way, since what we might think of as an *external* intervention k that could lead

to the same kind of bodily movement, is now part of the active states, while at the same time displaying the same formal properties as any putatively 'internal' cause of the movement (as the Bayesian network in Fig. 7 should make clear). This example exposes 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 the states into internal and external by means of a Markov blanket does not necessarily equate with the boundary between agent and environment found in sensorimotor loops, at least as these are intuitively or typically understood.

In other words, the co-parents of a child s in a Bayesian network include all other factors that could potentially cause, modulate or influence the occurrence of s. This puts pressure on the analogy between Markov blankets and sensorimotor boundaries on which Friston blankets are based. Including these co-parents in the Friston blanket will include states in the environment (like the doctor's hammer), forcing one to accept counterintuitive conclusions about the boundaries of an agent. Not including the co-parents, on the other hand, gives up on the idea that conditional independence and Markov blankets are the right kind of tools to delineate the boundaries of agents, calling into question the validity of the Friston blanket construct as a formal tool.

5.2 Conditional independence is model-relative

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.[xvi] 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. (p. 4)

As has been mentioned in the previous section, Bayesian models are often explicitly said to be instrumental tools that are not designed to develop a final and complete description of a system, but are rather best at capturing the dependencies between the element of a system and/or predicting its behavior, at a particular level of analysis (and relative to our current knowledge and resource constraints). What the 'right' Bayesian network is for the knee-jerk reaction might depend on the observed states that we are given, our background knowledge and assumptions, and more pragmatically, the problem we want to model, as well as the time and computational power that is at our disposal. Which, and how many, Markov blankets can be identified within this model will depend on all of these factors. This suggests that Bayesian networks are not the right kind of tool to delineate real ontological boundaries in a non-arbitrary way. Here we are talking about Bayesian models in general, but an important caveat is that Bayesian networks have been famously used as tools for decomposing physical systems. Importantly, however, such decomposition relies on treating the model as a map of the target system which is then used to direct interventions that can be modelled using Pearl's 'do-calculus' (Pearl

2009; cf. Woodward 2003). Such applications of Bayesian modelling rarely makes use of the Bayesian Occam's razor (mentioned in Section 2.4.), since the goal is not to predict the behavior of the system, but rather to depict how parts of the system influence each other.

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., *relevant* Friston blankets cannot simply be 'detected' in some objective way and then used to determine the boundary of a system.[xvii] 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, see Pearl 1988, or dynamic regimes at steady-state, see Friston et al. 2021a). 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, Wiese and Friston 2021; cf. Andrews 2020, Beni 2021, and Sánchez-Cañizares 2021 for some recent critical discussion).

While we do not want to try and solve all of these issues here, it is important to recognise 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 organisms, 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. We turn to this question in the next section, but it is important to note that even if a legitimate explanatory project can be defined for Friston blankets, the conceptual issues outlined in this section will also still need to be addressed.

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 as a tool for distinguishing 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 technical assumptions and of the general explanatory 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 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 Hipólito et al. (2021), 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. To give an initial example, Kirchhoff and Kiverstein (2021) 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)

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 metaphysically-laden Friston blanket which is supposed to be instantiated by some physical system. The ambition of the philosophical prospects of the Friston blanket construct is again made clear by Kirchhoff and Kiverstein (2021):

> 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 the boundaries of the mind cannot simply be assumed from the formal credibility of the original Pearl blanket construct. We should emphasise at this point that it is not only Kirchhoff and Kiverstein (2021) making this assumption, which is prevalent in much of the active inference literature that draws on Friston's (2013) "Life as we know it" paper discussed in section 4.1. In what follows we will consider the differences between the Pearl blanket and Friston blanket constructs in more detail, providing additional examples as we go.

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

We are now in a position to articulate what we perceive to be the central methodological difference between how the two notions of Markov blankets are applied in the literature. As we see it, applications of the two constructs should be understood as representing different research programs. The first, which we will call 'inference with a model', corresponds roughly to the use of Markov blankets (or Pearl blankets) described in Section 3 of this paper. The main thesis that drives this research program is that organisms perform variational inference to regulate perception and action. In doing so, they rely (implicitly or explicitly) on a model of their environment, which might feature something like Pearl blankets as an auxiliary statistical construct. The second research program, which we call 'inference within a model', constitutes the position we described in Section 4 of this paper, using Markov blankets (or Friston blankets) as a measure of the real ontological boundary between a system and its environment. The main thesis that drives this latter research programs is that living systems and their environments are dynamically coupled systems that can be represented using network models, and that modelling tools (like Markov blankets) can therefore be legitimately used to distinguish an agent from its environment. These are two very different projects, with different commitments,

aims, and tools (although both might fall broadly under the FEP framework). In the rest of this subsection we will briefly characterize both projects.

Inference with a model

As mentioned above, an important motivation for the free energy principle is the parallel between scientific inference and active inference. Like the scientist, the agent wants to know and control the states of some aspect of the world which remains hidden, while only having access to some limited set of observations. The agent can solve this problem by using a generative model of its environment. The agent uses (or appears to use) variational inference to obtain a recognition density which approximates the posterior density.

In model-based cognitive neuroscience, the two approaches have been stacked together. The explanatory project is to infer the details of the generative model an agent is using to infer the states of its environment. This seems to be one of the strongest potential empirical applications of the FEP and some of its related ideas, (Parr et al. 2019, Adams et al. 2013, and Pezzulo, Rigoli, and Friston 2018), and reflects a more general explanatory strategy in cognitive neuroscience (Lee and Mumford 2003, Rao and Ballard 1999). Although perhaps not directly empirically refutable (cf. Andrews 2020), this approach guides an active research program, whose quality will eventually determine its overall viability.

As we highlighted in section 3.1, Pearl blankets play an auxiliary role in projects of this kind. They describe conditional independence on random variables (represented for instance in Bayesian networks), and are not a literal feature of either the agent or its environment (or indeed, the boundary between the two). There has been some discussion of the status of the theoretical posits of this kind of research. Do agents really possess a model of their environment, or are they merely usefully *modeled* as such? These questions about realism and instrumentalism of cognitive constructs are interesting and have been extensively discussed in the recent literature on active inference (Colombo and Seriès, 2012; Ramstead, Kirchoff, and Friston 2020; Ramstead, Friston, and Hipólito 2020; van Es 2020), but

these discussions are not our main focus. The framing of the agent as a modeller of its environment has also led to an important but rather long-winded debate about whether, and in what sense, free energy minimizing agents should be seen as utilizing generative models as representations of their environment (Gładziejewski 2016; Clark 2015a, 2015b; Dołęga 2015; Kiefer and Hohwy 2018, Kirchhoff and Robertson 2018; Williams, 2018). Here we merely point out that this debate also allows for taking an instrumentalist or realist stance and, more importantly, that it is orthogonal to the distinction between inference with a model and inference within a model.

One complicating factor that is worth mentioning here is a potential disanalogy between scientific inference and active inference. In scientific inference, a scientist literally uses a model to make inferences out of observed data. The model itself is inert when not being used by an intentional agent. The same does not go for active inference. The agent does not *have* a model of its environment that it uses to perform inference, but rather the agent *is* a model of its environment (Friston 2013; 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 both scientist and model. Considerations of this kind have led some theorists to turn towards a different (and perhaps more ambitious) explanatory project, where Markov blankets also come to be seen as a literal part of the physical systems being studied.

Inference within a model

The 'primordial soup simulation' that we presented in Section 4.2 suggests a very different research direction for the 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 as the dynamics of the system reach equilibrium. Agent and environment are separated by each other through a Friston blanket. The Markov blanket formalism has subsequently been presented as not just being able to identify the boundaries of agents, but also of any supposedly self-organising system, including species (Ramstead et al. 2019) and biospheres (Rubin et al. 2020).

One could see the primordial soup simulation as an interesting toy model to investigate the emergence of sensorimotor boundaries in a highly idealized domain. This has long been a successful strategy in complex systems research. For example, Conway's Game of Life (Gardner, 1970) has been used to formalize concepts such as autopoiesis (Beer, 2004, 2014, 2020). Such toy models come with strong explanatory power but also forthright metaphysical modesty: they do not claim to directly model or capture real world phenomena. They are merely used as demonstrations of how certain concepts or principles could play out in a simplified system. This, however, is very different from how most active inference theorists frame their work, as we will now see.

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)

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. In order for the formalism to do such heavy metaphysical lifting, however, additional premises need to be in place. After all, cognitive systems (or other systems whose boundaries we might be interested in) exist in the physical world, while the original Markov blanket formalism operates on abstract mathematical entities. Hence, the question for proponents of the more ambitious FEP project is: how can the two kinds of entities map onto each other, such that conclusions about the boundaries of cognitive systems can be drawn based on the mathematical framework?

As we have hinted at before, there are three strategies available to the FEP theorist who wants to use Markov blankets in this way: a literalist, realist, and an instrumentalist one. The literalist position is roughly equivalent to the claim that the world just *is* a network consisting of interacting systems, which are themselves more fine-grained probabilistic networks, and so on, and this is why the Friston blanket formalism works as a way to demarcate real world boundaries. The realist position is still committed to the claim that Friston blankes do pick out real boundaries in the world, but they are taken to be representations of worldly features, rather than literally *being* such features themselves. Finally, the instrumentalist position holds that the world can merely be usefully modeled as a Bayesian network, and that this justifies using the Pearl blanket formalism as a guide to worldly boundaries. We think that both the literalist and realist positions have similar problems, while the instrumentalist position is less problematic but also less interesting. We will discuss each position in turn.

The literalist position entails that the mathematical structures posited by the FEP are not merely a map of self-organizing systems, but are themselves the territory (cf. Andrews 2020). In this case, the FEP framework might constitute something like a 'blanket-oriented ontology' (BOO): a view in which reality consists of a number of hierarchically nested Friston blankets. This might be an appealing picture for some, but it is certainly not something that can be simply read off the formalism itself. Rather, it is an additional assumption that must be explicitly stated and argued for. In a recent paper, Menary and Gillett (2020) point out the strong Platonist and Pythagorean attitude that would be necessary in order to motivate this kind of ontology. Such an approach is not without allure and could be made philosophically interesting, but it would certainly not be metaphysically agnostic. The FEP and Friston blankets would serve as a starting assumption of such an ontological project, rather than its end goal. At any rate, the resulting approach would be quite far removed from the empirical and naturalistic research program that FEP purports to be, and would certainly involve answering "ontological questions in an a priori manner" (Ramstead et al. 2019, p.3).

At first sight, the realist alternative might look less objectionable. Conclusions can be drawn about real-world systems because there is a systematic mapping between reality and our mathematical descriptions of reality in terms of Bayesian networks. After all, it is relatively easy to find some mapping between a given target and the assumed model domain. However, the difficulty lies in finding a non-arbitrary mapping that is privileged for principled reasons. In the literature on Bayesian inference, the gold standard for establishing what the right kind of model is for a given target domain is Bayesian model selection. This requires a set of observations which is then used to select the most parsimonious explanatory model of these observations (see Section 2.4). In turn, Friston blankets can be understood only relative to such a model (see Section 5.2). The puzzle then is that if one wants to use the Markov blanket formalism to demarcate the boundaries of, e.g., a cognitive agent, one needs to already have a principled justification for why to start from one particular model rather than a different one, at which point it is not clear that the Markov blanket formalism is doing much additional work.

Some authors have followed this path and advocated for the realist position by claiming that it is not the Markov blanket formalism alone, but rather the Markov blanket formalism plus the free energy principle, that provides the relevant demarcations of agent-environment boundaries. Only those Markov blankets that demarcate free energy minimizing systems (or the systems which minimize the most free-energy, see Hohwy 2016) can be taken to represent the boundaries of living or cognitive systems. This defense of Friston blankets might look appealing at first, but faces a serious obstacle by assuming that free-energy minimizing systems can be identified without the help of the assumptions behind the Friston blanket construct, such as the existence of unambiguously active or passive states. This is a problem because, as it turns out, it is not that difficult to characterize all sorts of systems as free energy minimizing systems. For example, Baltieri, Buckley, and Bruineberg (2020) show that even the humble Watt governor can be analysed as a free energy minimizing system. Elsewhere, Rubin et al. (2020) have proposed modelling the Earth's climate system as the planet's own Friston blanket, while Parr (2021) uses Friston blankets to model enzymatic reactions in biochemical networks. What these examples show is that the scope of the free energy formula is so broad that it is inadequate to pick out

only living or cognitive systems. One could bite the bullet and claim that planets and Watt governors are cognitive systems, but this would be a surprising result and few would be on board with such radical assumptions. Finally, as we saw in Section 2, the free energy principle already assumes a mathematical structure to be in place (be it a random dynamical system or a Bayesian network). Therefore, in and of itself, the free energy principle has nothing to say about how these mathematical structures should be mapped onto physical structures.

All of the above suggest that Bayesian networks are not the right kind of tools to delineate real world boundaries in an objective and non-question-begging way. Perhaps ultimately these problems are resolvable, but as far as we know, nearly no-one in the literature has thus far paid any attention to them (for a refreshing exception see Biehl 2017, and some of the references therein). These considerations have led some authors behind the more recent active inference literature to embrace instrumentalism about the whole framework, not just the Friston blanket construct. Some have suggested that the active inference framework should subscribe to a fundamentally instrumentalist approach to scientific investigation, such that the use of Markov blankets to demarcate organism-environment boundaries should be understand just as another feature of our (scientific) models, rather than making any ontological claims about the structure of the world (see e.g. Andrews 2020; Colombo, Elkin, and Hartmann 2018; Ramstead, Kirchoff, and Friston 2020; Ramstead, Friston, and Hipólito, 2020; van Es 2020). This kind of global scientific instrumentalism is fine so far as it goes, and of course has precedents elsewhere in the philosophical debates about scientific realism (see e.g. Chakravartty 2017 for a helpful overview), but we do not think that it is reflective of the attitude that most scientists (or even philosophers) take towards the kinds of claims being made about Friston blankets in the active inference literature. Such global instrumentalism definitely does not sit well with the blanket-oriented ontology described above, and seems to be incompatible with understanding FEP as providing a "formal ontology" (Ramstead et al., 2019). Nonetheless, we are happy to settle for a conditional conclusion here: insofar as one is a scientific realist, and treats the seemingly ontological claims made about Friston blankets in a realist manner, then some further metaphysical assumptions are needed in order to warrant these claims.

7. Conclusion

Despite all of the issues and ambiguities pointed out in our above treatment, the free energy principle and active inference framework have considerable following in the fields of neuroscience and biology, due in part to ambitious claims regarding their unificatory potential (Friston 2010; Friston 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 conceptual tool that could settle centuries-long disputes about the relationship between mind and life (Clark 2013, 2015a, 2020; Hohwy 2013; Friston, Wiese, and Hobson 2020). At the same time, different parts of the framework, have raised some important, and in some cases yet-to-be-answered, scientific and philosophical problems. Some of these problems has to do with the capacity of the framework to account for traditional folk psychological distinctions between belief and desire (see e.g., Dewhurst, 2017; 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, Clark 2020). Another, very common, kind of critique is that the framework either does not enjoy any empirical support, or that the FEP is empirically inadequate (Colombo and Wright 2018; Williams 2020; Colombo and Palacios 2021), and should therefore be considered to offer, at best, a redescription of 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 2021; Williams 2020; Colombo and Palacios 2021), or that it presents a false equivocation between probability and adaptive value (Colombo 2020). Other works, such as Di Paolo et al. (2021) and Raja et al. (2021) have recently disputed claims about the FEP representing a *general* unifying principle, claiming that it fails to account for different sensorimotor aspects of embodied and (autopoietic) enactive cognition

More relevant for what we have discussed here, 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. Finally, various technical aspects of the FEP are now under scrutiny in works such as Rosas et al. (2020), Biehl et al. (2021) and Aguilera et al. (2021). Rosas et al. (2020) define a new object, a 'causal blanket', based on ideas from computational mechanics, in an attempt to overcome assumptions about Langevin dynamics in a stationary/steady-state regime. Biehl et al. (2021) cast doubts on the inconsistent mathematical treatment of Markov blankets over the years, partially acknowledged by Friston et al. (2021a) who now address such differences and specifies new and more detailed constraints for a cohesive treatment of Markov blankets in the FEP (see endnote x). Aguilera et al. (2021), on the other hand, question the relevance of the FEP for sensorimotor accounts of living systems, given some of its assumptions and in particular the descprition of agents' behaviour in terms of free energy gradients on ensemble averages of trajectories, claiming that (under the mathematical assumptions presented in their paper) these "free energy gradients [are] uninformative about the behaviour of an agent or its specific trajectories" (see also Di Paolo et al. 2021 for a similar conceptual point, and Da Costa et al. 2021 and Parr et al. 2021 for possible counterarguments).

These latter 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 they are deployed to support metaphysical claims made by the free energy principle, especially in the context of active agents and living organisms. After having been initially developed 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), Markov blankets have been claimed by proponents of the free energy principle to clarify the boundaries of the mind (Hohwy 2017; Clark 2017; Kirchhoff and Kiverstein 2021), of living systems (Friston 2013; Kirchhoff 2018; Kirchhoff et al. 2018), and even of social systems (Ramstead, Badcock, and Friston 2018; Veissière et

al. 2020; Rubin et al. 2020; Fox, 2021). Interestingly, in these papers a system gets defined in terms of relations of independence made within a Bayesian network. 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 itself is taken to be a Bayesian network, with the Markov blankets defining what it is to be a 'thing' within this world (Friston 2013; Kirchhoff et al. 2018; Friston 2019; Hipólito et al. 2020). We then raised some possible issues with this approach, namely the question of whether Bayesian networks are merely an instrumental modelling tool for the free energy principle framework, or whether the framework presupposes some kind of more fundamental Bayesian graphical ontology.

All of this points towards a fundamental dilemma for anyone interested in using Markov blankets to make substantial philosophical claims about biological and cognitive systems (which is what we take proponents of the free energy principle to be wanting to do). On the one hand, Markov blankets can be used in their original Pearl blanket guise, as a formal mathematical construct for performing inference on a generative model. This usage is philosophically innocent but cannot, without further assumptions that need to be explicitly stated, justify the kinds of conclusions that it is sometimes used for in the FEP literature (see e.g. Hohwy 2017; Kirchhoff et al. 2018; Kirchhoff and Kiverstein 2021). On the other hand, Markov blankets can be used in a more ontologically robust fashion, as what we have called Friston blankets, to demarcate actual worldly boundaries. 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; Parr, Da Costa, and Friston 2020) and philosophical (Ramstead, Badcock, and Friston 2018; Friston, Wiese, and Hobson 2020; Hipólito 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 or organism to perform inference. In inference within a model the scientist disappears from the scene, becoming a mere spectator of the inference show unfolding before their eyes. Here the Friston blanket specifies the anatomy of the target system: it is a formalization of the boundary between this system and its environment.

Ultimately, the considerations presented in this paper leaves the FEP theorist with a choice. One can accept a rather technical and innocent conception of Markov blankets as an auxiliary formal concept that define what nodes are relevant for variational inference. This conception is admittedly scientifically useful but has not yet lead to any philosophically interesting conclusions about the nature of life or cognition. Alternatively, 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 agents and their environment. Such a strong realist reading cannot be justified by just 'doing the maths', but rather needs to be independently argued for, and no such argument has yet been offered.

Acknowledgments

The authors would like to thank Micah Allen, Mel Andrews, Martin Biehl, Daniel Dennett, Kevin Flowers, Hajo Greif, Julian Kiverstein, Richard Menary, Thomas Parr, Nina Poth, Maxwell Ramstead, Fernando Rosas, Matthew Sims, Filippo Torresan, Wanja Wiese, Tobias Schlicht and members of his research group, Marcin Miłkowski and members of his research group, and the Active Inference Lab for insightful and critical discussions and timely feedback on previous versions of the manuscript. The authors also like to thank the editor and eight reviewers for their time and effort. The manuscript has benefited enormously from their critical reports.

JB is funded by a Macquarie Research Fellowship. KD's work is funded by the Volkswagen Stiftung

grant no. 87 105. MB is a JSPS International Research Fellow supported by a KAKENHI Grant-in-Aid for Scientific Research (No. JP19F19809).

Conflicts of interest: none

References

- Adams, R. A., Stephan, K., Brown, H., Frith, C., and Friston, K. J. (2013). The computational anatomy of psychosis. *Frontiers in Psychiatry*, 4: 47.
- Aguilera, M., Millidge, B., Tschantz, A., & Buckley, C. L. (2021). How particular is the physics of the Free Energy Principle?. *arXiv preprint arXiv:2105.11203*.
- Allen, M. and Friston, K. J. (2018). From cognitivism to autopoiesis: towards a computational framework for the embodied mind. *Synthese*, 195(6): 2459–2482.
- Andrews, M. (2020). The Math is not the Territory: Navigating the Free Energy Principle. [Preprint] http://philsci-archive.pitt.edu/18315.
- Anderson, M. L. (2017). Of bayes and bullets: An embodied, situated, targeting-based account of predictive processing. In Wiese,W. and Metzinger, T. K., (Eds.), *Philosophy and predictive processing*, 4. Frankfurt am Main, Germany: MINDGroup.
- Attias, H. (2003). Planning by probabilistic inference. In C. M. Bishop and B. J. Frey, (Eds.), *Proc. of the 9th Int. Workshop on Artificial Intelligence and Statistics, 2003*.
- Baltieri, M. and Buckley, C. L. (2019). Generative models as parsimonious descriptions of sensorimotor loops. *Behavioral and Brain Sciences*, 42: e218.
- Baltieri, M., Buckley, C. L., and Bruineberg, J. (2020). Predictions in the eye of the beholder: an active inference account of Watt governors. In *Artificial Life Conference Proceedings* (pp. 121-129). Cambridge, MA: MIT Press.
- Barandiaran, X. E., Di Paolo, E., & Rohde, M. (2009). Defining agency: Individuality, normativity, asymmetry, and spatio-temporality in action. *Adaptive Behavior*, *17*(5), 367-386.
- Beal, M. J. (2003). *Variational algorithms for approximate Bayesian inference.* Doctoral dissertation, UCL (University College London).
- Beer, R. D. (2004). Autopoiesis and cognition in the game of life. *Artificial Life*, 10(3): 309–326.
- Beer, R. D. (2014). The cognitive domain of a glider in the game of life. *Artificial life*, 20(2): 183–206.
- Beer, R. D. (2020). An investigation into the origin of autopoiesis. *Artificial Life*, 26(1): 5–22.
- Beni, M.D. (2021). A critical analysis of Markovian monism. *Synthese*: https://doi.org/10.1007/s11229-021-03075-x.
- Biehl, M. (2017). *Formal approaches to a definition of agents*. Doctoral dissertation, University of Hertfordshire.
- Biehl, M., Guckelsberger, C., Salge, C., Smith, S. C., and Polani, D. (2018). Expanding the active inference landscape: More intrinsic motivations in the perception-action loop. *Frontiers in Neurorobotics*, 12: 45.
- Biehl, M., Pollock, F. A., and Kanai, R. (2021). A Technical Critique of Some Parts of the Free Energy Principle. *Entropy*, *23*(3), 293.
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer-Verlag New York.
- Blei, D.M., Kucukelbir, A., and McAuliffe, J.D. (2017). Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112:518, 859-877
- Boik, J. C. (2021). Science-Driven Societal Transformation, Part III: Design. *Sustainability*, 13(2), 726.
- Bogacz, R. (2017). A tutorial on the free energy framework for modelling perception and learning. *Journal of mathematical psychology*, 76: 198–211.
- Bruineberg, J., Kiverstein, J., and Rietveld, E. (2018). The anticipating brain is not a scientist: the free energy principle from an ecological-enactive perspective. *Synthese*, 195(6): 2417–2444.
- Buckley, C. L., Kim, C. S., McGregor, S., and Seth, A. K. (2017). The free energy principle for action and perception: A mathematical review. *Journal of Mathematical Psychology,* 14: 55–79.

- Cao, R. 2020. “New Labels for Old Ideas: Predictive Processing and the Interpretation of Neural Signals.” Review of Philosophy and Psychology 11 (3): 517–46
- Ciaunica, A., Constant, A., Preissl, H., and Fotopoulou, K. (2021). The first prior: From co-embodiment to co-homeostasis in early life. *Consciousness and Cognition*, *91*, 103117.
- Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and BrainSciences*, 36(03): 181–204.
- Clark, A. (2015a). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press.
- Clark, A. (2015b). Radical predictive processing. *The Southern Journal of Philosophy*, 53: 3–27.
- Clark, A. (2017). How to knit your own markov blanket. In Metzinger, T. K. and Wiese, W., (Eds.), *Philosophy and predictive processing*: 3. Open MIND. Frankfurt am Main: MIND Group.
- Clark, A. (2020). Beyond desire? Agency, choice, and the predictive mind. *Australasian Journal of Philosophy* 98(1): 1-15.
- Clark, A. and Chalmers, D. (1998). The extended mind. *Analysis*, 58(1):7–19.
- Chakravartty, A. (2017). Scientific Realism. *The Stanford Encyclopedia of Philosophy* (Summer 2017 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/sum2017/entries/scientific-realism/.
- Colombo, M. (2020). Maladaptive social norms, cultural progress, and the free-energy principle. *Behavioral and Brain Sciences* 43: e100.
- Colombo, M., Elkin, L., and Hartmann, S. (2018). Being Realist about Bayes, and the Predictive Processing Theory of Mind. *The British Journal for Philosophy of Science*, 72(1).
- Colombo, M., and Palacios, P. (2021). Non-equilibrium thermodynamics and the free energy principle in biology. *Biology & Philosophy*, *36*(5), 1-26.
- Colombo, M., and Seriès, P. (2012). Bayes in the brain—On Bayesian modelling in neuroscience. *The British Journal for the Philosophy of Science*, 63: 697–723.
- Colombo, M. and Wright, C. (2021). First principles in the life sciences: the free-energy principle, organicism, and mechanism. *Synthese* 198(14): 3463-3488.
- Da Costa, L., Parr, T., Sajid, N., Veselic, S., Neacsu, V., & Friston, K. (2020). Active inference on discrete state-spaces: a synthesis. *Journal of Mathematical Psychology*, *99*, 102447.
- Daunizeau, J. (2017). The variational laplace approach to approximate Bayesian inference. [preprint] arXiv:1703.02089.
- Dayan, P., Hinton, G. E., Neal, R. M., and Zemel, R. S. (1995). The Helmholtz Machine. *Neural computation*, 7(5): 889–904.
- Dewhurst, J. (2017). Folk Psychology and the Bayesian Brain. In *Philosophy and Predictive Processing*. Frankfurt am Main: MIND Group.
- Di Paolo, E., Thompson, E., & Beer, R. D. (2021). Laying down a forking path: Incompatibilities between enaction and the free energy principle.
- Dołęga, K. (2017). Moderate Predictive Processing. In Thomas K. Metzinger and Wanja Wiese (Eds.), *Philosophy and Predictive Processin*g: 10. MIND Group, Frankfurt am Main.
- Fausto-Sterling, A. (2021). A Dynamic Systems Framework for Gender/Sex Development: From Sensory Input in Infancy to Subjective Certainty in Toddlerhood. *Frontiers in Human Neuroscience*, *15*, 150.
- Feldman, H. and Friston, K. J. (2010). Attention, uncertainty, and free energy. *Frontiers in human neuroscience*, 4:215.
- Fox, S. (2021). Active Inference: Applicability to Different Types of Social Organization Explained through Reference to Industrial Engineering and Quality Management. *Entropy*, 23(2): 198.
- Friston, K. J. (2005). A theory of cortical responses. *Philosophical transactions of the Royal Society of London. Series B, Biologicalsciences*, 360(1456): 815–836.
- Friston, K. J. (2008). Hierarchical models in the brain. *PLoS Computational Biology*, 4(11).

- Friston, K. J. (2010). The free energy principle: a unified brain theory? *Nature reviews. Neuroscience*, 11(2):127–138.
- Friston, K.J. (2012). A Free Energy Principle for Biological Systems. *Entropy* 2012, 14, 2100-2121.
- Friston, K. J. (2013). Life as we know it. *Journal of the Royal Society Interface*, 10(86): 20130475.
- Friston, K. J. (2019). A free energy principle for a particular physics. [preprint] arXiv:1906.10184.
- Friston, K. J., Harrison, L., and Penny, W. (2003). Dynamic causal modelling. *Neuroimage*, 19.4: 1273-1302.
- Friston, K. J., Kilner, J., and Harrison, L. (2006). A free energy principle for the brain. *Journal of Physiology-Paris*, 100(1): 70–87.
- Friston, K. J., Mattout, J., Trujillo-Barreto, N., Ashburner, J., and Penny, W. (2007). Variational free energy and the Laplace approximation. *Neuroimage*, 34(1): 220–234.
- Friston, K. J., Trujillo-Barreto, N., and Daunizeau, J. (2008). DEM: A variational treatment of dynamic systems. *NeuroImage*, 41(3): 849–885.
- Friston, K. J., Daunizeau, J., Kilner, J., and Kiebel, S. J. (2010). Action and behavior: A free energy formulation. *Biological Cybernetics*, 102(3): 227–260.
- Friston, K. J. and Ao, P. (2012). Free energy, value, and attractors. *Computational and Mathematical Methods in Medicine,* Volume 2012, Article ID 937860.
- Friston, K., Sengupta, B., and Auletta, G. (2014). Cognitive dynamics: From attractors to active inference. *Proceedings of the IEEE*, 102(4): 427–445.
- Friston, K. J., Levin, M., Sengupta, B., and Pezzulo, G. (2015a). Knowing one's place: a free energy approach to pattern regulation. *Journal of The Royal Society Interface*, 12(105): 20141383.
- Friston, K. J., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., and Pezzulo, G. (2015b). Active inference and epistemic value. *Cognitive neuroscience*, 6(4): 187-214.
- Friston, K. J., FitzGerald, T., Rigoli, F., Schwartenbeck, P., and Pezzulo, G. (2017). Active inference: a process theory. *Neural Computation*, 29(1):1–49.
- Friston, K.J., Parr., T., and de Vries, B. (2017). The graphical brain: Belief propagation and active inference. *Network Neuroscience*, 1(4): 381-414.
- Friston, K. J., Wiese, W., and Hobson, J. A. (2020). Sentience and the origins of consciousness: From cartesian duality to markovian monism. *Entropy*, 22(5): 516.
- Friston, K.J.; Da Costa, L.; Parr, T. (2021a). Some Interesting Observations on the Free Energy Principle. *Entropy* 2021, *23*, 1076. https://doi.org/10.3390/e23081076
- Friston, K. J., Fagerholm, E. D., Zarghami, T. S., Parr, T., Hipólito, I., Magrou, L., and Razi, A. (2021b). Parcels and particles: Markov blankets in the brain. *Network Neuroscience*, 5(1): 211-251
- Gardner, M. (1970). Mathematical Games: The fantastic combinations of John Conway's new solitaire game "Life". *Scientific American*, 223: 120–123.
- Gładziejewski, P. (2016). Predictive coding and representationalism. *Synthese*, 193(2): 559–582.
- Gregory, R. L. (1980). Perceptions as hypotheses. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 290(1038): 181-197.
- Griffiths, T. L., and Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychol Sci.*, 17(9): 767-73.
- Grossberg, S. (1980). How does a brain build a cognitive code? *Psychological Review*, 87(1): 1–51.
- Hafner, V. V., Loviken, P., Villalpando, A. P., and Schillaci, G. (2020). Prerequisites for an artificial self. *Frontiers in Neurorobotics*, 14.
- Hesp, C., Ramstead, M., Constant, A., Badcock, P., Kirchhoff, M., and Friston, K. (2019). A multi-scale view of the emergent complexity of life: A free energy proposal. In *Evolution, Development and Complexity*. Springer, Cham.

- Hinton, G. E. and Zemel, R. S. (1994). Autoencoders, minimum description length, and Helmholtz free energy. In *Advances in Neural Information Processing Systems 6*. San Mateo, CA: Morgan Kaufmann.
- Hipólito, I., Ramstead, M. J. D., Convertino, L., Bhat, A., Friston, K. J., Parr, T. (2021). Markov blankets in the brain, *Neuroscience and Biobehavioral Reviews*, 125: 88-97.
- Hohwy, J. (2013). *The Predictive Mind*. Oxford: OUP.
- Hohwy, J. (2016). The self-evidencing brain. *Noûs*, 50(2): 259–285.
- Hohwy, J. (2017). How to entrain your evil demon. In Metzinger, T. K. and Wiese, W., (Eds.), *Philosophy and predictive processing*: 2. Open MIND. Frankfurt am Main: MIND Group.
- Jefferys, W. H., and Berger, J. O. (1991). Sharpening Occam's Razor on a Bayesian strop. *Bulletin of the Astronomical Society*, 23(3), 1259.
- Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine learning*, 37(2): 183–233.
- Kappen, H. J., Gómez, V., and Opper, M. (2012). Optimal control as a graphical model inference problem. *Machine learning*, 87(2): 159–182.
- Khezri, D. B. (2021). *Free energy Governance-Sensing, Sensemaking, and Strategic Renewal-Surprise-Minimization and Firm Survival*. Doctoral dissertation, Universität St. Gallen.
- Kiefer A., and Hohwy, J. (2018). Content and misrepresentation in hierarchical generative models. *Synthese*, 195: 2387–2415.
- Kirchhoff, M. D. (2018). Autopoiesis, free energy, and the life–mind continuity thesis. *Synthese*, 195(6): 2519–2540.
- Kirchhoff, M. D., Parr, T., Palacios, E., Friston, K. J., and Kiverstein, J. (2018). The markov blankets of life: autonomy, active inference and the free energy principle. *Journal of The Royal Society Interface*, 15(138): 20170792.
- Kirchhoff, M. D., and Robertson, I. (2018). Enactivism and predictive processing: a non-representational view. *Philosophical Explorations*, 21(2): 264-281.
- Kirchhoff, M. D. and Kiverstein, J. (2021). How to determine the boundaries of the mind: a markov blanket proposal. *Synthese*, https://doi.org/10.1007/s11229-019-02370-y.
- Kirchhoff, M.D., van Es, T. (2021). A universal ethology challenge to the free energy principle: species of inference and good regulators. *Biology & Philosophy* 36 : 8.
- Kiverstein, J., Kirchhoff, M., and Thacker, M. (2021). Why Pain Experience is not a Controlled Hallucination of the Body. [preprint] http://philsci-archive.pitt.edu/18770/
- Klein, C. 2018. "What Do Predictive Coders Want?" *Synthese* 195 (6): 2541–57.
- Knill, D. C. and Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, 27(12): 712–719.
- Knill, D. C. and Richards, W. (1996). *Perception as Bayesian inference*. Cambridge University Press.
- Körding, K. and Wolpert, D. (2004). Bayesian integration in sensorimotor learning. *Nature*, 427: 244–247.
- Kuchling, F., Friston K. J., Georgiev G., and Levin M. (2020). Morphogenesis as Bayesian Inference: A Variational Approach to Pattern Formation and Control in Complex Biological Systems. *Phys Life Rev* 33: 88–108.
- Lee, T. S. and Mumford, D. (2003). Hierarchical bayesian inference in the visual cortex. *JOSA A*, 20(7): 1434–1448.
- Liwtin, P., and Miłkowski, M. (2020). Unification by Fiat: Arrested Development of Predictive Processing. *Cognitive Science* 44.
- MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge, UK: CUP.
- Maturana, H. R., & Varela, F. J. (1972). *Autopoiesis and cognition: The realization of the living* (Vol. 42). Springer Science & Business Media.

- Menary, R. and Gillett, A. J. (2020). Are Markov blankets real and does it matter? In Mendon ca, D., Curado, M., and Gouveia, S. S., editors, *The Philosophy and Science of Predictive Processing.* Bloomsbury Academic.
- Millidge, B., Tschantz, A., Seth, A. K., & Buckley, C. L. (2020). On the relationship between active inference and control as inference. In *International Workshop on Active Inference* (pp. 3-11). Springer, Cham.
- Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.
- Oaksford, M., and Chater, N. (2001). The probabilistic approach to human reasoning. *Trends in Cognitive Sciences*, 5(8): 349-357.
- Opper, M. and Archambeau, C. (2009). The variational Gaussian approximation revisited. *Neural computation*, 21(3):786–792.
- Palacios, E. R., Razi, A., Parr, T., Kirchhoff, M., & Friston, K. (2020). On Markov blankets and hierarchical self-organisation. *Journal of theoretical biology*, *486*, 110089.
- Parisi, G. (1988). *Statistical field theory*. Addison-Wesley.
- Parr, T. (2021). Message Passing and Metabolism. *Entropy*, *23*(5), 606.
- Parr, T., Mirza, M. B., Cagnan, H., and Friston, K. J. (2019). Dynamic causal modelling of active vision. *Journal of Neuroscience*, 39(32): 6265–6275.
- Parr, T., Da Costa, L., & Friston, K. (2020). Markov blankets, information geometry and stochastic thermodynamics. *Philosophical Transactions of the Royal Society A*, *378*(2164), 20190159.
- Parr, T., Da Costa, L., Heins, C., Ramstead, M. J. D., & Friston, K. J. (2021). Memory and Markov Blankets. *Entropy*, *23*(9), 1105.
- Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.
- Pearl, J. (2009). *Causality*. Cambridge university press.
- Penny, W. D., Friston, K. J., Ashburner, J., Kiebel, S., Nichols, T. (Eds.) (2011). *Statistical parametric mapping: the analysis of functional brain image*s. Elsevier.
- Pezzulo, G., Rigoli, F., and Friston, K. J. (2018). Hierarchical active inference: a theory of motivated control. *Trends in CognitiveSciences*, 22(4): 294–306.
- Poirier, P., Faucher, L., & Bourdon, J. N. (2021). Cultural blankets: Epistemological pluralism in the evolutionary epistemology of mechanisms. *Journal for General Philosophy of Science*, *52*(2), 335-350.
- Raja, V., Valluri, D., Baggs, E., Chemero, A., & Anderson, M. L. (2021). The markov blanket trick: On the scope of the free energy principle and active inference.
- Ramstead, M. J. D., Badcock, P. B., and Friston, K. J. (2018). Answering schrödinger's question: A free energy formulation. *Physics of life reviews*, 24:1–16.
- Ramstead, M. J., Kirchhoff, M. D., Constant, A., and Friston, K. J. (2019). Multiscale integration: beyond internalism and externalism. *Synthese*, 198:41–70.
- Ramstead, M. J., Friston, K. J., and Hipólito, I. (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 et al. (2020a)
- Ramstead, M. J., Kirchhoff, M. D., and Friston, K. J. (2020). A tale of two densities: Active inference is enactive inference. *Adaptive Behavior*, *28*(4), 225-239. – Ramstead et al. (2020b)
- Ramstead, M. J., Hesp, C., Tschantz, A., Smith, R., Constant, A., & Friston, K. (2020). Neural and phenotypic representation under the free-energy principle. *Neuroscience & Biobehavioral Reviews*. – Ramstead et al. (2020c)
- Rao, R. P. and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1): 79–87.
- Rosas, F. E., Mediano, P. A., Biehl, M., Chandaria, S., & Polani, D. (2020). Causal blankets: Theory and algorithmic framework. In *International Workshop on Active Inference*. Springer, Cham.
- Rubin, S., Parr, T., Da Costa, L., and Friston, K. J. (2020). Future climates: Markov blankets and active inference in the biosphere. *Journal of the Royal Society Interface,* 17: 20200503.

- Sajid, N., Ball, P. J., Parr, T., and Friston, K. J. (2021). Active inference: demystified and compared. *Neural Computation,* 33(3): 674-712.
- Sánchez-Cañizares J. The Free Energy Principle: Good Science and Questionable Philosophy in a Grand Unifying Theory. *Entropy*. 2021; 23(2): 238. https://doi.org/10.3390/e23020238.
- Seth, A., Millidge, B., Buckley, C. L., and Tschantz, A. (2020). Curious inferences: Reply to Sun and Firestone on the dark room problem. *Trends in Cognitive Sciences*, 24(9): 681–683.
- Sims, M. (2020). How to count biological minds: symbiosis, the free energy principle, and reciprocal multiscale integration. *Synthese*. https://doi.org/10.1007/s11229-020-02876-w.
- Stephan, K. E., Penny, D., Daunizeau, W. J., Moran, R. J., and Friston, K. J. (2009). Bayesian model selection for group studies. *NeuroImage*, 46(4): 1004–1017.
- Stephan, K. E., Penny, W. D., Moran, R. J., den Ouden, H. E., Daunizeau, J., & Friston, K. J. (2010). Ten simple rules for dynamic causal modeling. *Neuroimage*, *49*(4), 3099-3109.
- Sun, Z. and Firestone, C. (2020a). The dark room problem. *Trends Cogn. Sci.*, 24: 346–348.
- Sun, Z. and Firestone, C. (2020b). Optimism and pessimism in the predictive brain. *Trends Cogn. Sci.*, 24: 683–685.
- Tenenbaum, J. B., Kemp, C., Griffiths, T. L., Goodman, N. D. (2011). How to Grow a Mind: Statistics, Structure, and Abstraction. *Science*, 331(6022): 1279-1285.
- Tishby, N., & Polani, D. (2011). Information theory of decisions and actions. In *Perception-action cycle* (pp. 601-636). Springer, New York, NY.
- Tschantz, A., Seth, A. K., and Buckley, C. L. (2020). Learning action-oriented models through active inference. *PLOS Computational Biology*, 16(4):e1007805.
- Van de Cruys, S., J. Friston, K., and Clark, A. (2020). Controlled optimism: Reply to Sun and Firestone on the dark room problem. *Trends in Cognitive Science.*, 24(9):680–681.
- van Es, T. (2021). Living models or life modelled? On the use of models in the free energy principle. *Adaptive Behavior*. doi:10.1177/1059712320918678
- van Es, T., and Kirchhoff, M.D. (2021). Between pebbles and organisms: weaving autonomy into the Markov blanket. *Synthese*. https://doi.org/10.1007/s11229-021-03084-w
- Veissière, S. P., Constant, A., Ramstead, M. J., Friston, K. J., and Kirmayer, L. J. (2020). Thinking through other minds: A Variational approach to cognition and culture. *Behavioral and Brain Sciences*, 43.
- Vowels, M. J., Camgoz, N. C., & Bowden, R. (2021). D'ya like DAGs? A Survey on Structure Learning and Causal Discovery. *arXiv preprint arXiv:2103.02582*.
- Wiese W, and Friston KJ. (2021). Examining the Continuity between Life and Mind: Is There a Continuity between Autopoietic Intentionality and Representationality? *Philosophies*. 6(1):18. https://doi.org/10.3390/philosophies6010018.
- 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, D. (2018). Predictive Processing and the Representation Wars. *Minds and Machines*, 28: 141–172.
- Woodward, J. (2003). *Making Things Happen*. Oxford: OUP.
- Yon, D., Heyes, C., and Press, C. (2020). "Beliefs and Desires in the Predictive Brain." *Nature Communications* 11: 4404.
- Zhang, C., Bütepage, J., Kjellström, H., & Mandt, S. (2018). Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, *41*(8), 2008-2026.

[i] There are also other graphical formalisms commonly adopted in the literature outside of the ones proposed by Pearl, showing advantages in highlighting other features, for instance factor graphs (Bishop, 2006), but here the focus will be solely on Bayesian networks.

[ii] It should be noted that in its initial definition (Pearl, 1988) Markov blankets represented *all* possible sets of nodes shielding another node from the rest of the network, while the notion of a Markov *boundary* was used to characterise the smallest Markov blanket. Over time, however, the two definitions have often come to be used interchangeably to describe the minimal set of nodes, see for instance Bishop (2006), Murphy (2012). Here we will thus use 'Markov blanket' to refer to this latter notion.

[iii] Although Markov Blankets are typically presented visually as drawn on a Bayesian graph, the conditional independencies required for a Markov blanket can be obtained directly from the probability distribution.

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

[v] Note that the time index t is different from the time horizon τ used to describe instead the number of future steps to take into account when one optimizes a policy of τ-steps.

[vi] However see Millidge et al. (2020) for a treatment about the differences with more traditional frameworks for control as inference.

[vii] Unlike the 'naïve' or fully factorized mean-field (Zhang et al. 2018) where all latent variables are assumed to be independent, a structured mean-field imposes, as the name suggests, some non-trivial structure, i.e., independencies across partitions of hiddens variables rather than single ones.

[viii] Notice that the number of states identified as internal due to their coupling could have been smaller or larger, depending on the cut-off point for the metric of coupling used. It seems that in the original paper this was mostly an arbitrary choice following pragmatic, if somewhat unclear, considerations (Biehl, 2017 and Friston et al. 2021b).

[ix] Crucially, Friston blankets should be understood in the context of stochastic processes (i.e., time-indexed collections of random variables) rather than random variables for which Pearl blankets are usually defined. This implies the presence of an extra step whereby the nodes in the third panel ought be interpreted as part of a 'time slice' of a stochastic process after it has reached its non-equilibrium steady-state (NESS) (Friston et al., 2021a-b). Conditional independence is thus defined at the level of a single time slice of the NESS density, under the strong assumption that such density is a useful depiction of an agent-environment coupled system. Subsequently, and under a number of further non-trivial assumptions (Friston et al., 2021a, or see next note), this conditional independence is then applied to the dynamical couplings across different variables of the process.

[x] As highlighted by Biehl et al. (2021), the definition of Markov blankets using the adjacency matrix is ambiguous, and necessitates further, nontrivial constraints, i.e., independencies on different partitions of the variables now specified in Friston et al. (2021a), to be formally consistent with Pearl's notion of blankets. As Friston et al. (2021a) note, the use of the adjacency matrix (dynamical coupling or flow) has no direct relation to Pearl blankets, beyond a somewhat contrived version of conditional independence. In light of our discussion here, how-

ever, this aspect is not central, as we aim to showcase different issues in the use of Pearl blankets advocated under the free energy principle and active inference implementations, i.e., 'Friston blankets', even in their most recent formulations (Friston 2019, Friston et al 2021a-b).

[xi] The passage in Allen and Friston (2018) is part of a paragraph discussing relations between Friston blankets and the concept of *autopoiesis* for systems that 'self-create', maintaining their own existence over time via relational and operational constraints (Maturana and Varela, 1972, see also Beer 2004, 2014, 2020). This paragraph uses the paradigmatic example of an autopietic system: the living cell. The notion of physical boundary is thus interpreted following the given example, i.e., a cell membrane.

[xii] This apparent reversal can also be seen, for instance, in the following passages:

- Ramstead et al. (2019), "a Markov blanket induces a statistical partitioning between internal (systemic) and external (environmental) states" even though (and they do not specify the details) "Markov blankets are a result of the system's dynamics" (pp.43-4)
- Hesp et al. (2019), "The notion of a Markov blanket, and the independencies between states it induces, can be directly applied to […]" (p.198)
- Hipólito et al. (2019), "This figure highlights the conditional independencies induced by the presence of a Markov blanket." (p.14 of preprint, the same sentence also appears verbatim in Kirchhoff and Kiverstein 2021)
- Kirchhoff and Kiverstein (2019), "the Markov blanket for a cell […] renders the internal states of the cell statistically independent from its surroundings, and vice versa" (p.69), "The Markov blanket concept […] provides a statistical partitionioning of internal and external states" (p.71), and "The presence of a Markov blanket renders internal and external states conditionally independent of one another" (p.71)
- Ramstead et al. (2020a), "The presence of a Markov blanket induces a conditional independence between internal and external variables" (p.7)
- Ramstead et al. (2020c), "By inducing conditional independence (Pearl, 1988), Markov blankets enable us to define the boundaries between a system and its environment, and thereby delimit the system as such (Friston, 2013, 2020; Friston et al., 2015)." and "The existence of a Markov blanket induces certain conditional independences: the presence of the blanket partitions the system into […]" (p.11)
- Hipólito et al. (2021), "Ultimately, the dependencies induced by Markov blankets create a […]" (p.90)

[xiii] Note that specifications of these kinds do not require that anyone *literally* believe that the world itself is composed of Bayesian graphs, nodes and arrows (and we are certainly not accusing anyone of this), but rather just that they posit a direct, non-arbitrary mapping between a Markov blanket in a statistical model and a real, and in some ways meaningful, boundary in the world. This non-arbitrary mapping is sometimes attributed the status of a *structure-preserving* mapping, or isomorphism, for instance by Palacios et al. (2020) where "[t]he isomorphism between a statistical and spatial boundary rests on spatially dependent interactions among internal and external states." Although some formulations do suggest a literalist understanding of Markov Blankets, it is the latter kind of project that we think is particularly widespread in the contemporary literature and are criticising here.

[xiv] As highlighted in Friston et al. (2010), the notion of "command" in active inference is best understood in terms of proprioceptive predictions, with action seen in terms of minimising proprioceptive prediction errors. Here we stick to widely accepted nomenclature for the sake of simplicity.

[xv] The sensorimotor perspective is inherent in *active* inference formulations with, for instance, "[t]he treatment of neurons as if they were active agents" (Hipólito et al., 2021).

[xvi] Note that the problem of distinguishing proximal from distal interactions is different from simillar worries in philosophy of causation and in debates over internalisim and externalism. Here the problem is specific to the postulate of using Markov blankets as tools for picking out active agents from the environments in which they are embedded.

[xvii] In most cases, one might consider a relevant Friston blanket to be a structure that can be used to characterize a cell membrane as opposed to, say, a structure that maps to an arbitrary fraction of a cell split into five parts, where relations of conditional independence can nonetheless be identified using different thresholds (cf. Friston et al. 2021b). This choice of relevance is nonetheless a choice that has to be made at some point in the modelling process, and cannot simply be read off the model itself. Friston et al. (2021b) elegantly describe the problem: "The nonuniqueness of the particular partition is a key practical issue. There is no pretense that there is any unique particular partition. There are a vast number of particular partitions for any given coupled dynamical system. In other words, by simply starting with different internal states —or indeed the number of internal states per particle—we would get a different particular partition." (p.245-6).