Behavioral and Brain Sciences

cambridge.org/bbs

The Emperor's New Markov Blankets

Jelle Bruineberg[a], Krzysztof Dołęga[b], Joe Dewhurst[c] and Manuel Baltieri[d]

[a]Department of Philosophy, Macquarie University, Sydney, NSW 2109, Australia; [b]Institut für Philosophie II, Fakultät für Philosophie und Erziehungswissenschaft, Ruhr-Universität Bochum, 44801 Bochum, Germany; [c]Fakultät für Philosophie, Wissenschaftstheorieund Religionswissenschaft, Munich Center for Mathematical Philosophy, Ludwig-Maximilians-Universität München, 80539 Munich, Germany and [d]Laboratory for Neural Computation and Adaptation, RIKEN Centre for Brain Science, 351-0106 Wako City, Japan

jelle.bruineberg@mq.edu.au
krzysztof.dolega@rub.de
joseph.e.dewhurst@gmail.com
manuel.baltieri@riken.jp

Target Article

Cite this article: Bruineberg J, Dołęga K, Dewhurst J, Baltieri M. (2022) The Emperor's New Markov Blankets. *Behavioral and Brain Sciences* **45**, e183: 1–76. doi:10.1017/S0140525X21002351

Target Article Accepted: 16 October 2021
Target Article Manuscript Online: 22 October 2021
Commentaries Accepted: 24 February 2022

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

What is Open Peer Commentary? What follows on these pages is known as a Treatment, in which a significant and controversial Target Article is published along with Commentaries (p. 21) and an Author's Response (p. 66). See bbsonline.org for more information.

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 programmes: "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.

1. Introduction

The last 20 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 Clark, 2013; Griffiths & Tenenbaum, 2006; Knill & Pouget, 2004; Körding & Wolpert, 2004; Oaksford & Chater, 2001; Tenenbaum, Kemp, Griffiths, & Goodman, 2011). 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, 2010, 2019; Friston, Daunizeau, Kilner, & Kiebel, 2010; Friston, FitzGerald, Rigoli, Schwartenbeck, & Pezzulo, 2017a; Friston, Kilner, & Harrison, 2006), 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; Lee & Mumford, 2003; Rao & Ballard, 1999) 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., 2017a). 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, Levin, Sengupta, & Pezzulo, 2015a).

© The Author(s), 2021. Published by Cambridge University Press

CAMBRIDGE UNIVERSITY PRESS

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

while also offering (apparently) new insights on:

- the (trans)formation and survival of social and societal organizations (Boik, 2021; Fox, 2021; Khezri, 2021),
- climate systems and planetary-scale self-organization and autopoiesis (Rubin, Parr, Da Costa, & Friston, 2020),
- the notions of "self" and "individual," with studies on the sense of agency and on body ownership (Hafner, Loviken, Villalpando, & Schillaci, 2020), (in utero) co-embodiment (Ciaunica, Constant, Preissl, & Fotopoulou, 2021), pain experience (Kiverstein, Kirchhoff, & 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., 2019; Ramstead, Badcock, & Friston, 2018; Veissière, Constant, Ramstead, Friston, & Kirmayer, 2020) and possibly evolve (Poirier, Faucher, & Bourdon, 2021).

The formalisms deployed by the FEP (as outlined in sect. 3 and 4 of this paper) are sometimes explicitly presented as replacing older (and supposedly outdated) philosophical arguments (Ramstead, Kirchhoff, Constant, & Friston, 2019; Ramstead, Friston, & Hipólito, 2020a), 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 e.g., the exchange between Seth, Millidge, Buckley, & Tschantz [2020]; Sun & Firestone [2020a]; Van de Cruys, Friston, & Clark [2020]; as well as Sun & 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 modelling, 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 (Bishop, 2006; Murphy, 2012; Pearl, 1988). 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

Jelle Bruineberg is currently a Research Fellow at the Department of Philosophy of Macquarie University. In his doctoral research at the University of Amsterdam, he investigated the philosophical implications of active inference and the free energy principle and developed the argument that the free energy principle can be understood as an extension of embodied theories of the mind. In 2020, he received a Macquarie Research Fellowship to work on topics related to embodied cognition, digital technology, and the attention economy.

Krzysztof Dołęga is a Postdoctoral Fellow at the Ruhr-Universität Bochum, where he is currently realizing the Volkswagen Foundation sponsored research project "Why do people believe weird things? Bayesian Brain, Conspiracy Theories, and Intellectual Vices." Apart from probabilistic models of cognition, his work focuses on topics related to intentionality, mental representation, consciousness, and artificial intelligence. He is the co-editor, together with Joulia Smortchkova and Tobias Schlicht, of the 2020 Oxford University Press collection titled "What Are Mental Representations?"

Joe Dewhurst is currently a Postdoctoral Fellow at the Munich Center for Mathematical Philosophy, where he works on topics to do with computation, mechanistic explanation, and more recently formal approaches to causation and emergence in complex systems. His earlier doctoral research at the University of Edinburgh looked at the relationship between common-sense intuitions and scientific theories in contemporary cognitive science, and he worked at Edinburgh as a Teaching Assistant before moving to Munich in 2018.

Manuel Baltieri is a Researcher at Araya Inc. and a Visiting Researcher at the University of Sussex. Previously, he was a JSPS/Royal Society Postdoctoral Fellow in the Laboratory for Neural Computation and Adaptation at the RIKEN Centre for Brain Science while working on the target article. His research spans different areas of artificial intelligence, embodied cognition, control theory, Bayesian inference methods, and more generally, theories of agency, individuality, and agent–environment interactions in biology and artificial life.

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

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 FEP. 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 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 normalizing 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), normalized 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., Friston, Harrison, & Penny, 2003; Gregory, 1980; Penny, Friston, Ashburner, Kiebel, & Nichols, 2011).

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 (Beal, 2003; Bishop, 2006; MacKay, 2003). 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 (Beal, 2003; Bishop, 2006; Blei, Kucukelbir, & McAuliffe, 2017; Hinton & Zemel, 1994; Jordan, Ghahramani, Jaakkola, & Saul, 1999; MacKay, 2003; Zhang, Bütepage, Kjellström, & Mandt, 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 variables, traditionally defined with analytical methods, more intuitive and easy to track.[1] 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

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

(Friston, Parr, & de Vries, 2017b). Relationships between the variables are often described using genealogical terms, with *pa*(*a*) being the *parents* (or "ancestors") of their *children* (or "descendants") node *a* and *copa(a)* being the co-parents: nodes with which *a* has a child in common. In Figure 1 below, *m* is the target variable, *c* and *b* are the parents of *m*, *a* is the child of *m*, and *e* is *m*'s co-parent since they have *a* in common as a child. 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.

Let us introduce a simple textbook example that will help familiarize 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:

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

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[2] 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,[3] that is, it *shields* that node from all other nodes. Formally, a Markov blanket for a set of variables x_i is thus equivalent to:

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

where $\text{pa}(x_i)$ corresponds to the parents of x_i, $\text{ch}(x_i)$ to the children, and $\text{copa}(x_i)$ 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 *m* or *mb*(*m*). It is composed of *m*'s parents (*c* and *b*), its child (*a*), and its children's other parents (*e*). The *mb*(*j*) shown in Figure 1b, on the other hand, is composed of just two nodes (*a* and *t*), hence:

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

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 $q^*(x)$ that is closest to the true posterior. The candidate distribution is often called the recognition or variational density, because the methods used employ variational

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

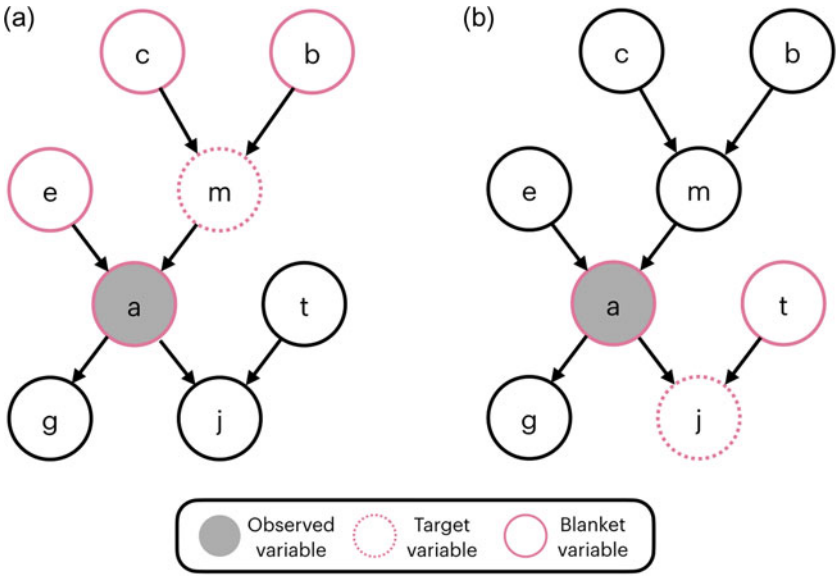


Figure 1. The "alarm" network with examples of Markov Blankets for two different variables. The target variables are indicated with a dashed pink circle, while the variables that are part of the Markov blanket are indicated with a solid pink circle.

calculus, that is, 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}):

$$q^*(x) = \arg\min_{q(x)\in Q} D_{KL}(q(x) \parallel p(x|y)). \tag{5}$$

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 minimization of a quantity that is called variational free energy (see Bishop, 2006; Murphy, 2012):

$$\begin{aligned} q^*(x) &= \arg\min_{q(x)\in Q} \int q(x) \ln \frac{q(x)}{p(y, x)} dx \\ &= \arg\min_{q(x)\in Q} F(x). \end{aligned} \tag{6}$$

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 & 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 to 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, that is, 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, Camgoz, & Bowden [2021] for a recent review). In this family of methods, the class of score-based approaches (Vowels et al., 2021) is of particular

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

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 (Penny et al., 2011; Stephan, Penny, Daunizeau, Moran, & Friston, 2009). 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 (Friston et al., 2017b; Penny et al., 2011; Stephan et al., 2009). This intuition can be formalized via a process of model comparison using different criteria, for example, the Akaike information criterion, the Bayesian information criterion, or variational free energy (via the maximization of model evidence, equivalent to the minimization of surprisal), but there is a general agreement that Bayesian methods offer a quantification of Occam's razor (Jefferys & 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 Bishop, 2006; Daunizeau, 2017; Friston, 2010; MacKay, 2003).

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 (Baltieri & Buckley, 2019; Stephan et al., 2010). 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 (Penny et al., 2011; Stephan et al., 2010). The same phenomenon can be analysed using different sources of data. For example, in a study of decision making one can include only behavioural data, or add neural measurements as well. The choice of relevant dimensions in model 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."[4]

3. Pearl blankets in the active inference framework

The specific application of the FEP 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 Fig. 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.

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, 2013, 2019; Friston et al., 2010, 2015b, 2017a). The core assumption underlying active inference is that living organisms can be thought of as systems whose fundamental imperative is to minimize free energy (this constitutes the so-called free energy principle). Active inference attempts to explain action, perception, and other aspects of cognition under

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

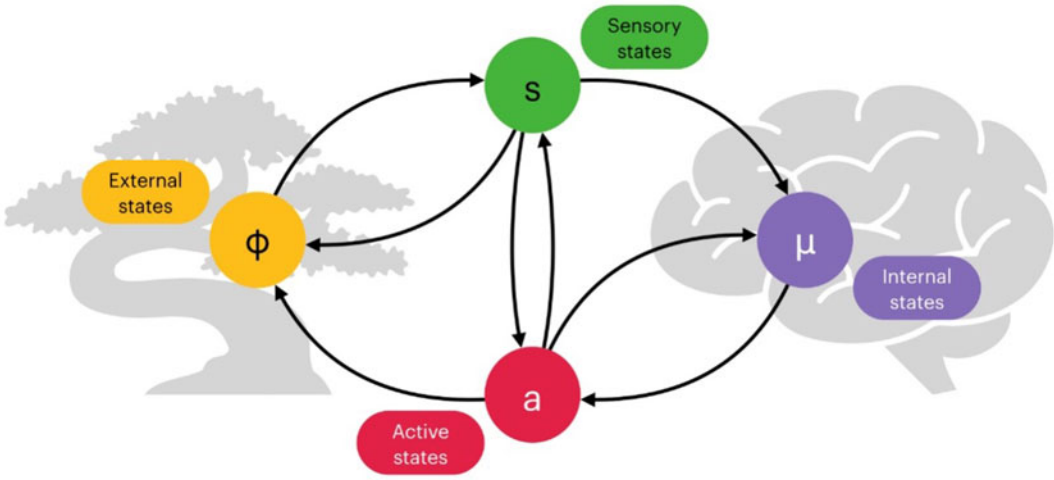


Figure 2. The Markov blanket as a sensorimotor loop (adapted from Friston, 2012). A diagram representing possible dependences between different components of interest: sensory states (green), internal states (violet), active states (red), and external states (yellow). Notice that although this figure 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).

the umbrella of variational (and expected) free energy minimization (Feldman & Friston, 2010; Friston et al., 2010, 2017a). From this perspective, perception can be understood as a process of optimizing 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, Hinton, Neal, and Zemel, 1995; Friston, 2005; Knill & Richards, 1996; Lee & Mumford, 2003; Rao & Ballard, 1999). At the same time, action is conceptualized 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, & Opper, 2012), with agents changing the world to better meet their expectations.

Active inference integrates a more general framework where minimizing expected free energy accounts for more complex processes of action and policy selection (Friston et al., 2015b, 2017a; Tschantz, Seth, & 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., Biehl, Guckelsberger, Salge, Smith, & Polani, 2018; Bogacz, 2017; Buckley, Kim, McGregor, & Seth, 2017; Da Costa et al., 2020; Friston et al., 2017b; Sajid, Ball, Parr, & Friston, 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 $0 \leq \tau \leq T$, that is, $\pi = [u_1, u_2, \ldots, u_\tau]$. This will allow us to formulate perception and action as variational problems in active inference. Perception is the minimization (at each time step t)[5] of the following equation:

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

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

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

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

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

This describes action selection as a minimization 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; Kappen et al., 2012).[6] 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 (Beal, 2003; Bishop, 2006; MacKay, 2003).

Some formulations of active inference, especially the earlier ones (Friston, 2008; Friston, Mattout, Trujillo-Barreto, Ashburner, & Penny, 2007; Friston, Trujillo-Barreto, & Daunizeau, 2008), have explicitly relied on a set of assumptions similar to the ones mentioned in the previous section: a mean-field approximation and

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

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 minimization 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 et al. (2008), and Friston (2008), however, make use of a "structured" mean-field assumption,[7] 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, 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 arises 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 behavioural and neural data) to infer the most likely model that the agent's brain might be implementing. For example, Parr, Mirza, Cagnan, and Friston (2019) investigate the generative models that underlie active vision. They use both MEG and eye-tracking to disambiguate a number of potential generative 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 behaviour. The most likely model (i.e., the one that best explains the data in the most parsimonious way) is selected by scoring each model based on its accuracy in predicting neural dynamics and oculomotor behaviour 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 modelling 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 modelling 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 FEP, Markov blankets have been assigned a role that they 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 modelled 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 (Fig. 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.

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

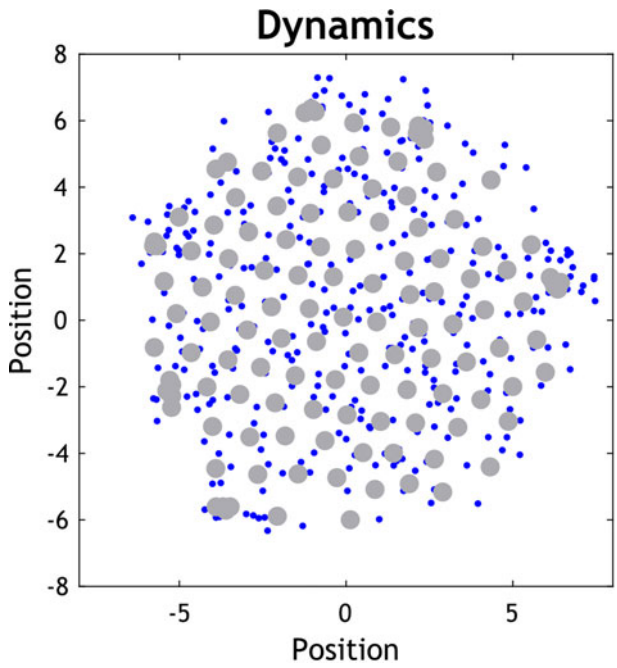


Figure 3. The "primordial soup" (adapted from Friston [2013] using the code provided). The larger (grey) dots represent the location of each particle, which are assumed to be observed by the modellers. There are three smaller (blue) dots associated with each particle, representing the electrochemical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

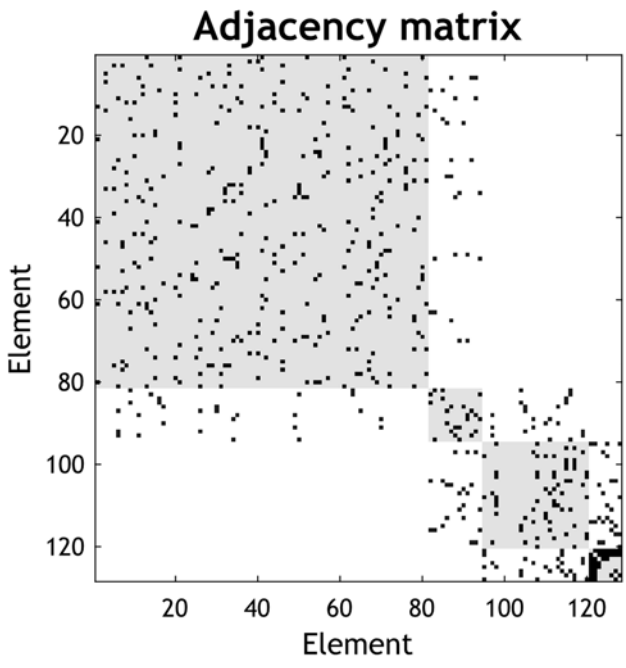


Figure 4. The adjacency matrix of the simulated soup at steady-state (from Friston, 2013). Element *i, j* has value 1 (a dark square) if and only if subsystem *i* electrochemically affects subsystem *j*. The four grey squares from top left to bottom right represent the hidden states, the sensory states,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

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 (Fig. 4). Because of the way the simulation is set up, the network will not be symmetrical (since a third of the randomly selected particles will not electrochemically affect the remaining ones).

Spectral graph theory is then used to identify the eight most densely coupled nodes, which are stipulated to be the "internal" states.[8] 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 Fig. 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

Applied to the primordial soup simulation, each particle can be coloured to indicate which of these sets it has been assigned to (see Fig. 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 labelled 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.

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-trivial assumptions, some of which are now included in Friston et al., (2021a, 2021b), 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

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

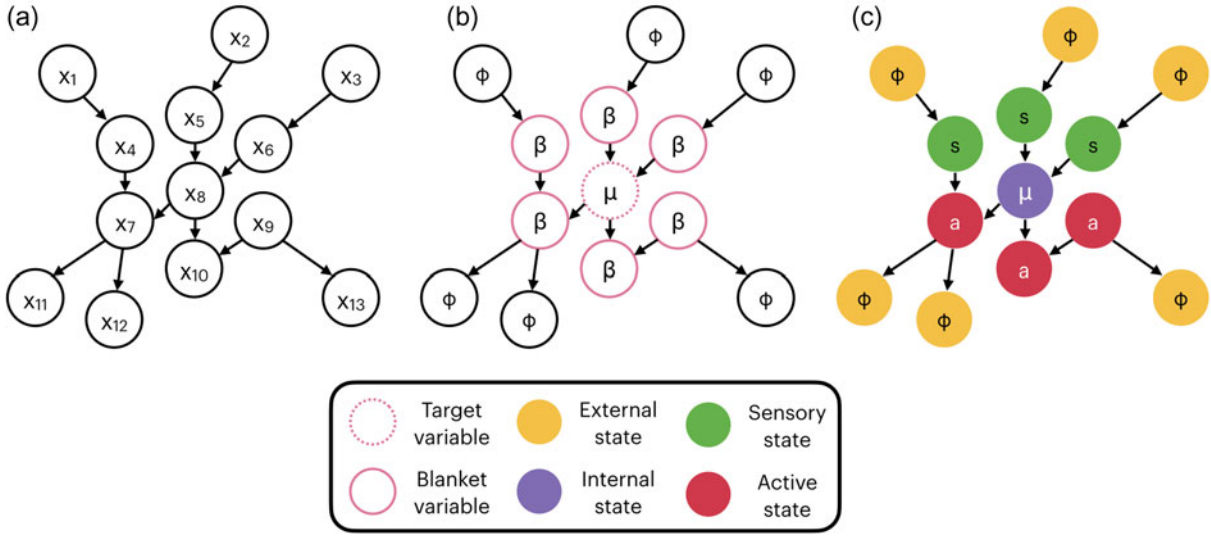


Figure 5. The Friston blanket. The three diagrams representing the stages of identifying a Friston blanket described in section 4.1. A system of interest is represented in the form a directed graph (a). Next the variable of interest is identified and a Markov blanket of shielding variables β is delineated 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).[9]

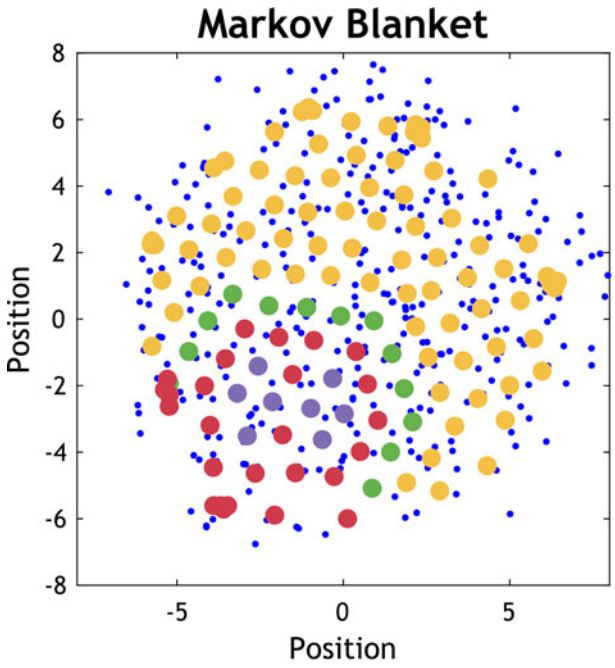


Figure 6. The Markov blanket of the simulated soup at steady-state (adapted from Friston [2013] using the code provided). Similarly to Figure 3, particles are indicated by larger dots. Particles that belong to the set of sensory states are in green, active states are in red. Internal states are violet,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

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 Fig. 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.[10] A first interpretative step is taken when labelling 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 that, 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 modelled as having a blanket (see

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

for instance Friston, 2013; Palacios, Razi, Parr, Kirchhoff, & Friston, 2020). Without considering the further assumptions explained in Biehl, Pollock, and Kanai (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.[11] Markov blankets here 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, that is, 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, for example, 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).[12] 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 characterization 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 & 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, that is, the sensory surface. Hohwy even goes as far as to define the philosophical position of epistemic internalism in terms of a Markov blanket:

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

> 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 et al. (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.[13] 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 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 that 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 Figure 7.

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*, that is, 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*).[14] 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 that 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) and 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[15] (see Barandiaran, Di Paolo, & Rohde, 2009; Tishby & Polani, 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 Figure 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,

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

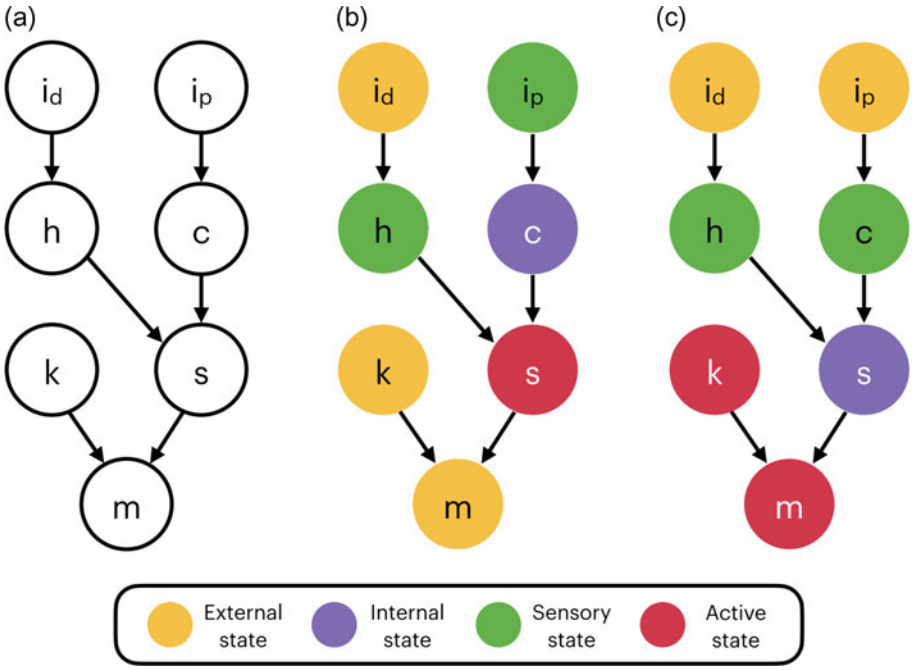


Figure 7. Conditions leading up to the knee-jerk reflex. On the left, a Bayesian network where i_d and i_p denote the motor intentions of the doctor and the patient respectively. Node *s* denotes the spinal neurons that are directly responsible for causing the kicking movement *m*. Node *h* indicates 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, for example, by someone else kicking it to move it mechanically. The middle (b) and the right figures (c) with the coloured-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.

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, that is, 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.[16] 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 behaviour, 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 make use of the Bayesian Occam's razor (mentioned in section 2.4.), since the goal is not to predict the behaviour 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 modelling choices, that is, *relevant* Friston blankets cannot simply be "detected" in some objective way and then used to determine the boundary of a system.[17] This can be easily seen by the fact that Markov blankets are defined only in relation to a set of conditional (in)dependencies,

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

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., cf. Andrews, 2020; Beni, 2021; Friston et al., 2020; Hohwy, 2016; Sánchez-Cañizares, 2021; Wiese & Friston, 2021 for some recent critical discussion).

While we do not want to try and solve all of these issues here, it is important to recognize 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 (2020a); 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. 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 emphasize 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 programmes. 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 programme 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 programme, 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 programmes 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.

6.1.1 *Inference with a model*

As mentioned above, an important motivation for the FEP 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 that 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 that approximates the posterior density.

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

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 (Adams, Stephan, Brown, Frith, & Friston, 2013; Parr et al., 2019; Pezzulo, Rigoli, & Friston, 2018), and reflects a more general explanatory strategy in cognitive neuroscience (Lee & Mumford, 2003; Rao & Ballard, 1999). Although perhaps not directly empirically refutable (cf. Andrews, 2020), this approach guides an active research programme, 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 *modelled* 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 & Seriès, 2012; Ramstead, Kirchhoff, & Friston, 2020b; Ramstead et al., 2020a; van Es, 2021), 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 (Clark, 2015a, 2015b; Dołęga, 2017; Gładziejewski, 2016; Kiefer & Hohwy, 2018; Kirchhoff & 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 (Baltieri & Buckley, 2019; Bruineberg, Kiverstein, & Rietveld, 2018; Friston, 2013; Friston, 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.

6.1.2 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-organizing 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 blankets 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 modelled 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

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

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 programme 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 that is then used to select the most parsimonious explanatory model of these observations (see sect. 2.4). In turn, Friston blankets can be understood only relative to such a model (see sect. 5.2). The puzzle then is that if one wants to use the Markov blanket formalism to demarcate the boundaries of, for example, 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 alone, but rather the Markov blanket plus the FEP, that provides the relevant demarcations of agent–environment boundaries. Only those Markov blankets that demarcate free energy minimizing systems (or the systems that 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 FEP 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 FEP 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 understood 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, & Hartmann, 2018; Ramstead et al., 2020a, 2020b; van Es, 2021). 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 BOO 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 FEP 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, 2019; Friston et al., 2017a; Hesp et al., 2019; Kuchling, Friston, Georgiev, & Levin, 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; Friston et al., 2020; Hohwy, 2013). 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 have 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, & Press, 2020), although its defenders have argued that it can account for desire in a novel way (Clark, 2020; Wilkinson, Deane, Nave, & Clark, 2019). 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 & Palacios, 2021; Colombo & Wright, 2021; Williams, 2021), and should therefore be considered to offer, at best, a redescription of existing data (see e.g., Cao, 2020; Colombo et al., 2018; Liwtin & Miłkowski, 2020).

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

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 & Palacios, 2021; Colombo & Wright, 2021; Williams, 2021), or that it presents a false equivocation between probability and adaptive value (Colombo, 2020). Other works, such as Di Paolo, Thompson, and Beer (2021) and Raja, Valluri, Baggs, Chemero, and Anderson (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 (2021) have recently argued against a realist interpretation of the mathematical models described by FEP, which are claimed to be better interpreted instrumentally. Along the same lines, Baltieri et al. (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 minimization. Finally, various technical aspects of the FEP are now under scrutiny in works such as Rosas, Mediano, Biehl, Chandaria, and Polani (2020); Biehl et al. (2021); and Aguilera, Millidge, Tschantz, and Buckley (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 9). 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 description 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, Friston, Heins, & Pavliotis [2021] and Parr, Da Costa, Heins, Ramstead, & Friston [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 FEP, 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 FEP to clarify the boundaries of the mind (Clark, 2017; Hohwy, 2017; Kirchhoff & Kiverstein, 2021), of living systems (Friston, 2013; Kirchhoff, 2018; Kirchhoff et al., 2018), and even of social systems (Fox, 2021; Ramstead et al., 2018; Rubin et al., 2020; Veissière et al., 2020). 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; Friston, 2019; Hipólito et al., 2021; Kirchhoff et al., 2018). We then raised some possible issues with this approach, namely the question of whether Bayesian networks are merely an instrumental modelling tool for the FEP 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 FEP 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 & 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 (Biehl et al., 2021; Friston, 2019; Parr, Da Costa, & Friston, 2020) and philosophical (Friston et al., 2020; Hipólito et al., 2021; Ramstead et al., 2018) 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

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

authors also thank the editor and eight reviewers for their time and effort. The manuscript has benefited enormously from their critical reports.

Financial Support. 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).

Conflict of Interest. None.

Notes

1. 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.
2. 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 characterize 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.
3. 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.
4. 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.
5. 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.
6. However, see Millidge, Tschantz, Seth, and Buckley (2020) for a treatment about the differences with more traditional frameworks for control as inference.
7. 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, that is, independencies across partitions of hidden variables rather than single ones.
8. 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; Friston et al., 2021b).
9. 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 to 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, 2021b). 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.
10. As highlighted by Biehl et al. (2021), the definition of Markov blankets using the adjacency matrix is ambiguous, and necessitates further, non-trivial constraints, that is, 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, however, 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, that is, "Friston blankets," even in their most recent formulations (Friston, 2019; Friston et al., 2021a, 2021b).
11. 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 & 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, that is, a cell membrane.
12. 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. 43–44)
- Hesp et al. (2019), "The notion of a Markov blanket, and the independencies between states it induces, can be directly applied to […]" (p. 198)
- 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. (2021), "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, 2019; Friston et al., 2015a)." 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)

13. 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 to 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.
14. 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 minimizing proprioceptive prediction errors. Here we stick to widely accepted nomenclature for the sake of simplicity.
15. 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).
16. Note that the problem of distinguishing proximal from distal interactions is different from similar worries in philosophy of causation and in debates over internalism 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.
17. 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." (pp. 245–246)

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

References

Adams, R. A., Stephan, K., Brown, H., Frith, C., & 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., & Friston, K. J. (2018). From cognitivism to autopoiesis: Towards a computational framework for the embodied mind. *Synthese*, *195*(6), 2459–2482.

Anderson, M. L. (2017). Of Bayes and bullets: An embodied, situated, targeting-based account of predictive processing. In W. Wiese & T. K. Metzinger (Eds.), *Philosophy and predictive processing* (Vol. *4*, pp. 1–14). MINDGroup.

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

Attias, H. (2003). Planning by probabilistic inference. In Bishop, C. M., & Frey, B. J. (Eds.), *Proc. of the 9th Int. Workshop on artificial intelligence and statistics, 2003* (pp. 9–16). PMLR.

Baltieri, M., & Buckley, C. L. (2019). Generative models as parsimonious descriptions of sensorimotor loops. *Behavioral and Brain Sciences*, *42*, e218.

Baltieri, M., Buckley, C. L., & Bruineberg, J. (2020). Predictions in the eye of the beholder: An active inference account of Watt governors. *Artificial life conference proceedings* (pp. 121–129). 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*, *199*, 6407–6427. 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., & 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., & 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.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, *112*(518), 859–877.

Bogacz, R. (2017). A tutorial on the free energy framework for modelling perception and learning. *Journal of Mathematical Psychology*, *76*, 198–211.

Boik, J. C. (2021). Science-driven societal transformation, part III: Design. *Sustainability*, *13*(2), 726.

Bruineberg, J., Kiverstein, J., & Rietveld, E. (2018). The anticipating brain is not a scientist: The free energy principle from an ecological-enactive perspective. *Synthese*, *195*(6), 2417–2444.

Buckley, C. L., Kim, C. S., McGregor, S., & 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), 517–546.

Chakravartty, A. (2017). Scientific realism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (summer 2017 edition). https://plato.stanford.edu/archives/sum2017/entries/scientific-realism/

Ciaunica, A., Constant, A., Preissl, H., & 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*(3), 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 T. K. Metzinger & W. Wiese (Eds.), *Philosophy and predictive processing*: 3. Open MIND (pp. 1–19). MIND Group.

Clark, A. (2020). Beyond desire? Agency, choice, and the predictive mind. *Australasian Journal of Philosophy*, *98*(1), 1–15.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, *58*(1), 7–19.

Colombo, M. (2020). Maladaptive social norms, cultural progress, and the free-energy principle. *Behavioral and Brain Sciences*, *43*, e100.

Colombo, M., Elkin, L., & 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., & Palacios, P. (2021). Non-equilibrium thermodynamics and the free energy principle in biology. *Biology & Philosophy*, *36*(5), 1–26.

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

Colombo, M., & Wright, C. (2021). First principles in the life sciences: The free-energy principle, organism, and mechanism. *Synthese*, *198*(14), 3463–3488.

Da Costa, L., Friston, K., Heins, C., & Pavliotis, G. A. (2021). Bayesian mechanics for stationary processes. *Proceedings of the Royal Society A*, *477*(2256), 20210518.

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., & 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* (pp. 1–13). 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 T. K. Metzinger & W. Wiese (Eds.), *Philosophy and predictive processing* (pp. 1–19). MIND Group.

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., & 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., Sengupta, B., & Auletta, G. (2014). Cognitive dynamics: From attractors to active inference. *Proceedings of the IEEE*, *102*(4), 427–445.

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., & Ao, P. (2012). Free energy, value, and attractors. *Computational and mathematical methods in medicine*, Volume 2012, Article ID 937860.

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., Daunizeau, J., Kilner, J., & Kiebel, S. J. (2010). Action and behavior: A free energy formulation. *Biological Cybernetics*, *102*(3), 227–260.

Friston, K. J., Fagerholm, E. D., Zarghami, T. S., Parr, T., Hipólito, I., Magrou, L., & Razi, A. (2021b). Parcels and particles: Markov blankets in the brain. *Network Neuroscience*, *5*(1), 211–251.

Friston, K. J., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017a). Active inference: A process theory. *Neural Computation*, *29*(1), 1–49.

Friston, K. J., Harrison, L., & Penny, W. (2003). Dynamic causal modelling. *Neuroimage*, *19*(4), 1273–1302.

Friston, K. J., Kilner, J., & Harrison, L. (2006). A free energy principle for the brain. *Journal of Physiology-Paris*, *100*(1), 70–87.

Friston, K. J., Levin, M., Sengupta, B., & 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., Mattout, J., Trujillo-Barreto, N., Ashburner, J., & Penny, W. (2007). Variational free energy and the Laplace approximation. *Neuroimage*, *34*(1), 220–234.

Friston, K. J., Parr, T., & de Vries, B. (2017b). The graphical brain: Belief propagation and active inference. *Network Neuroscience*, *1*(4), 381–414.

Friston, K. J., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., & Pezzulo, G. (2015b). Active inference and epistemic value. *Cognitive Neuroscience*, *6*(4), 187–214.

Friston, K. J., Trujillo-Barreto, N., & Daunizeau, J. (2008). DEM: A variational treatment of dynamic systems. *NeuroImage*, *41*(3), 849–885.

Friston, K. J., Wiese, W., & Hobson, J. A. (2020). Sentience and the origins of consciousness: From Cartesian duality to Markovian monism. *Entropy*, *22*(5), 516.

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.

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

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., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, *17*(9), 767–773.
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., & Schillaci, G. (2020). Prerequisites for an artificial self. *Frontiers in Neurorobotics*, *14*, 1–10.
Hesp, C., Ramstead, M., Constant, A., Badcock, P., Kirchhoff, M., & Friston, K. (2019). A multi-scale view of the emergent complexity of life: A free energy proposal. In *Evolution, development and complexity* (pp. 195–227). Springer.
Hinton, G. E., & Zemel, R. S. (1994). Autoencoders, minimum description length, and Helmholtz free energy. In *Advances in neural information processing systems* (pp. 3–10). 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 University Press.
Hohwy, J. (2016). The self-evidencing brain. *Noûs*, *50*(2), 259–285.
Hohwy, J. (2017). How to entrain your evil demon. In T. K. Metzinger & W. Wiese (Eds.), *Philosophy and predictive processing*: 2. Open MIND (pp. 1–15). MIND Group.
Jefferys, W. H., & Berger, J. O. (1991). Sharpening Occams razor on a Bayesian strop. *Bulletin of the Astronomical Society*, *23*(3), 1259.
Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., & Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning*, *37*(2), 183–233.
Kappen, H. J., Gómez, V., & 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., & Hohwy, J. (2018). Content and misrepresentation in hierarchical generative models. *Synthese*, *195*, 2387–2415.
Kiefer, A. B. (2020). Psychophysical identity and free energy. *Journal of The Royal Society Interface*, *17*(169), 20200370. http://dx.doi.org/10.1098/rsif.2020.0370
Kirchhoff, M. D. (2018). Autopoiesis, free energy, and the life–mind continuity thesis. *Synthese*, *195*(6), 2519–2540.
Kirchhoff, M. D., & Kiverstein, J. (2019). *Extended consciousness and predictive processing: A third-wave view*. Routledge.
Kirchhoff, M. D., & Kiverstein, J. (2021). How to determine the boundaries of the mind: A Markov blanket proposal. *Synthese*, *198*(5), 4791–4810. https://doi.org/10.1007/s11229-019-02370-y
Kirchhoff, M. D., Parr, T., Palacios, E., Friston, K. J., & 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., & Robertson, I. (2018). Enactivism and predictive processing: A non-representational view. *Philosophical Explorations*, *21*(2), 264–281.
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., & 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), 2541–2557.
Knill, D. C., & 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., & Richards, W. (1996). *Perception as Bayesian inference*. Cambridge University Press.
Körding, K., & Wolpert, D. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*, 244–247.
Kuchling, F., Friston, K. J., Georgiev, G., & Levin, M. (2020). Morphogenesis as Bayesian inference: A variational approach to pattern formation and control in complex biological systems. *Physics of Life Reviews*, *33*, 88–108.
Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *JOSA A*, *20*(7), 1434–1448.
Litwin, P., & Miłkowski, M. (2020). Unification by fiat: Arrested development of predictive processing. *Cognitive Science*, *44*, e12867.
MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. CUP.
Maturana, H. R., & Varela, F. J. (1972). *Autopoiesis and cognition: The realization of the living* (Vol. 42). Springer Science & Business Media.
Menary, R., & Gillett, A. J. (2020). Are Markov blankets real and does it matter? In D. Mendonca, M. Curado, & S. S. Gouveia (Eds.), *The philosophy and science of predictive processing* (pp. 39–58). Bloomsbury Academic.
Millidge, B., Tschantz, A., Seth, A. K., & Buckley, C. L. (2020). On the relationship between active inference and control as inference. *International workshop on active inference* (pp. 3–11). Springer.
Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT press.
Oaksford, M., & Chater, N. (2001). The probabilistic approach to human reasoning. *Trends in Cognitive Sciences*, *5*(8), 349–357.
Opper, M., & 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., 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.
Parr, T., Mirza, M. B., Cagnan, H., & Friston, K. J. (2019). Dynamic causal modelling of active vision. *Journal of Neuroscience*, *39*(32), 6265–6275.
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 images*. Elsevier.
Pezzulo, G., Rigoli, F., & Friston, K. J. (2018). Hierarchical active inference: A theory of motivated control. *Trends in Cognitive Sciences*, *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. *Physics of Life Reviews*, *39*(2), 49–72. doi:10.1016/j.plrev.2021.09.001
Ramstead, M. J., Friston, K. J., & Hipólito, I. (2020a). Is the free energy principle a formal theory of semantics? From variational density dynamics to neural and phenotypic representations. *Entropy*, *22*(8), 889.
Ramstead, M. J., Hesp, C., Tschantz, A., Smith, R., Constant, A., & Friston, K. (2021). Neural and phenotypic representation under the free-energy principle. *Neuroscience & Biobehavioral Reviews*, *120*, 109–122. https://www.sciencedirect.com/science/article/pii/S0149763420306643
Ramstead, M. J., Kirchhoff, M. D., Constant, A., & Friston, K. J. (2019). Multiscale integration: Beyond internalism and externalism. *Synthese*, *198*, 41–70.
Ramstead, M. J., Kirchhoff, M. D., & Friston, K. J. (2020b). A tale of two densities: Active inference is enactive inference. *Adaptive Behavior*, *28*(4), 225–239.
Ramstead, M. J. D., Badcock, P. B., & Friston, K. J. (2018). Answering Schrödinger's question: A free energy formulation. *Physics of Life Reviews*, *24*, 1–16.
Rao, R. P., & 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. M., Biehl, M., Chandaria, S., & Polani, D. (2020). Causal blankets: Theory and algorithmic framework. In T. Verbelen, P. Lanillos, C. L. Buckley, & C. De Boom (Eds.), *Active Inference. IWAI 2020. Communications in Computer and Information Science* (Vol. *1326*, pp. 187–198). Springer.
Rubin, S., Parr, T., Da Costa, L., & 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., & Friston, K. J. (2021). Active inference: Demystified and compared. *Neural Computation*, *33*(3), 674–712.
Sánchez-Cañizares, J. (2021). The free energy principle: Good science and questionable philosophy in a grand unifying theory. *Entropy*, *23*(2), 238. https://doi.org/10.3390/e23020238
Seth, A., Millidge, B., Buckley, C. L., & 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*, *199*, 2157–2179. https://doi.org/10.1007/s11229-020-02876-w
Stephan, K. E., Penny, D., Daunizeau, W. J., Moran, R. J., & 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., & Firestone, C. (2020a). The dark room problem. *Trends in Cognitive Sciences*, *24*, 346–348.
Sun, Z., & Firestone, C. (2020b). Optimism and pessimism in the predictive brain. *Trends in Cognitive Sciences*, *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 (New York, N.Y.)*, *331*(6022), 1279–1285.
Tishby, N., & Polani, D. (2011). Information theory of decisions and actions. *Perception-action cycle* (pp. 601–636). Springer.
Tschantz, A., Seth, A. K., & Buckley, C. L. (2020). Learning action-oriented models through active inference. *PLoS Computational Biology*, *16*(4), e1007805.
Van de Cruys, S., Friston, K. J., & Clark, A. (2020). Controlled optimism: Reply to Sun and Firestone on the dark room problem. *Trends in Cognitive Science*, *24*(9), 680–681.

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

van Es, T. (2021). Living models or life modelled? On the use of models in the free energy principle. *Adaptive Behavior*, *29*(3), 315–329. https://doi.org/10.1177/1059712320918678

van Es, T., & Kirchhoff, M. D. (2021). Between pebbles and organisms: Weaving autonomy into the Markov blanket. *Synthese*, *199*, 6623–6644. https://doi.org/10.1007/s11229-021-03084-w

Veissière, S. P., Constant, A., Ramstead, M. J., Friston, K. J., & Kirmayer, L. J. (2020). Thinking through other minds: A variational approach to cognition and culture. *Behavioral and Brain Sciences*, *43*, 1–21.

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, 1–35.

Wiese, W., & Friston, K. J. (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, S., Deane, G., Nave, K., & Clark, A. (2019). Getting warmer: Predictive processing and the nature of emotion. *The value of emotions for knowledge* (pp. 101–119). Palgrave Macmillan.

Williams, D. (2018). Predictive processing and the representation wars. *Minds and Machines*, *28*, 141–172

Williams, D. (2021). Is the brain an organ for free energy minimisation? *Philosophical Studies*, *195*, 2459. http://dx.doi.org/10.1007/s11098-021-01722-0

Woodward, J. (2003). *Making things happen*. Oxford University Press.

Yon, D., Heyes, C., & 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.

Open Peer Commentary

Recurrent, nonequilibrium systems and the Markov blanket assumption

Miguel Aguilera and Christopher L. Buckley

School of Engineering and Informatics, University of Sussex, Falmer, Brighton BN1 9QJ, UK
sci@maguilera.net, C.L.Buckley@sussex.ac.uk
https://maguilera.net/, https://christopherlbuckley.com/

doi:10.1017/S0140525X22000309, e184

Abstract

Markov blankets – statistical independences between system and environment – have become popular to describe the boundaries of living systems under Bayesian views of cognition. The intuition behind Markov blankets originates from considering acyclic, atemporal networks. In contrast, living systems display recurrent, nonequilibrium interactions that generate pervasive couplings between system and environment, making Markov blankets highly unusual and restricted to particular cases.

In the target article, Bruineberg and colleagues disrupt current debates about the role of Markov blankets in demarcating the boundaries between living systems and their environments. The authors accurately describe the gap between a Markov blanket as a useful property for statistical inference and the more ontologically loaded concept in the free energy principle (FEP), as a boundary *within* which Bayesian inference occurs. While the arguments pursued by the target article are both correct and important, we think that a fundamental concern remains unaddressed as the paper tacitly accepts (as generally the FEP does) that Markov blankets can be identified largely on the basis of the *structural* connectivity of a system (as opposed to its *functional* connectivity).

A Markov blanket is defined as a set of variables (the "blanket") that separates two other sets of variables within a system, rendering them *conditionally independent*. That is, if the state of the blanket is fixed, the first set of variables (e.g., an agent) becomes independent of the second set (e.g., the environment). This property, also known as the *global Markov condition* (Richardson & Spirtes, 1996), depends on the (conditional) functional couplings describing the statistical interdependencies of a system. Markov blankets were initially introduced in the context of Bayesian networks (Pearl, 1988), which take the form of *directed acyclic graphs*. In such acyclic models, Markov blankets can be directly identified by applying a simple rule to the structural connectivity alone (e.g., Fig. 1A). In particular, the Markov blanket of a set of nodes x contains the parent nodes of x, the children nodes of x and the parents of each child (where children and parents of x are defined as the nodes with incoming/outgoing connections from/to x). This specific sparse structural connectivity is defined as the *local Markov condition* (Richardson & Spirtes, 1996).

The FEP derives much of its intuitions about Markov blankets from acyclic models. However, the theory takes the idea much further, both philosophically and mathematically. The FEP often considers the local Markov condition sufficient for a Markov blanket (Friston, 2013, 2019), suggesting that a boundary between system and environment arises naturally from this sparse structural connectivity as in directed acyclic graphs, without considering functional dynamics. Recent works have refined this argument, and justify a similar equivalence of Markov blankets and structural connectivity under an asymptotic approximation to a weak-coupling equilibrium (Friston et al., 2021b, see Eq. [S8] in Supplementary material). Under this assumption, some works have predicted that Markov blankets will be commonplace in adaptive systems, for example, in brain networks (Friston et al., 2021b; Hipólito et al., 2021).

Under assumptions of either acyclic models or asymptotic equilibrium, previous works have focused solely on the structural connectivity between system elements. However, living systems present two crucial properties that make the occurrence of Markov blankets difficult: (1) they display *cycles* in the form of both reentrant connectivity and loops of interaction with their environment, and (2) they behave *far from equilibrium*, usually exhibiting asymmetric interactions both between components and with the environment. These properties make the relationship between structural and functional connectivity non-trivial.

In a recent article (Aguilera, Millidge, Tschantz, & Buckley, 2022), we studied analytically the existence of Markov blankets in nonequilibrium linear systems with recurrent connections. In these systems, their cyclic, asymmetric structure propagates reverberant activity system-wide, generating couplings beyond their structural connectivity. As a consequence, for most parameter configurations of a system, the sparse connectivity of the local Markov condition does not result in a Markov blanket. That is, even if a system only interacts with the environment via a physical boundary (e.g., a cell membrane or a perception–action interface), it will in general not display the conditional independence associated with a Markov blanket, a crucial issue that has been ignored in the FEP literature until very recently (Aguilera et al., 2022; Biehl, Pollock, & Kanai, 2021).

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

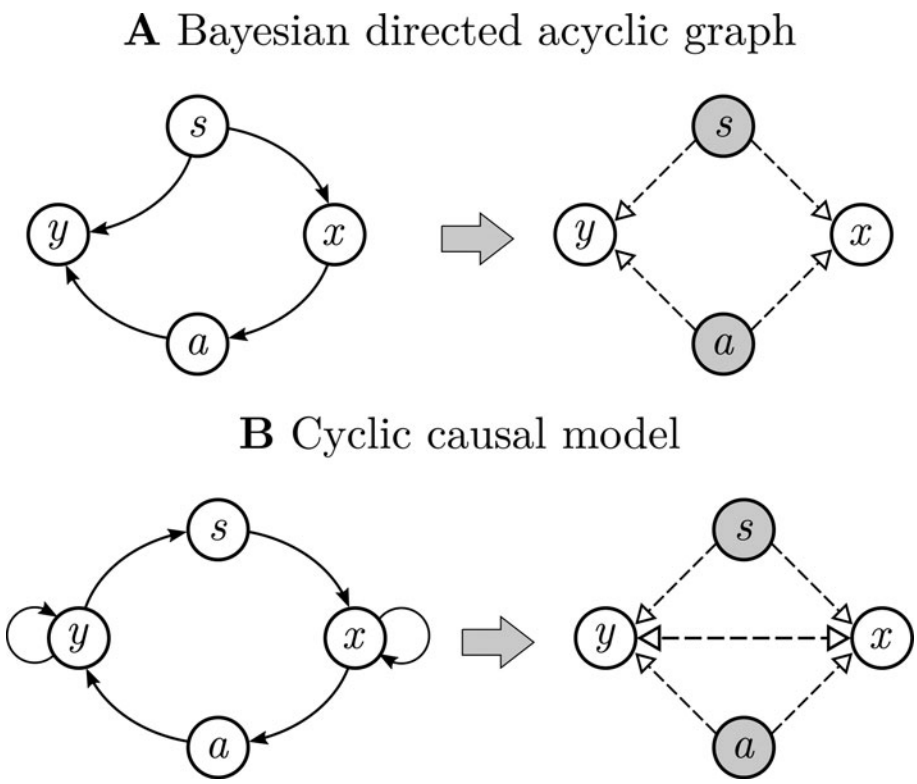


Figure 1 (Aguilera and Buckley). Structural and functional couplings in cyclic versus acyclic networks. The left-hand figures show the structural connectivity of directed graphs. The right-hand figures show the conditional functional couplings of the system when the state of the "blanket" *s*, *a* is fixed. In directed acyclic graphs (A), the structural and functional couplings are directly related, and fixing the boundary results in conditional independence of *x*, *y*, yielding a Markov blanket. In directed cyclic graphs (B), the recurrent structural connections result in additional functional couplings between variables, generating a new coupling between *x*, *y* that "crosses" the boundary, therefore not resulting in a Markov blanket in general.

These results raise fundamental and primary concerns about the frequent use of Markov blankets, not only in the FEP, but more generally as an explanatory concept for natural phenomena. A recent article (Friston, Da Costa, & Parr, 2021a) has suggested that additional conditions (a sparsity of solenoidal couplings, a type of dynamical flows driving systems out of equilibrium) guarantee the emergence of Markov blankets, our study shows that these additional conditions become even more unlikely in the presence of recurrent connectivity in the studied non-equilibrium dynamics (Aguilera et al., 2022). It is important to note that these studies were restricted to linear systems and the generalization of these conclusions to nonlinear systems is yet to be studied. However, one could expect that nonlinear interactions might create a larger gap between intuitions drawn from structural considerations and actual functional couplings in the system.

These results do not imply that recurrent, nonequilibrium systems can never display Markov blankets. Our point, however, is that this only happens for highly specific cases, and certainly does not straightforwardly follow from the identification of a physical boundary. Therefore, it cannot be taken for granted that biological systems operate in this narrow parameter space. Such a finding would have significant implications for the physics of biological systems. Nevertheless, without evidence of this, debates about the implications of Markov blankets for living systems seem presumptuous and risk relegating the role of Markov blankets in elucidating the properties of living systems to the level of a potentially misleading metaphor rather than a verifiable hypothesis.

Financial support. M. A. is funded by the European Commission's under a Marie Skłodowska-Curie Action (grant agreement 892715). C. L. B. is supported by BBRSC grant BB/P022197/1.

Conflict of interest. None.

References

Aguilera, M., Millidge, B., Tschantz, A., & Buckley, C. L. (2022). How particular is the physics of the free energy principle? *Physics of Life Reviews*, *40*, 24–50. https://doi.org/10.1016/j.plrev.2021.11.001

Biehl, M., Pollock, F. A., & Kanai, R. (2021). A technical critique of some parts of the free energy principle. *Entropy*, *23*(3), 293. https://doi.org/10.3390/e23030293

Friston, K. J. (2013). Life as we know it. *Journal of the Royal Society Interface*, *10*(86), 20130475. https://doi.org/10.1098/rsif.2013.0475

Friston, K. J. (2019). A free energy principle for a particular physics. *ArXiv*:1906.10184 [q-Bio]. http://arxiv.org/abs/1906.10184

Friston, K. J., Da Costa, L., & Parr, T. (2021a). Some interesting observations on the free energy principle. *Entropy*, *23*(8), 1076. https://doi.org/10.3390/e23081076

Friston, K. J., Fagerholm, E. D., Zarghami, T. S., Parr, T., Hipólito, I., Magrou, L., & Razi, A. (2021b). Parcels and particles: Markov blankets in the brain. *Network Neuroscience*, *5*(1), 211–251. https://doi.org/10.1162/netn_a_00175

Hipólito, I., Ramstead, M. J. D., Convertino, L., Bhat, A., Friston, K., & Parr, T. (2021). Markov blankets in the brain. *Neuroscience & Biobehavioral Reviews*, *125*, 88–97. https://doi.org/10.1016/j.neubiorev.2021.02.003

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

Richardson, T. S., & Spirtes, P. (1996). *Automated discovery of linear feedback models*. Carnegie Mellon (Department of Philosophy).

The seductive allure of cargo cult computationalism

Micah Allen[a,b,c]

[a]Aarhus Institute of Advanced Studies, Aarhus University, 8000 Aarhus, Denmark; [b]Center of Functionally Integrative Neuroscience, Aarhus University, 8000 Aarhus, Denmark and [c]Cambridge Psychiatry, University of Cambridge, Cambridge CB2 8AH, UK
micah@cfin.au.dk
https://www.the-ecg.org/

doi:10.1017/S0140525X22000279, e185

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

Abstract

Bruineberg and colleagues report a striking confusion, in which the formal Bayesian notion of a "Markov blanket" has been frequently misunderstood and misapplied to phenomena of mind and life. I argue that misappropriation of formal concepts is pervasive in the "predictive processing" literature, and echo Richard Feynman in suggesting how we might resist the allure of cargo cult computationalism.

> *The first principle is that you must not fool yourself – and you are the easiest person to fool.*
>
> — *Richard Feynman (1974)*

In their compelling arguments, Bruineberg and colleagues reveal how the mathematical, Bayesian construct of a "Markov blanket" has become a go-to *explanans* in topics as far-reaching as neuroscience, sociology, the philosophy of mind, and epistemology. Through careful analyses of the prerequisite formalisms, they reveal how many of these works confuse a realist understanding of Markov blankets with their actual properties as defined by formal mathematics. A consequence of this confusion is that Markov blankets are ascribed properties they do not possess and are frequently leveraged to explain phenomena for which they have little direct relevance. Indeed, it has been argued Markov blankets demarcate the definition of life (Kirchhoff, Parr, Palacios, Friston, & Kiverstein, 2018), can guide an Asimov-like attempt at psychohistory (Allen, 2018; Ramstead, Badcock, & Friston, 2018; Veissière, Constant, Ramstead, Friston, & Kirmayer, 2019), and even extend cognition to plants (Calvo & Friston, 2017). Bruineberg and colleagues argue that these and many similar arguments fail to capture what is and is not offered by Markov blankets and proffer a helpful framework for understanding and applying Markovian concepts, based on an informed analysis of their actual formal properties.

As I read the target article, I could not help but think of the late Richard Feynman's now infamous remarks, delivered to the Caltech class of 1974, in which he warned of the dangers of "cargo cult science" (Feynman, 1974). Prior to the commencement address, Feynman visited the Esalen Institute, a well-known nexus for "alternative" science. He recounted how, to his surprise, many of the advocates of esoteric mysticism and parapsychology he met there inevitably presented their ideas as scientific, when they were clearly anything but. But perhaps more worrying, he also remarked how many of the styles of argumentation he encountered could also be found in mainstream fields of psychology, neuroscience, and even physics. Feynman dubbed these trends as "cargo cult science" and outlined how to identify and avoid becoming one.

What then, is cargo cult science? Cargo cults were first described during the Second World War, when Melanesian and other Pacific Islanders sought to capture the technological and economic powers of the Allied and Japanese forces who would frequently land there to trade cargo for goods. Convinced by their spiritual leaders that such awesome wealth and technological power would be shared with them, the islands formed ritualistic cults fetishizing outward characteristics of the foreign powers. By wearing their uniforms, making totem rifles, and marching around the beach, these cults hoped that the gods would also bestow upon them the same powers of those they imitated.

Much like these namesake cults, Feynman described cargo cult science as generally being that which sought the appeal and authority of the scientific method, but which failed to live up to its standard in several key regards. First and foremost, a key quality lacking in cargo cult science was a radical commitment to scientific integrity – a commitment to acting as one's own harshest critic. Other key signs of possible cultism included: (1) a failure to engage critically with both the strengths and faults of any theoretical or empirical postulate, (2) an insistence on doing pseudo-experiments that could not have come out otherwise, (3) a kind of ahistorical perspective in which key data points or advances are overlooked or ignored entirely, and crucially, (4) a surface appeal to explanatory devices or scientific concepts without a deeper engagement.

In recent years, I have observed a steady growth of these sorts of errors in the predictive processing literature. Chief among these is a casual, devil-may-care appropriation of computational concepts and the outward appearances of computationalism without a deeper engagement. This takes several forms: For example, the description of pseudo-equations as Bayesian "models,"[1] or the frequent introduction of new psychological "theories" rehashing concepts such as "priors," "prediction errors," "precision," or "Markov Blankets" as explanatory in and of themselves.[2] A more recent trend is found in the burgeoning volume of so-called "*in silico*" demonstrations, wherein an off the shelf Markovian model is merely re-parameterized and then described as a new "model" of some complex phenomenon – emotion (Hesp et al., 2021), ecological niche construction (Bruineberg, Rietveld, Parr, van Maanen, & Friston, 2018), or interoception (Allen, Levy, Parr, & Friston, 2019) are all salient examples. Typically, such demonstrations involve minimal reshaping of the underlying models themselves, which is remarkable considering the breadth of topics to which they are applied.

A similar error can be found frequently in empirical studies of various psychological phenomena – a kind of computational prestidigitation, in which a construct such as "precision" is appealed to, an experiment is conducted, and then a paper produced that proudly claims to have provided evidence for the underlying computational theory. For example, experiment in which attention to the body is manipulated, and some consequent alteration in an ambiguous data feature is observed, which is then interpreted unambiguously as "evidence for precision weighting" (Petzschner et al., 2019). While predictive processing will certainly claim the credit here, it seems obvious that in the absence of an actual model fitting procedure, just about any psychological or computational theory could explain the obtained results. This trick is pervasive in a new flood of psychological and neuroscientific experiments in which attention, expectation, confidence, or other concepts with a similar sounding cousin in formal theory of can be found and manipulated, a high impact paper produced, and no attempt at true falsification made.

The issue is of course that across all these examples there is a failure to engage critically and directly with the underlying formal constructs, and a commensurate failure to apply appropriate computational methods to enable falsification and ultimately, safeguard against pseudoscience. Critical steps for establishing Feynman's radical honesty – such as model cross-validation,

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

model falsification, or even fitting to empirical data at all, are few and far between (Palminteri, Wyart, & Koechlin, 2017; Wilson & Collins, 2019). It is salient then that in the same lecture, Feynman famously warned of his "first principle." Too many of us increasingly risk violating these ideals, perhaps in hopes of riding the Bayesian wave to the promised land of high impact computational neuroscience papers. I applaud Bruineberg and colleagues for showing us how to leave the cargo cult behind.

Financial support. MA is supported by a Lundbeckfonden Fellowship (under Grant R272-2017-4345), and the AIAS-COFUND II fellowship programme that is supported by the Marie Skłodowska-Curie actions under the European Union's Horizon 2020 (under Grant 754513), and the Aarhus University Research Foundation.

Conflict of interest. None.

Notes

1. See, e.g., "This model is formalized by the following equation: P(Mommy|Interoception, Exteroception) ∝ P(Mommy) × P(Interoception|Mommy) × P(Exteroception|Mommy)," from Atzil, Gao, Fradkin, & Barrett (2018). See another similar example in Allen and Tsakiris (2018).
2. Aptly dubbed the "Bayes Glaze" by anonymous twitter commentator, @Neuroskeptic.

References

Allen, M. (2018). The foundation: Mechanism, prediction, and falsification in Bayesian enactivism: Comment on "Answering Schrödinger's question: A free-energy formulation" by Maxwell James Désormeau Ramstead et al. *Physics of Life Reviews*, *24*, 17–20. https://doi.org/10.1016/j.plrev.2018.01.007

Allen, M., Levy, A., Parr, T., & Friston, K. J. (2019). In the body's eye: The computational anatomy of interoceptive inference. *BioRxiv*, 603928. https://doi.org/10.1101/603928

Allen, M., & Tsakiris, M. (2018). The body as first prior: Interoceptive predictive processing and the primacy. In M. Tsakiris, & H. De Preester (Eds.), *The interoceptive mind: From homeostasis to Awareness* (p. 27). Oxford Univeristy Press.

Atzil, S., Gao, W., Fradkin, I., & Barrett, L. F. (2018). Growing a social brain. *Nature Human Behaviour*, *2*(9), 624–636. https://doi.org/10.1038/s41562-018-0384-6

Bruineberg, J., Rietveld, E., Parr, T., van Maanen, L., & Friston, K. J. (2018). Free-energy minimization in joint agent-environment systems: A niche construction perspective. *Journal of Theoretical Biology*, *455*, 161–178. https://doi.org/10.1016/j.jtbi.2018.07.002

Calvo, P., & Friston, K. (2017). Predicting green: Really radical (plant) predictive processing. *Journal of the Royal Society Interface*, *14*(131), 20170096. https://doi.org/10.1098/rsif.2017.0096

Feynman, R. P. (1974). Cargo cult science. *Engineering and Science*, *37*(7), 10–13.

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

Kirchhoff, M., Parr, T., Palacios, E., Friston, K., & Kiverstein, J. (2018). The Markov blankets of life: Autonomy, active inference and the free energy principle. *Journal of the Royal Society Interface*, *15*(138), 20170792. https://doi.org/10.1098/rsif.2017.0792

Palminteri, S., Wyart, V., & Koechlin, E. (2017). The importance of falsification in computational cognitive modeling. *Trends in Cognitive Sciences*, *21*(6), 425–433. https://doi.org/10.1016/j.tics.2017.03.011

Petzschner, F. H., Weber, L. A., Wellstein, K. V., Paolini, G., Do, C. T., & Stephan, K. E. (2019). Focus of attention modulates the heartbeat evoked potential. *NeuroImage*, *186*, 595–606. https://doi.org/10.1016/j.neuroimage.2018.11.037

Ramstead, M. J. D., Badcock, P. B., & Friston, K. J. (2018). Answering Schrödinger's question: A free-energy formulation. *Physics of Life Reviews*, *24*, 1–16. https://doi.org/10.1016/j.plrev.2017.09.001

Veissière, S. P. L., Constant, A., Ramstead, M. J. D., Friston, K. J., & Kirmayer, L. J. (2019). Thinking through other minds: A variational approach to cognition and culture. *Behavioral and Brain Sciences*, *43*, e90. https://doi.org/10.1017/S0140525X19001213

Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *ELife*, *8*, e49547. https://doi.org/10.7554/eLife.49547

Making reification concrete: A response to Bruineberg et al.

Mel Andrews

Department of Philosophy, The University of Cincinnati, Cincinnati, OH 45221, USA
mel.andrews@tufts.edu
www.mel-andrews.com

doi:10.1017/S0140525X22000310, e186

Abstract

The principal target of this article is the reification Bruineberg et al. perceive of formalism within the literature on the variational free energy minimization (VFEM) framework. The authors do not provide a definition of reification, as none yet exists. Here I offer one. On this definition, the objects of the authors' critiques fall short of full-blown reification – as do the authors themselves.

Scientific modelling is a bit like playing a game of pretend. We play soldiers and pretend our sticks to be guns; we play explorers and pretend the sand to be lava. Misrepresentations, distortions, and untruths run rampant in scientific models. Such idealisations are harmless, so long as we do not forget that we have made them; so long as we do not forget that the sand is really sand and our friends really our friends.

Imagining the developing organism to be a generative model, or the brain a predictive engine can be scientifically fecund. It requires, however, a bit of suspension of disbelief. It requires the scientist to dream up a mapping between two things which are fundamentally unalike; to envision the world as though it really were a directed acyclic graph, a phase space, Shannon information measures over a Riemannian manifold. We come to talk as though these were one and the same.

Formal modelling, as with science in general, aims at uncovering causal patterns (Potochnik, 2017). Philosophers of science hold scientific models to consist of interpreted structure (Weisberg, 2013). Take structure to be a mathematical system; an interpretation or construal thereof relates this formal structure to a target phenomenon of interest. Model structures bear no intrinsic representation relations to targets (Nguyen & Frigg, 2021). We stipulate a relation of partial representation between model structure and target in order to put the model to work. Crucially, there will be features of any model structure that do not map onto any feature of the systems we utilise them to investigate. This makes modelling prone to undergo reification.

The mere employment of an idealisation or the application of an idealised model is not yet a reification, however; nor is the mixing of math and metaphorical language. Reification is the mismapping of formal structure onto target phenomena – or theoretical representation thereof – in a manner that leads us to misapprehend the causal structure of nature. Reification is definitionally an epistemic error.

The central move in Bruineberg et al. is to distinguish between a strictly formal reading of the Markov blanket construct, as developed in Pearl (1988), and a conceptually – even metaphysically – laden notion, as wielded under the variational free energy

minimization (VFEM) framework. It is this "reification of the Markov blanket construct" in the research programme centred around this second notion that they denounce.

> "[M]any authors in the field," they write, "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."

One way to view this is as a multifold reification. But I think that there is a much more natural reading available, namely that the Markov blanket is simply a mathematical fixture of the VFEM framework – held as contentless formal modelling framework. It is utilised to build many different models that differ with respect to their conceptual content.

I have argued (Andrews, 2021) that the VFEM framework is a mathematical model structure and can do no conceptual or philosophical work on its own. The interpreted formal structure of VFEM models, however, can. Such models typically operate in an intuition-pumping capacity; they function to shape our conceptual grasp on the causal structure of target systems of interest, ranging from microphage to macrocosm (Kirchhoff, Parr, Palacios, Friston, & Kiverstein, 2018; Kuchling, Friston, Georgiev, & Levin, 2019; Rubin, Parr, Da Costa, & Friston, 2020). What Bruineberg et al. take to be assertions about nature are mere modelling gambits.

There is a crucial distinction to be drawn between claims made about the causal structure of nature that result from a modelling exercise and idealisations or gambits involved in the modelling procedure which may be literally untrue in reference to the model's target – and which, indeed, hold no pretence to truth (Potochnik, 2017). Not all scientific modelling efforts aim to furnish such assertions, however. Some modelling work furthers our epistemic aims of latching onto causal patterns in the world by enabling us to reenvision the form that such causal architectures could take. All modelling aims at facilitating human understanding, but not all modelling aims at truth. This entails that we will have different degrees of commitment to the posits involved in our modelling efforts.

Interpreted model structure does not consist of knowledge about the natural world. The interpretation of a formal model merely maps it onto our conceptual representations of things which we hold to exist in nature. Facts, or tentative facts, are only generated when a model is put into appropriate coordination with an empirical measurement procedure. Not all scientific modelling, however, is in dialogue with data in this manner. We run into trouble when we treat modelling practices not aimed at knowledge – conceived as sufficiently true causal patterns – as generating knowledge.

Game-theoretic modelling in the social sciences or optimality modelling in biology do not on their own generate knowledge of natural systems, because they are not coordinated with the results of empirical measurement procedures. The use of VFEM models occupies a similar epistemic position. So long as we keep this context in view, the conceptual exploration facilitated by such models is harmless.

Bruineberg et al. understand VFEM models to be after claims about the causal structure of systems in the world because they implicitly take the epistemic utility of modelling to reside solely in offering truth-evaluable assertions about nature. The loose, analogical reasoning of the literature surrounding the VFEM formalism strikes them as would-be statements of fact. Were this the case, it would indeed be a reification, and an egregious one, at that. I want us to resist, however, the temptation to read the literature in this light.

Indeed, if reification were merely talk of the world as though it had formal properties, or talk of formal structure as though it had empirical or conceptual content, Bruineberg et al. would themselves be guilty of the reification fallacy, for they speak of Pearl's (1988) Markov boundary as being "substantiated by the empirical literature," and address its "scientific credibility." The existence and use of a formal tool, however, cannot receive empirical substantiation.

Thus the takeaway of this exposition is not that the VFEM framework is unscientific or bad science, or that it does not deliver what its proponents and architects want of it, but rather that its ambitions were (like most science) far less ambitious to begin with than philosophers had hoped. The other key lesson is to caution VFEM modellers against use of overly suggestive language, lest they send the philosophers into an even deeper state of befuddlement. Such conceptual promiscuity, especially in the treatment of heavily idealised models, opens the door to reification.

Conflict of interest. None.

References

Andrews, M. (2021). The math is not the territory: Navigating the free energy principle. *Biology & Philosophy*, *36*(3), 1–19.

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

Kuchling, F., Friston, K., Georgiev, G., & Levin, M. (2019). Morphogenesis as Bayesian inference: A variational approach to pattern formation and control in complex biological systems. *Physics of Life Reviews*, *33*, 88–108.

Nguyen, J., & Frigg, R. (2021). Mathematics is not the only language in the book of nature. *Synthese*, *198*(24), 5941–5962.

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

Potochnik, A. (2017). *Idealization and the aims of science*. University of Chicago Press.

Rubin, S., Parr, T., Da Costa, L., & Friston, K. (2020). Future climates: Markov blankets and active inference in the biosphere. *Journal of the Royal Society Interface*, *17*(172), 20200503.

Weisberg, M. (2013). *Simulation and similarity: Using models to understand the world*. Oxford University Press.

Markov blankets and Bayesian territories

Jeff Beck

Department of Neurobiology, Duke University, Durham, NC 27710, USA
jeff.beck@duke.edu

doi:10.1017/S0140525X22000619, e187

Abstract

Bruineberg et al. argue that one ought not confuse the map (model) for the territory (reality) and delineate a distinction between innocuous Pearl blankets and metaphysically laden Friston blankets. I argue that all we have are models, all knowledge is conditional, and that if there is a Pearl/Friston distinction, it is a matter of the domain of application: latents or

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

observations. This suggests that, if anything, Friston blankets may inherit philosophical significance previously assigned to observations.

Models models everywhere

"Do not mistake the map for the territory." This critique is meant to be a generic criticism of drawing conclusions about the real world from one's scientific model. The map is the model and the territory is the target of the modeling endeavor.

This instrumentalist view is consistent with the Bayesian view, which holds that models serve only to represent or summarize statistical regularities present in observations. A model is never correct or true, it is simply more likely than other models under consideration. Finite data and failure to consider all possible models render metaphysical (*a priori* true) conclusions unavailable. Paraphrasing Laplace and Maxwell, *all knowledge is conditional*. Bayesians have priors and likelihoods with relative truth values that are conditioned on the data. Mathematicians have axioms (priors) and tools for generating additional truths conditioned on those axioms (likelihoods with logical structure). Philosophers specify relationships between and properties of intuitive "concepts" (likelihoods) and then prune and organize them to fit into axiomatic logical decision trees (priors). For example, Russell's paradox arises from the intuition that sets can be defined solely by a list of properties. His theory of types resolves this by introducing a hierarchal generative model for sets and sets of sets (Russell, 1903). Zermelo's solution was to say that elements of a set can have properties that further divide them into subsets. This corresponds to generative models for subsets that have the philosophically preferred logical tree structure (Zermelo, 1908).

Regardless, *all we have are models. This is as true for the philosopher as it is for the mathematician and scientist*. The above criticism reduces to the true but irrelevant statement that metaphysical (*a priori*) conclusions cannot be drawn. Ultimately, I believe the problem here is that "reality" is an evocative but poorly defined term that should be avoided. Territory, on the other hand, may be a sensible notion and Markov blankets can play a role in its identification. In the Bayesian framework, models, hypotheses, parameters, latents all receive the same treatment. There is only one quantity that has a special role: the observations or data (D). A Bayesian doesn't care about which model is correct or what values the parameters and latents take. This is explicit in posterior predictive modeling, which marginalizes out these details to generate a summary of previously observed data in the form of a prediction about future or unobserved data given the set of models under consideration, that is, $p(D'|D,\{M\})$. If observations are the territory, then this posterior predictive object has the appealing property that it is conditioned on the territory and its domain is the territory. Alternatively, it takes in something that can be called "facts" and generates predictions about future or unobserved "facts." Model details simply provide a language to talk about relationships between observations. For example, suppose some observations of d_1 and d_2 are linearly correlated. A model with independent latent z that linearly drives d_1 and d_2 provides a compact language that describes the observed correlation. The posterior predictive formulation also demonstrates that *science is ultimately concerned with prediction and data compression and nothing else. Models simply provide a language for talking about relationships between observations*.

Markov blankets define objects

Markov blankets have two domains of application: latent variables of a model or observations. Historically, Pearl blankets are applied to latent variabes in a model. This allows us to define macroscopic objects (collections of latents) within a model and establish a taxonomy of objects defined by the statistics of their boundaries. It also has the potential to establish a language for discussing similarities between models in the same way we discuss motifs of connections between latents. When applied directly to observations, as in "The Markov blankets of life" (Kirchhoff, Parr, Palacios, Friston, & Kiverstein, 2018), Markov blankets identify macroscopic objects in the territory.

By virtue of the epistemic seal, the blanket also defines the territory of the macroscopic object. As with Pearl blankets, this can lead to a taxonomy of objects defined by the statistics of their boundaries. The presence of a Markov blanket in our observations also tells us that we can ignore the territory within the boundary without consequence to our ability to predict what is going on outside the boundary, so long as the boundary persists and is observed. This is why it is generally irrelevant that protons are made up of a zoo of more fundamental things. Regardless, this line of reasoning establishes the link between Markov blankets and "thingness" and defines a thing by the relationship between the statistics of its boundary and the statistics of its environment. *It also suggests that if there is any philosophical distinction to be made between Friston and Pearl blankets, that distinction is derived or inherited from the domain of application*.

Regarding the notion of "inference with a model." I do not view this as a categorically unique thing. The free-energy principle offers up a normative description of the behavior of objects (defined by their Markov blankets) in the language of agent–environment interaction. That is, objects (1) form beliefs about the external world, (2) use that information to predict changes in the boundary, and (3) act to affect the boundary (and indirectly the external world) in a way that drives boundary statistics to a desired stationary distribution. This is necessary because boundary maintenance and object identity are inexorably linked.

However, *from the complete class theorem, we know that the language of agent–environment interaction is uniformly applicable to coupled systems. Thus, the relatively innocuous statement that objects are defined by the statistics of the interactions between their boundary and their environment is equivalent to the statement that objects perform inference with a generative model and "act" to enforce a particular statistical relationship between boundary and environment*. That said, the moral of this story seems to be that the simple language of object–environment interactions should be preferred over the language of agent–environment interactions because the latter tends to generate confusion and unnecessary philosophical reaction. Still, it is just two ways of saying the same thing (Ramstead, Friston, & Hipólito, 2020).

Financial support. I received no funding for this effort.

Conflict of interest. None.

References

Kirchhoff, M., Parr, T., Palacios, E., Friston, K., & Kiverstein, J. (2018). The Markov blankets of life: Autonomy, active inference and the free energy principle. *Journal of the Royal Society Interface*, *15*(138), 20170792. doi:10.1098/rsif.2017.0792

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

Ramstead, M. J. D., Friston, K. J., & Hipólito, I. (2020). Is the free-energy principle a formal theory of semantics? From variational density dynamics to neural and phenotypic representations. *Entropy (Basel)*, *22*(8), 889. doi:10.3390/e22080889

Russell, B. (1903). Appendix B: The doctrine of types. In B. Russell (Ed.), *The principles of mathematics* (pp. 523–528). Cambridge University Press.

Zermelo, E. (1908). Untersuchungen über die Grundlagen der Mengenlehre I. *Mathematische Annalen*, *65*(2), 261–281.

Redressing the emperor in causal clothing

Victor J. Btesh[a], Neil R. Bramley[b] and David A. Lagnado[a]

[a]Experimental Psychology Department, University College London, London WC1H 0AP, UK and [b]Psychology Department, University of Edinburgh, Edinburgh EC8 9JZ, UK
victor.btesh.19@ucl.ac.uk; neil.bramley@ed.ac.uk; d.lagnado@ucl.ac.uk
https://www.bramleylab.ppls.ed.ac.uk/;
https://www.ucl.ac.uk/lagnado-lab/david_lagnado.html

doi:10.1017/S0140525X22000176, e188

Abstract

Over-flexibility in the definition of Friston blankets obscures a key distinction between observational and interventional inference. The latter requires cognizers form not just a causal representation of the world but also of their own boundary and relationship with it, in order to diagnose the consequences of their actions. We suggest this locates the blanket in the eye of the beholder.

Bruineberg et al. argue for a crucial distinction between inference *with* and *within* a model, with Pearl blankets pertaining to the former and Friston blankets the latter. However, any set of variables in a graphical model possesses a Pearl blanket (which therefore says nothing about system boundaries), while Friston blankets are taken to pick out living subsystems of a larger ecosystem. Unfortunately, Friston blankets have been applied almost as liberally as their statistical counterparts, including to individual neurons (Palacios, Isomura, Parr, & Friston, 2019), body substructures such as the brain (Seth & Friston, 2016), and eyes (Parr & Friston, 2018) as well as larger organisms (Buckley, Kim, McGregor, & Seth, 2017; Veissière, Constant, Ramstead, Friston, & Kirmayer, 2019). This plurality of *blankets* is acknowledged by Parr (2020) and celebrated by Kirchhoff, Parr, Palacios, Friston, and Kiverstein (2018) as evidence for the ubiquity of the free-energy principle (FEP). We contend that this flexibility in what is cast as internal, external, sensory, or active states, is dangerously confused; it gives the false impression that the theory can recruit causal concepts, for example, Markov blankets, without committing to the full implications of a causal model-based understanding of perception and action.

The causal nature of the world is implicit in active inference, where sensory states are depicted as caused by external states that are, in turn, causally influenced by active states (Friston, Daunizeau, & Kiebel, 2009; 2011). However, Friston et al. (2009) propose that agents do not represent the world as such, but simply as a statistical coupling between the distribution of internal and external states through the blanket states. Worryingly, FEP theorists assume this is sufficient for agents to evaluate the consequences of their actions (Ramstead, Kirchhoff, & Friston, 2020), and do everything else associated with cognition such as thinking, planning, imagining, and explaining (Sloman & Lagnado, 2015). While Ramstead et al. (2021) claim that the recognition density (the agents' approximate distribution over external states conditional on sensory states) represents the world, nothing is said about how this density encodes causal relations that are separable from actions and sensations. If the self-evidencing agent only represents relationships between their active and sensory states, and not the external world of causes that give rise to these, how can they arbitrate between inputs caused by their own actions and those that "would have happened anyway," for example, those caused by ongoing dynamics out in the world? How too are they to do the myriad other things we associate with cognition?

In other words, active inference seems to conflate two different forms of inference. One is simply conditioning one's internal model on observations to update probabilities and make predictions. This includes both inferring likely consequences of observations – if the light turns on, we predict that the room is illuminated – but also their likely causes – that someone else must be home and have turned on the switch. A much-discussed limitation of such "passive" learning is that it struggles to answer questions about causal directionality (Bramley, Dayan, Griffiths, & Lagnado, 2017; Lagnado & Sloman, 2004, 2006; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003). Thus, a second form of inference is through active interventions, local alterations to the world that allow the learner to identify causal effects – for example, that the switch controls the light rather than the reverse. Clearly, if they then conclude that the light coming on means someone else is home, or that turning on the light would make someone else appear, they would have made a foundational mistake. Learning from intervention, or imagining actions, requires updating one's model in a more sophisticated way than simply conditioning on observations (Pearl, 2009). One must represent one's own action as coming from outside the system being modelled. This is a subtlety that active inference overlooks but one that humans are highly sensitive to (Bramley, Lagnado, & Speekenbrink, 2015; Bramley et al., 2017; Bramley, Gerstenberg, Mayrhofer, & Lagnado, 2018, 2019; Hagmayer, Sloman, Lagnado, & Waldmann, 2007; Lagnado & Sloman, 2004; Rothe, Deverett, Mayrhofer, & Kemp, 2018; Sloman & Lagnado, 2005). Even rats are sensitive to the distinction between light or noise as signals (for food) or as consequences of their own action, that is, pressing a button (Blaisdell, Sawa, Leising, & Waldmann, 2006; Clayton & Dickinson, 2006). To avoid interpreting the consequences of their own actions as signals for food, rats must treat themselves as independent from the light–food system. Critically, whether a sensory input is perceived as observational or interventional is agent-relative. One agent's intervention is, from the perspective of another agent, a worldly cause. This highlights that deciding what falls inside or outside a system's boundaries is a modelling choice that depends on the goal of the modeller and so does not resolve questions about actual physical boundaries.

To exhibit adaptive behaviour in a causal world, cognizers should not only approximate the expected observational distribution of external states but also the expected distribution under potential actions. This latter task requires that cognizers treat themselves as separate from the system they are learning about. To choose and evaluate the effect of its actions, an agent must perform inference *with* a model encoding asymmetric causal relations – in the sense that only actions on causes influence effects but not the reverse (Griffiths & Tenenbaum, 2005, 2009; Lagnado, Waldmann, Hagmayer, & Sloman, 2007; Tenenbaum, Griffiths, & Kemp, 2006) and should exhibit behaviour aimed at disambiguating these asymmetries. As

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

such, we suggest that the notion of Markov blankets is critical to the agent's model of its own interactions with the world. In this sense, both the agent and the theorist describing it are performing inference *with* a model, and the cognition-relevant blankets are those that are properties of self-world representations rather than ontological features of living systems.

To sum up, we agree that casting behaviour as action–perception loops has yielded theoretical insights into self-regulatory (Barrett, 2017; Pezzulo, Rigoli, & Friston, 2015; Seth & Friston, 2016) and habitual behaviour (Friston et al., 2015, 2016). However, we fear that inattention to causal representational structure means active inference suffers the same pitfalls as predictive processing (Sloman, 2013), and behaviourism before it, consigned to explain only simple autonomic or reflex behaviours and not those that make intelligent systems such fascinating and unique parts of the natural world.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Conflict of interest. None.

References

Barrett, L. F. (2017). The theory of constructed emotion: An active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, *12*(1), 1–23. https://doi.org/10.1093/scan/nsw154

Blaisdell, A. P., Sawa, K., Leising, K. J., & Waldmann, M. R. (2006). Causal reasoning in rats. *Science (New York, N.Y.)*, *311*(5763), 1020–1022. https://doi.org/10.1126/science.1121872

Bramley, N. R., Dayan, P., Griffiths, T. L., & Lagnado, D. A. (2017). Formalizing Neurath's ship: Approximate algorithms for online causal learning. *Psychological Review*, *124*(3), 301–338. https://doi.org/10.1037/rev0000061

Bramley, N. R., Gerstenberg, T., Mayrhofer, R., & Lagnado, D. A. (2019). Intervening in time. In S. Kleinberg (Ed.), *Time and Causality across the Sciences* (pp. 86–115). Cambridge University Press. https://doi.org/10.1017/9781108592703.006

Bramley, N. R., Lagnado, D. A., & Speekenbrink, M. (2015). Conservative forgetful scholars: How people learn causal structure through sequences of interventions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 708–731.

Bramley, N. R., Gerstenberg, T., Mayrhofer, R., & Lagnado, D. A. (2018). Time in causal structure learning. *Journal of Experimental Psychology: Learning Memory and Cognition*, *44*(12), 1880–1910. https://doi.org/10.1037/xlm0000548

Buckley, C. L., Kim, C. S., McGregor, S., & Seth, A. K. (2017). The free energy principle for action and perception: A mathematical review. *Journal of Mathematical Psychology*, *81*, 55–79. https://doi.org/10.1016/j.jmp.2017.09.004

Clayton, N., & Dickinson, A. (2006). Rational rats. *Nature Neuroscience*, *9*(4), 472–474. https://doi.org/10.1038/nn0406-472

Friston, K., Daunizeau, J., & Kiebel, S. J. (2009). Reinforcement learning or active inference? *PLoS ONE*, *4*(7), e6421. https://doi.org/10.1371/journal.pone.0006421

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., O'Doherty, J., & Pezzulo, G. (2016). Active inference and learning. *Neuroscience and Biobehavioral Reviews*, *68*, 862–879. https://doi.org/10.1016/j.neubiorev.2016.06.022

Friston, K., Mattout, J., & Kilner, J. (2011). Action understanding and active inference. *Biological Cybernetics*, *104*(1–2), 137–160. https://doi.org/10.1007/s00422-011-0424-z

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., & Pezzulo, G. (2015). Active inference and epistemic value. *Cognitive Neuroscience*, *6*(4), 187–224. https://doi.org/10.1080/17588928.2015.1020053

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*, 334–384. https://doi.org/10.1016/j.cogpsych.2005.05.004

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review*, *116*(4), 661–716. https://doi.org/10.1037/a0017201

Hagmayer, Y., Sloman, S. A., Lagnado, D. A., & Waldmann, M. R. (2007). Causal reasoning through intervention. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 86–100). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195176803.003.0007

Kirchhoff, M., Parr, T., Palacios, E., Friston, K., & Kiverstein, J. (2018). The Markov blankets of life: Autonomy, active inference and the free energy principle. *Journal of the Royal Society Interface*, *15*(138), 20170792. https://doi.org/10.1098/rsif.2017.0792

Lagnado, D. A., & Sloman, S. A. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning Memory and Cognition*, *30*(4), 856–876. https://doi.org/10.1037/0278-7393.30.4.856

Lagnado, D. A., & Sloman, S. A. (2006). Time as a guide to cause. *Journal of Experimental Psychology: Learning Memory and Cognition*, *32*(3), 451–460. https://doi.org/10.1037/0278-7393.32.3.451

Lagnado, D. A., Waldmann, M. R., Hagmayer, Y., & Sloman, S. A. (2007). Beyond covariation: Cues to causal structure. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (Vol. *44*, pp. 154–172). https://doi.org/10.1093/acprof:oso/9780195176803.003.0011

Palacios, E., Isomura, T., Parr, T., & Friston, K. (2019). The emergence of synchrony in networks of mutually inferring neurons. *Scientific Reports*, *9*(1), 1–14. https://doi.org/10.1038/s41598-019-42821-7

Parr, T. (2020). Choosing a Markov blanket. *Behavioral and Brain Sciences*, *43*, E112. http://dx.doi.org/10.1017/S0140525X19002632

Parr, T., & Friston, K. J. (2018). Active inference and the anatomy of oculomotion. *Neuropsychologia*, *111*(October 2017), 334–343. https://doi.org/10.1016/j.neuropsychologia.2018.01.041

Pearl, J. (2009). *Models, reasoning, and inference* (2nd ed.). Cambridge University Press.

Pezzulo, G., Rigoli, F., & Friston, K. (2015). Active inference, homeostatic regulation and adaptive behavioural control. *Progress in Neurobiology*, *134*, 17–35. https://doi.org/10.1016/j.pneurobio.2015.09.001

Ramstead, M. J. D., Kirchhoff, M. D., Constant, A., & Friston, K. J. (2021). Multiscale integration: Beyond internalism and externalism. *Synthese*, *198*, 41–70. doi: https://doi.org/10.1007/s11229-019-02115-x

Ramstead, M. J. D., Kirchhoff, M., & Friston, K. (2020). A tale of two densities: Active inference is enactive inference. *Adaptive Behavior*, *28*(4), 225–239. https://doi.org/10.1177/1059712319862774

Rothe, A., Deverett, B., Mayrhofer, R., & Kemp, C. (2018). Successful structure learning from observational data. *Cognition*, *179*(March 2017), 266–297. https://doi.org/10.1016/j.cognition.2018.06.003

Seth, A. K., & Friston, K. J. (2016). Active interoceptive inference and the emotional brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371* (1708), 20160007. https://doi.org/10.1098/rstb.2016.0007

Sloman, A. (2013). What else can brains do? *Behavioral and Brain Sciences*, *36*(3), 230–231. https://doi.org/10.1017/S0140525X12002439

Sloman, S. A. & Lagnado, D. A. (2005). Do we "do"? *Cognitive Science*, *29*, 5–39.

Sloman, S. A., & Lagnado, D. A. (2015). Causality in thought. *Annual Review of Psychology*, *66*(1), 223–247. https://doi.org/10.1146/annurev-psych-010814-015135

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E. J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, *27*(3), 453–489. doi: https://doi.org/10.1016/S0364-0213(03)00010-7

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*(7), 309–318. https://doi.org/10.1016/j.tics.2006.05.009

Veissière, S. P. L., Constant, A., Ramstead, M. J. D., Friston, K., & Kirmayer, L. J. (2019). Thinking through other minds: A variational approach to cognition and culture. *Behavioral and Brain Sciences*, *43*, e90: 1–75. doi: https://doi.org/10.1017/S0140525X19001213

The map, the territory, and the cartographer: Linking the "pure" formal models to the "murky" material world

Anna Ciaunica[a,b]

[a]Centre for Philosophy of Science, University of Lisbon, Campo Grande, Edificio C4 - 1749-016 Lisbon, Portugal and [b]Institute of Cognitive Neuroscience, University College London, London WC1N 3AR, UK
a.ciaunica@ucl.ac.uk

doi:10.1017/S0140525X22000590, e189

Abstract

Assigning to Pearl blankets an instrumental, a "pure" formal role, tacitly delegates the thorny question of mapping the "murky" territory to empirical sciences. But this move sidelines the problem, and does not offer a solution to the question: How do we relate the formal properties of an agent's model of the world to the real properties of the world itself?

This interesting paper aims at disentangling two distinct yet related interpretations of what it means to make an inference (a) *within* a model and (b) *with* a model. The authors note that insights resulting from the formal description in (b) are "smuggled in" via (a), leading to unwarranted metaphysical assumptions about what a Market blanket is or does in the real world. In short, the paper convincingly argues that the map should not be conflated with the territory.

How do we connect the map with the territory? And more importantly who makes this mapping and why? Here I suggest that the distinction between making an inference (a) *within* a model and (b) *with* a model, while important, leaves out the models' makers (i.e., the cartographers) and what the formalisms are for (i.e., the "real" world).

While maps/maths (i.e., formal models) can be viewed as significantly more "pure" and less "messy" than the murky empirical territory, without the latter and a cartographer to interpret them, not only they would be worthless, they wouldn't even exist. Maps/maths are indeed artefacts, while the territory is, as the authors acknowledge, "real."

Another way to express this idea is to say that maps/maths are real insofar they also part of the territory as being constructed, made, by someone for a particular purpose. In a way, everything belongs to the territory, even the numbers that a mathematician writes on a paper or on a screen. But in which sense "that" reality is different from the reality of a snake that bites your skin?

The interesting debate sparked by this thought-provoking paper brings us back to at least two classical debates in philosophy. One is between Plato and Aristotle around the notion of "matter" and "form," which is key for our discussion here. Indeed, "Pearl blankets" are after all a formal property of nodes in a Bayesian network where the latter is used as "useful and compact graphical abstractions for studying complex phenomena" (sect. 1, para. 5). These complex phenomena may be things like the behaviour of atomic particles or stock markets, both of which belong to the category "matter."

Aristotle introduces matter and form as contrasting notions, distinct causes, which together make up every ordinary object (Ainsworth, 2020). In doing so, he distances himself from Plato's theory of forms, which exist quite apart from the material world. He does so in part by insisting that "his own forms are somehow enmeshed in matter (*Metaphysics* vi 1 and vii 11, and *De Anima* i 1). He also maintains that all natural forms are like something which is snub, where something is snub only if it is concavity-realized-in-a-nose (*Physics* ii 2; cf. *Sophistical Refutations* 13 and 31)" (Ainsworth, 2020).

This is an important distinction because while in Plato's view forms can exist apart from the material reality (in an "ideal" world), for Aristotle they are intertwined. Coming back to the target paper: The only way to keep the Pearl blankets on the safe territory of pure abstractions as useful tools to tackle complex phenomena is to detach from the murky territory, as Plato does. This is an option, but it is not very informative one. Because what matters in the end is *what we do* with maps/maths in the real world, and how the cartographer enmeshes them with the territory. At some point, the Pearl blankets formalisms will need to meet the territory if they want to provide some useful information about our "real" world, for example, to help us stay safe from snakes and predict stock markets.

A couple of 2000 years or so later, a second classical debate which echoes nicely the debate captured in the target paper, opposed within the Vienna Circle, Schlick to Neurath on the foundations of human knowledge. On the one hand, Schlick notes that "all great attempts at establishing a theory of knowing arise from the problem of the reliability of human cognition, and this problem in turn originates in the wish for *absolute certainty*" (1959, p. 209, my italics). Scientific attempts are in search of "an unshakeable, indubitable foundation, a firm basis on which the uncertain structure of our knowledge could rest (…) the bedrock, which exists prior to all construction and does not itself vacillate" (*ibid.*). One may argue that maths/maps provide such basis via the formal models.

Against this view, Neurath pointed out that our cognitive situation is that of a sailor who "far out at sea, transforms the shape of their clumsy vessel from a more circular to a more fishlike one (…) and must rebuild their ship upon the open sea, never able to dismantle it in dry dock or to reconstruct it there from the best materials" (Neurath, 1944, p. 47). Or to use again our metaphor: One cannot leave the "shacky" territory in order to make a "pure" abstract map/model of it. Not only we cannot leave the territory, but our maps are also parts of the territory, and highly influenced by it.

Here I suggest that both Pearl and Markov blankets have in common the fact that they are maps/models constructed by a cartographer with the purpose of making sense of an open and constantly moving sea. However, while the former aims at sticking to the abstract formalism insisting on its instrumental role, the latter crosses the boundaries between map and territory and dives into the sea, by making ontological claims. These claims may be wrong, the same way many former theories were proven throughout centuries of scientific endeavour. Yet, we do know with certainty that sticking to the maps/maths only will not give us interesting information about the territory, especially if we disregard the cartographer behind it and her relation to the map itself.

Financial support. This work was supported by an FCT grant 2020.02773.CEECIND and an FCT project INTERSELF PTDC/FER-FIL/4802/2020 to AC.

Conflict of interest. None.

References

Ainsworth, T. (2020). *Form vs matter* – Sandford Encyclopaedia of Philosophy. First published February 8, 2016; substantive revision March 25, 2020. Retrieved from https://plato.stanford.edu/entries/form-matter/.

Neurath, O. (1944). Foundations of the social sciences. In O. Neurath, R. Carnap & C. Morris (Eds.), *International encyclopedia of unified science* (Vol. *2*, No. 1, pp. iii+50). University of Chicago Press.

Schlick, M. (1959). The foundation of knowledge. In A. J. Ayer (Ed.), *Logical positivism* (pp. 209–227). The Free Press.

Return of the math: Markov blankets, dynamical systems theory, and the bounds of mind

Lincoln John Colling

School of Psychology, University of Sussex, Falmer BN1 9QH, UK
l.colling@sussex.ac.uk
http://research.colling.net.nz

doi:10.1017/S0140525X2200022X, e190

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

Abstract

Bruineberg and colleagues highlight work using Markov blankets to demarcate the bounds of the mind. This echoes earlier attempts to demarcate the bounds of the mind from a dynamical systems perspective. Advocates of mechanistic explanation have challenged the dynamical approach to independently motivate the application of the formalism, a challenge that Markov blanket theorists must also meet.

The target article highlights work by, for example, Kirchhoff and Kiverstein (2021) and Ramstead, Kirchhoff, Constant, and Friston (2021), where attempts have been made to invoke the mathematical formalism of Markov blankets to justify claims about the bounds of the mind. The strategy highlighted in the target article is, however, not a new one. Rather, it echoes earlier approaches by advocates of dynamical systems theory to similarly use mathematical models of a system's behaviour to justify claims about the bounds of the mind. By placing the target article in the context of this earlier work it might be possible to bring into sharper relief the issues raised in the target article and, more positively, sketch a way forward for Markov blanket theorists.

One well-known model from the dynamical approach to cognition is the Haken–Kelso–Bunz model (Haken, Kelso, & Bunz, 1985), which arose out of empirical work on the dynamics of inter-limb coordination. For example, Kelso (1984) observed that when people rotated their wrists in an anti-phase pattern while gradually increasing the cycling speed a critical point was reached after which there was a rapid breakdown in the anti-phase pattern with coordination rapidly being reestablished in an in-phase pattern. The model was able to formalise this phenomena using the tools of dynamical modelling resulting in a model containing two parameters: one for cycling speed and one for the relative phase between the two limbs. The power of this model is its generality. Not only can it capture the dynamics of inter-limb coordination but it can also be applied to phenomena like socially coordinated motor behaviour. For example, Schmidt, Carello, and Turvey (1990) found that two people seated next to each other, and engaging in anti-phase leg swings, exhibit the same dynamics as inter-limb coordination.

The Haken–Kelso–Bunz model is an example of a nonlinear dynamical system. As Chemero (2011) argues, only linear systems are decomposable while nonlinear systems are non-decomposable. The upshot of this non-decomposability is that the dynamics of nonlinear systems must be modelled in terms of global collective variables, and that it is not possible to model the system dynamics in terms of separate component parts. That is, the mathematical formalism enforces viewing the two people engaged in inter-personal coordination as a single unified system. The boundaries of this system are not located within a single person but encompass both people. Drawing the boundaries of the system at the edge of a person's skull amounts to splitting the system, a move prohibited by the formalism. The move here is echoed by Kirchhoff and Kiverstein (2021) and Ramstead et al. (2021) outlined by the target article. The mathematical formalism of nonlinearly coupled dynamical systems is replaced with the Markov blanket formalism, but the consequence is the same. The formalism used to model the system defines the bounds of the system.

This move is not unproblematic. The explanatory status of dynamical models has been questioned by those advocating for *mechanistic explanation* (e.g., Colling & Williamson, 2014; Kaplan & Bechtel, 2011; Kaplan & Craver, 2011). Although it's not possible to replay these arguments here, I will pick out one point that turns on this explanatory worry. The coupling-induced synchronisation observed in intra- and inter-personal limb movements is also found in other physical systems including ostensibly non-cognitive systems like pendulum clocks. Although the model accurately predicts the dynamics of these systems, the model itself, by avoiding reference to the physical facts of the system, does not allow one to predict which systems might exhibit the relevant dynamics. However, if one does examine the physical facts of the system then it is evident why some systems exhibit the relevant dynamics and others do not. In the example of coupled pendulum clocks an explanation that makes reference to the physical facts of the system, their parts, and their interactions – that is, a mechanistic explanation – provides reasons why certain arrangements of pendulum clocks exhibit the relevant dynamics while other arrangements do not. For example, a sketch of a mechanistic explanation might reference vibrations produced by the clocks and the role the wall plays in transmitting vibrations between clocks. This in turn provides an explanation for why clocks placed side-by-side on the same wall exhibit the relevant dynamics while clocks placed on opposite walls do not. Mechanistic explanations might similarly be furnished for why particular limb movements or interpersonal actions exhibit the relevant dynamics. The fact that the system exhibits these dynamics is only part of the story. What is missing is an explanation of why the system should exhibit these dynamics in the first place. The upshot of this is that what licenses application of the model (and what licenses demarcating the boundaries of the system) are some set of facts about the mechanism.

The move by Kirchhoff and Kiverstein (2021) and Ramstead et al. (2021) to demarcate the boundaries of the mind gives rise to a similar worry. Is there some set of explanatory facts that licences placing the Markov blanket at the brain, the skin, or at any other "boundary"? On this, the answer is not clear. For example, Kirchhoff and Kiverstein (2021) reject the idea of explanation dependent boundaries while Ramstead et al. (2021) appear to at least partially endorse the idea. Kirchhoff, Parr, Palacios, Friston, and Kiverstein (2018) go further and explicitly reference a mechanism sketch in deciding on the location of the Markov blanket (using the example of coupled pendulums). But as the example from dynamical systems theory shows, the formalism itself does not license predictions about which systems are amenable to the formalism and which are. Rather, these predictions are made independently of the formalism on, for example, mechanistic grounds. The Markov blanket theorist is presented with the same challenge. That is, to provide an explanation or prediction of which systems are amenable to the formalism – or because the formalism is applicable to every "thing" (Friston, 2019), which systems are amenable to specific applications of the formalism independent of the particular application of the formalism itself.

Conflict of interest. None.

References

Chemero, A. (2011). *Radical embodied cognitive science*. MIT Press.

Colling, L. J., & Williamson, K. (2014). Entrainment and motor emulation approaches to joint action: Alternatives or complementary approaches? *Frontiers in Human Neuroscience*, *8*, 754. https://doi.org/10.3389/fnhum.2014.00754

Friston, K. (2019). A free energy principle for a particular physics. Retrieved January 3, 2022, from http://arxiv.org/abs/1906.10184

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

Haken, H., Kelso, J. A. S., & Bunz, H. (1985). A theoretical model of phase transitions in human hand movements. *Biological Cybernetics*, *51*(5), 347–356. https://doi.org/10.1007/BF00336922

Kaplan, D. M., & Bechtel, W. (2011). Dynamical models: An alternative or complement to mechanistic explanations? *Topics in Cognitive Science*, *3*(2), 438–444. https://doi.org/10.1111/j.1756-8765.2011.01147.x

Kaplan, D. M., & Craver, C. F. (2011). The explanatory force of dynamical and mathematical models in neuroscience: A mechanistic perspective. *Philosophy of Science*, *78*(4), 601–627. https://doi.org/10.1086/661755

Kelso, J. A. (1984). Phase transitions and critical behavior in human bimanual coordination. *The American Journal of Physiology*, *246*, R1000–R1004. https://doi.org/10.1152/ajpregu.1984.246.6.R1000

Kirchhoff, M. D., & Kiverstein, J. (2021). How to determine the boundaries of the mind: A Markov blanket proposal. *Synthese*, *198*, 4791–4810.

Kirchhoff, M. D., Parr, T., Palacios, E., Friston, K., & Kiverstein, J. (2018). The Markov blankets of life: Autonomy, active inference and the free energy principle. *Journal of the Royal Society Interface*, *15*(138), 20170792. https://doi.org/10.1098/rsif.20170792.

Ramstead, M. J. D., Kirchhoff, M. D., Constant, A., & Friston, K. J. (2021). Multiscale integration: Beyond internalism and externalism. *Synthese*, *198*(1), 41–70. https://doi.org/10.1007/s11229-019-02115-x

Schmidt, R. C., Carello, C., & Turvey, M. T. (1990). Phase transitions and critical fluctuations in the visual coordination of rhythmic movements between people. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(2), 227–247. https://doi.org/10.1037/0096-1523.16.2.227

Nothing but a useful tool? (F)utility and the free-energy principle

Matteo Colombo

Tilburg Center for Logic, Ethics and Philosophy of Science, Tilburg University, 5000 LE Tilburg, The Netherlands
m.colombo@uvt.nl
https://mteocolphi.wordpress.com/

doi:10.1017/S0140525X22000346, e191

Abstract

Bruineberg and collaborators distinguish three philosophical positions about the status of Markov blankets in the context of active inference modelling, namely: literalism, realism, and instrumentalism. They criticize the first two positions and suggest that instrumentalism is "less problematic but also less interesting" (sect. 6.1.2, para. 5) Here, I sketch how literalists and realists might reply to Bruineberg et al.'s criticisms, and I explain why instrumentalism is more interesting and contentious than what Bruineberg and collaborators suggest.

Bruineberg and collaborators distinguish three philosophical positions about the status of Markov blankets in active inference modelling, namely: literalism, realism and instrumentalism (sect. 6.1). Literalism is the view that the world is fundamentally a Markov blanket; so, everything in it – including your brain, my tortoises, and the symbols you are currently reading on this page – is a Markov blanket or grounded in a Markov blanket. Realism concerns the relation of active inference models to the world. It says that, for any active inference model, there is some mapping between some of its theoretical posits and some worldly features. Particularly, it says that the Markov blankets posited by any of these models can be mapped onto some boundaries of some objects. Instrumentalism concerns the relation of Markov blankets in active inference models both to the world and to their users. It presumes that at least some active-inference models are useful for achieving some epistemic or pragmatic goal.

Though these *-isms* might come across as futility, they foreground very general questions about the point of the free-energy principle (FEP) and its relevance for understanding life and mind, the utility and correct way of interpreting active inference modelling, and the rational epistemic attitude towards the content of the theories and models grounded in the FEP.

Bruineberg et al. criticize literalism because it reifies abstract mathematical structures and is "removed from the empirical and naturalistic research programme that FEP purports to be" (sect. 6.1.2, para. 6). Yet literalists may reply they are engaged in revisionary metaphysics that makes no appeal to the supernatural. Noticing the FEP is a research programme in mathematical physics quite removed from empirical data (Colombo & Palacios, 2021), literalists may point out reification (or Platonism) is widespread among mathematicians to make sense of their achievements, and may also draw an analogy with computation, showing how cellular automata have been stripped of their "metaphysical modesty" for arguing that the universe is fundamentally a cellular automaton (e.g., Wolfram, 2002; Zuse, 1970).

Bruineberg et al. criticize realism because the Markov blanket formalism doesn't tell modellers how to find "a non-arbitrary mapping that is privileged for principled reasons" (sect. 6.1.2, para. 7). As this criticism accepts that idealized active inference models can be evaluated for their accuracy, realists may insist that somebody's finding a boundary (counter)intuitive has no bearing on when an active inference model is accurate. Insofar as a given active inference model *is* accurate and the Markov blankets it posits enjoy referential success, then the system being modelled does possess the properties to which the formal structure of the model successfully refers – whether counterintuitively or not.

Perhaps, Bruineberg et al.'s criticism is not that realism and literalism are counterintuitive or incoherent, but that these positions in the FEP literature are often based on bad (or no) arguments uninformed by relevant results in metaphysics and the philosophy of modelling about, say, the boundaries of objects (e.g., Varzi, 2011), the notions of *structure*, *thing*, and *fundamentality* (e.g., Sider, 2020) or the bearing of imagination and fiction on scientific modelling (e.g., Levy & Godfrey-Smith, 2019). Or perhaps, Bruineberg et al.'s criticism is that, to account for the achievements of active inference modelling, we should focus attention on the utility of such models rather than their accuracy; and to explain their utility, it's irrelevant whether Markov blankets are "fundamental," "real," or "fictitious."

This last idea seemingly coheres with Bruineberg et al.'s recommended instrumentalism. But any plausible, instrumentalist position towards a scientific model is premised on the success of that model in furthering some scientific aim. The problem, in the context of the FEP, is that it's unclear how active inference modelling is successful.

It's uncontroversial that Markov blankets, and other statistical and algorithmic tools for causal search, discovery, and inference often play an incredibly useful role in helping scientists to represent and study systems of interest at a suitable scale, make reliable predictions, discover causal mechanisms, and facilitate interventions in the world (e.g., Marinescu, Lawlor, & Kording, 2018; Spirtes, Glymour, Scheines, & Heckerman, 2000). It's also uncontroversial that some tools such as the digital computer (and methods for inferential statistics) have historically inspired new theories like the computational theory of mind (Gigerenzer,

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

1991). What is contentious is whether successful active inference models are *nothing but* instruments for prediction, control, or achieving some other aim, and whether the theoretical claims of these models constitute knowledge of the world (including of its unobservable aspects). Traditionally, instrumentalists don't limit themselves to make the banal claim that maths and stats are useful tools to do science. They want to make sense of successful scientific practices, accounting for what warrants scientists' reliance on empirically successful models in inquiry (cf. Psillos, 1999; Stanford, 2006). So, any plausible instrumentalist position towards an active inference model, or FEP more generally, should presume the model under consideration is useful, successful, or furthers some scientific aim.

While computational neuroscientists, as well as modellers in other sciences have a diversity of aims (Kording, Blohm, Schrater, & Kay, 2018; Potochnik, 2017), widely shared modelling aims include empirical adequacy (Van Fraassen, 1980, pp. 11–13), novel predictions (Lakatos, 1978, pp. 31–34), and guidance on how to intervene in the world (Cartwright, 2007, Ch. 3; Woodward, 2003, pp. 7–9). If it's unclear that active inference models are empirically adequate, make novel predictions, and guide cognitive and life scientists to successfully intervene in the world, then there is no obvious question for instrumentalists (and scientific realists) to address about the epistemic status of these models.

In fact, when they describe Baltieri, Buckley, and Bruineberg's (2020) active inference model of the Watt governor, Bruineberg et al. give us reason to believe that active inference modelling is *not* a useful approach for studying and understanding any self-organizing system. As Bruineberg et al. themselves recognize, it's unclear "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 minimization" (sect. 7, para. 2). If the utility and empirical successes of active-inference modelling are contentious, then an instrumentalist position towards the FEP and the modelling practices it grounds cannot be as unproblematic and uninteresting as Bruineberg et al. suggest.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Conflict of interest. None.

References

Baltieri, M., Buckley, C. L., & Bruineberg, J. (2020). Predictions in the eye of the beholder: an active inference account of Watt governors. *arXiv preprint arXiv:2006.11495*.

Cartwright, N. (2007). *Hunting causes and using them: Approaches in philosophy and economics*. Cambridge University Press.

Colombo, M., & Palacios, P. (2021). Non-equilibrium thermodynamics and the free energy principle in biology. *Biology & Philosophy*, *36*(5), 1–26.

Gigerenzer, G. (1991). From tools to theories: A heuristic of discovery in cognitive psychology. *Psychological Review*, *98*,254–267.

Kording, K., Blohm, G., Schrater, P., & Kay, K. (2018). Appreciating diversity of goals in computational neuroscience. Preprint: https://doi.org/10.31219/osf.io/3vy69

Lakatos, I. (1978). Falsification and the methodology of scientific research programmes. In J. Worrall & G. Currie (Eds.). *The methodology of scientific research programmes* (philosophical papers I) (pp. 8–101). Cambridge University Press.

Levy, A., & Godfrey-Smith, P. (Eds.). (2019). *The scientific imagination*. Oxford University Press.

Marinescu, I. E., Lawlor, P. N., & Kording, K. P. (2018). Quasi-experimental causality in neuroscience and behavioural research. *Nature Human Behaviour*, *2*(12), 891–898.

Potochnik, A. (2017). *Idealization and the aims of science*. University of Chicago Press.

Psillos, S. (1999). *Scientific realism: How science tracks truth*. Routledge.

Sider, T. (2020). *The tools of metaphysics and the metaphysics of science*. Oxford University Press.

Spirtes, P., Glymour, C. N., Scheines, R., & Heckerman, D. (2000). *Causation, prediction, and search*. MIT Press.

Stanford, P. K. (2006). Instrumentalism. In S. Sarkar & J. Pfeifer (Eds.), *The philosophy of science: An encyclopedia* (pp. 400–405). Routledge.

Van Fraassen, B. C. (1980). *The scientific image*. Oxford University Press.

Varzi, A. C. (2011). Boundaries, conventions, and realism. In J. K. Campbell, M. O'Rourke, & M. H. Slater (Eds.), *Carving nature at its joints* (pp. 129–153). MIT Press.

Wolfram, S. (2002). *A new kind of science*. Wolfram Media.

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford University Press.

Zuse, K. (1970). *Calculating space*. MIT Press.

Markov blankets and the preformationist assumption

Mads Dengsø[a], Ian Robertson[a] and Axel Constant[b]

[a]Faculty of the Arts, Social Sciences and Humanities, University of Wollongong, Wollongong, NSW 2522, Australia and [b]Charles Perkins Centre, The University of Sydney, Sydney, NSW 2006, Australia
madsdengsoe@gmail.com
ianrob@uow.edu.au
axel.constant.pruvost@gmail.com

doi:10.1017/S0140525X22000358, e192

Abstract

Bruineberg and colleagues argue that a realist interpretation of Markov blankets inadvertently relies upon unfounded assumptions. However, insofar as their diagnosis is accurate, their prescribed instrumentalism may ultimately prove insufficient as a complete remedy. Drawing upon a process-based perspective on living systems, we suggest a potential way to avoid some of the assumptions behind problems described by Bruineberg and colleagues.

Bruineberg and colleagues contend that so-called "Friston blankets" introduce a number of "non-arbitrary assumptions" in applying Markov blankets to the boundaries of living systems (sect. 4, paras. 1 and 2). The application of Markov blankets to living systems requires prior observations providing "a principled justification for why to start from one particular model rather than a different one" (sect. 6, para. 1). In this sense, they conclude, Markov blankets owe part of their explanatory power to these prior assumptions to a point where "it is not clear that the Markov blanket formalism is doing much additional work" (sect. 6.1.2, para. 7).

If the application of Markov blankets to living systems is indeed determined by such underlying assumptions, this would seem to imply that at least some of the confusions that Bruineberg and colleagues have set out to untangle run deeper than our attitudes toward Markov blankets.

If so, then a strong instrumentalism about Markov blankets may itself be insufficient as a measure to untangle the root causes of the confusions between realist and instrumentalist readings of Markov blankets (see Andrews [2021] and Kirchhoff, Kiverstein, & Robertson [2022] for recent discussion of realism and instrumentalism *qua* free-energy principle [FEP] models). Besides the eternal vigilance demanded by our models and metaphors, we

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

may need to reevaluate some of the starting observations informing their application.

The assumption that the organism and the environment constitute two conditionally independent interactants defines many Bayesian approaches to living systems, including the Fristonian one targeted by Bruineberg and colleagues. This guiding assumption behind designating living systems in terms of an inner organism contraposed by an outer environment may be interpreted as a variant of preformationism: The notion that organisms and environments constitute and should be evaluated in our theorizing as separate entities with inherent properties, and whose interaction is essentially secondary to their independent existence (see Anderson, 2017; see also Oyama, 2000).

This assumption, of a pre-established conditional independence between organisms and their respective environments, presents a potential point of theoretical (Colombo & Wright, 2021) and empirical (Aguilera, Millidge, Tschantz, & Buckley, 2021) incongruity between Markov blankets and the essentially coupled character of sensorimotor interfaces. It has moreover been brought into question by more recent accounts emphasizing the constitutive role that interaction plays in producing and sustaining the separate forms of organism and environment (see, e.g., Bruineberg & Rietveld, 2019; Gallagher & Allen, 2018; Kirchhoff & Kiverstein, 2019, 2021).

We believe that the risk of preformationism echoes earlier debates within the literature in that it "force[s] us to recognize that the picture of biological agents as free-energy-minimizing systems requires something closer to a process-based (rather than a static or state-based) ontology" (Clark, 2017, p. 17). In this regard, existing accounts have already shown how temporally deep hierarchical models provide for adaptive models with less sharp distinctions between organisms and their environments (see Kirchhoff, Parr, Palacios, Friston, and Kiverstein, 2018).

What we here want to briefly suggest is that a process-based perspective may furthermore avoid preformationism not only in the application of Markov blankets, but also at the level of the underlying assumptions that inform this application.

The sort of process-based perspective that we have in mind serves to preclude preformationism specifically by reconceptualizing stabilized forms on either side of the (Markov) boundary as products of ongoing exchanges that serve to perpetuate the living system. That is, under a process-based perspective on living systems, we may understand the organism and its respective environment not as a preformed substance but as an ensemble of processes (e.g., metabolism). The process view we refer to echoes the view of process ontology that takes processes – instead of substantive forms – as the fundamental unit of analysis in biology. Process ontology seeks to reverse the explanatory relation between entities and processes: Rather than explaining processes in terms of interactions between distinct entities, process ontology explains entities as relatively stable phases of continuous processes (Nicholson & Dupré, 2018; see also Griffiths & Stotz, 2018).

Narratively, as applied to active inference, a process-based perspective conceptualizes organismic boundaries as "hard-won achievements" of living systems (Kirchhoff & Kiverstein, 2019; see also Kirchhoff, 2015; Sutton, 2010). This reversal is decisive for at least one of the underlying assumptions that Bruineberg and colleagues ascribe to Friston blankets: It eliminates the need for the assumption of a preformed organism *qua* model and environment *qua* modeled distal world, which arguably commits Friston blankets (and other Bayesian accounts) to a particular variant of substantialist realism. In its stead, processes are what is taken to be the fundamental unit of biological analysis. Under a process-based view, then, one need not assume the organism and environment since these may be derived from the continuous exchanges.

While Bruineberg and colleagues' prescribed strong instrumentalism might still furnish us with helpful resources for clearing up confusions surrounding the application of Markov blankets to living systems, we find that some such confusions may still be traced to the prior observations that inform this application. We believe that a process-based perspective may aid us in upending a central assumption that prefigures some of the forms of confusion targeted by Bruineberg and colleagues. While a far cry from absolving us of the duty to attend to other crucially important issues pointed out by Bruineberg and colleagues in their insightful target article, we nonetheless believe that critically assessing the starting assumptions underlying these issues may ultimately prove to be indispensable in their resolution.

Financial support. This work was supported by the Australian Research Council Discovery Project Minds in Skilled Performance (IR, grant number DP170102987) and by the Australian Laureate Fellowship Project A Philosophy of Medicine for the 21st Century (AC, grant number FL170100160) and by a Social Sciences and Humanities Research Council (SSHRC) doctoral fellowship (AC, grant number 752-2019-0065).

Conflict of interest. None.

References

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.

Anderson, M. L. (2017). Of Bayes and bullets: An embodied, situated, targeting-based account of predictive processing. In T. Metzinger & W. Wiese (Eds.), *Philosophy and predictive processing: 4*. MIND Group. https://doi.org/10.15502/9783958573055

Andrews, M. (2021). The math is not the territory: Navigating the free energy principle. *Biology & Philosophy*, *36*(3), 1–19.

Bruineberg, J., & Rietveld, E. (2019). What's inside your head once you've figured out what your head's inside of. *Ecological Psychology*, *31*(3), 198–217.

Clark, A. (2017). How to knit your own Markov blanket: Resisting the second law with metamorphic minds. In T. Metzinger & W. Wiese (Eds.), *Philosophy and predictive processing: 3*. MIND Group. https://doi.org/10.15502/9783958573031

Colombo, M., & Wright, C. (2021). First principles in the life sciences: The free-energy principle, organicism, and mechanism. *Synthese*, *198*(14), 3463–3488.

Gallagher, S., & Allen, M. (2018). Active inference, enactivism and the hermeneutics of social cognition. *Synthese*, *195*(6), 2627–2648.

Griffiths, P., & Stotz, K. (2018). Developmental systems theory as a process theory. In D. J. Nicholson & J. Dupré (Eds.), *Everything flows: Towards a processual philosophy of biology* (pp. 225–245). Oxford University Pres.

Kirchhoff, M., Kiverstein, J., & Robertson, I. (2022). *The literalist fallacy & the free energy principle: Model-building, scientific realism and instrumentalism*. [Preprint].

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

Kirchhoff, M. D. (2015). Species of realization and the free energy principle. *Australasian Journal of Philosophy*, *93*(4), 706–723.

Kirchhoff, M. D., & Kiverstein, J. (2019). *Extended consciousness and predictive processing: A third-wave view*. Routledge.

Kirchhoff, M. D., & Kiverstein, J. (2021). How to determine the boundaries of the mind: A Markov blanket proposal. *Synthese*, *198*(5), 4791–4810.

Nicholson, D. J., & Dupré, J. (2018). A manifesto for a processual philosophy of biology. In D. J. Nicholson & J. Dupré (Eds.), *Everything flows: Towards a processual philosophy of biology* (pp. 4–45). Oxford University Press.

Oyama, S. (2000). *The ontogeny of information*. Duke University Press.

Sutton, J. (2010). Exograms and interdisciplinarity: History, the extended mind, and the civilizing process. In R. Menary (Ed.), *The extended mind* (pp. 189–225). MIT Press.

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

There is no "inference within a model"

Marco Facchin

Department of Human and Life Sciences, L&PIC (Linguistic and Philosophy IUSS Center), University School for Advanced Studies (IUSS) Pavia, Pavia 27100, Lombardy, Italy
marco.facchin@iusspavia.it
marco.facchin.marco.facchin@gmail.com

doi:10.1017/S0140525X22000085, e193

Abstract

I argue that there is no viable development of the instrumentalist *inference within a model* research program. I further argue that both Friston and Pearl blankets are not the right sort of tool to settle debates on philosophical internalism and externalism. For these reasons, the *inference within a model* program is far less promising than the target article suggests.

In this commentary, I want to focus on the *inference within a model* research program, and briefly argue for two claims. If correct, these claims suggest that Bruineberg and co-authors present the inference within a model research program in a far too bright light.

First claim: Bruineberg and co-authors mischaracterize the instrumentalist development of that research program. They say that it is viable, but uninteresting. I think the opposite is true: It is not a viable development, but, if viable, it would be interesting.

Recall that, as the target article makes amply clear, (a) *Pearl* blankets only capture the patterns of (in)dependencies between variables in a model, which *need not* correspond to real boundaries in the world, and (b) *Pearl* blankets are model-dependent. Now, the instrumentalist inference within a model program uses Pearl blankets as a guide to find real boundaries in the world. Given (a) and (b), the success of such a program seems predicated on having *at least* a reliable rule of thumb to identify patterns of conditional independencies corresponding to real, worldly, systemic boundaries, as well as a reliable rule of thumb to identify the models that accurately capture the real structure of the modeled phenomena (as opposed to merely providing a parsimonious account of the data observed).

These rules of thumb would be interesting epistemic tools in their own right, and their usage would allow us to learn a great deal about the world. Moreover, while possessing these rules of thumb would not allow us to vindicate the most ambitious claims concerning *Friston* blankets in the literature on the free-energy principle (e.g., the claim that physical systems "possess" or "instantiate" Friston blankets, see Friston, 2013), possessing these rules of thumb *would* be sufficient to allow *Pearl* blankets to play the boundary-defining role Friston blankets are currently supposed to play in the philosophical literature on the free-energy principle (Hohwy, 2016, 2017; Kirchhoff & Kiverstein, 2021; Ramstead, Kirchhoff, Constant, & Friston, 2019). This could (in principle) allow us to solve hotly discussed philosophical problems only by "doing the math" as many supporters of the free energy principle claim, which would be an interesting development.

Yet the instrumentalist inference within a model program is not really viable, because it is subtly circular. Recall that a Pearl blanket is defined as the union of three sets: the sets of parents, co-parent, and children *of a target variable*. This means that, in order to identify the Pearl blanket of a variable (or set thereof), we must have already identified the target, "blanketed" variable. If this is correct, the identification of the target variable *logically precedes* the identification of its Pearl blanket. This means that, in order to identify the Pearl blanket of a real-world system, we must have already identified the variable(s) mapping over that system. Hence we *cannot* identify systems, and the various variables describing their behavior, *by identifying their Pearl blankets*, on the pain of circularity. So, although the usage of Pearl blanket to identify the boundaries of a system suggested by the instrumentalist inference within a model *would* have interesting consequences is viable, it does not really seem viable.

Second claim: Neither Friston nor Pearl blankets can be used to *satisfactorily* solve the disputes surrounding various forms of philosophical internalism and externalism.

Consider that Friston and/or Pearl blankets have been used to "identify in a principled manner" all the relevant factors constituting some phenomena of interest (see, e.g., Clark, 2017; Hohwy, 2016; Kirchhoff & Kiverstein, 2021). Consider further that in the relevant literature the presence of a Friston and/or Pearl blanket *defines* what counts as internal or external in the relevant sense. What is "surrounded" by the blanket counts as internal in the relevant sense, the rest counts as external (see Hohwy, 2017, pp. 6–7).

It is easy to see that the conjunction of these two ideas entails that all the factors constituting a phenomenon of interest count, by definition of "internal," as internal. But this means that the conjunction of these two claims makes internalism true *by definition*. If this is a solution to the philosophical internalism/externalism debate, it is not a *satisfactory* solution.

For one thing, the truth value of some forms of internalism and externalism seems to depend on contingent matters of fact. For example, the debate concerning externalism/internalism about the vehicles of cognition and consciousness would surely be solved by the existence of non-biological props able to mimic cerebral processes sufficiently well (see Adams & Aizawa, 2010, p. 78; Vold, 2015). It is hard to see how such a debate, concerning at least in part matters of fact, could be solved by definition.

The same verdict holds for other debates concerning externalism and internalism. The truth value of externalism and internalism about knowledge and mental content, for example, depends on what content and knowledge are; that is, on their nature (cf. Bonjour, 2005; Egan, 2009; Segal, 2009). But the nature of content and knowledge is not something that can be satisfactorily settled *by definition*. Surely no internalist or externalist should be allowed to win the day just by *defining* certain factors as internal or external!

Summarizing: in this commentary, I have tried to argue that the target article is wrong on the instrumentalist development of the inference within a model research program: it would be an exciting research program, if it were viable. But it is not viable. Furthermore, I have tried to argue that Friston and Pearl blankets are the wrong kind of tools when it comes to providing a satisfactory solution to the philosophical debates concerning externalism and internalism. If the arguments I have provided here are correct, the target article mischaracterizes the inference within a model research program: Its chances of success are far slimmer than Bruineberg and co-authors suggest.

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sector.

Conflict of interest. None.

References

Adams, F., & Aizawa, K. (2010). Why the mind is still in the head. In P. Robbins & M. Aydede (Eds.), *The Cambridge handbook of situated cognition* (pp. 78–96). Cambridge University Press.

Bonjour, L. (2005). Internalism and externalism. In P. K. Moser (Ed.), *The Oxford handbook of epistemology* (pp. 234–263). Oxford University Press.

Clark, A. (2017). How to knit your own Markov blanket. In T. Metzinger & W. Wiese (Eds.), *Philosophy and predictive processing: 3* (pp. 1–19). The MIND Group. https://doi.org/10.15502/9783958573031

Egan, F. (2009). Wide content. In A. Beckermann, B. P. McLaughlin & S. Walter (Eds.), *The Oxford handbook of philosophy of mind* (pp. 406–425). Oxford University Press..

Friston, K. (2013). Life as we know it. *Journal of the Royal Society Interface*, *10*(86), 20130475.

Hohwy, J. (2016). The self-evidencing brain. *Noûs*, *50*(2), 259–285.

Hohwy, J. (2017). How to entrain your evil demon. In T. Metzinger & W. Wiese (Eds.), *Philosophy and predictive processing: 2* (pp. 1–15). The MIND Group. https://doi.org/10.15502/9783958573048

Kirchhoff, D. M., & Kiverstein, J. (2021). How to determine the boundaries of the mind: A Markov blanket proposal. *Synthese*, *198*(5), 4791–4810.

Ramstead, M. J., Kirchhoff, D. M., Constant, A., & Friston, K. J. (2019). Multiscale integration: Beyond externalism and internalism. *Synthese*, *198*(1), 41–70.

Segal, G. (2009). Narrow content. In A. Beckermann, B. P. McLaughlin & S. Walter (Eds.), *The Oxford handbook of philosophy of mind* (pp. 425–440). Oxford University Press.

Vold, K. (2015). The parity argument for extended consciousness. *Journal of Consciousness Studies*, *22*(3–4), 16–33.

Practical implications from distinguishing between Pearl blankets and Friston blankets

Stephen Fox

VTT Technical Research Centre of Finland, FI-02150 Espoo, Finland
stephen.fox@vtt.fi

doi:10.1017/S0140525X22000061, e194

Abstract

Analysis provided in The Emperor's New Markov Blankets reveals that there is limited potential for practical application of Pearl and Friston blankets. However, Bruineberg and colleagues' analysis includes a simple diagram that has potential to better enable shared understanding of interactions between free energy principle constructs during the design and implementation of biosocial–technical systems.

A major contribution of Bruineberg and colleagues' analysis is to reduce potential time and confusion involved in trying to work out what might be done in practice with Markov blankets. In particular, they provide succinct comparative analyses of the many different descriptions of Markov blankets. Importantly, their comparative analysis includes illustrative cases, which is a proven strategy for improving learning (Roelle & Berthold, 2015). However, Bruineberg and colleagues' analysis reveals that both Pearl and Friston blankets are of limited usefulness for practice. In particular, Bruineberg and colleagues clarify that Pearl blankets is only an auxiliary construct that can be used to describe conditional independence on random variables. They also clarify that what can be done with Friston blankets in 2021 is a matter of ongoing debate that includes dispute over conceptual issues and mathematical details.

To use an automotive vehicle analogy for brevity, Bruineberg and colleagues' analysis can be distilled into the following summary: Pearl blankets can contribute to describing what is going on "under the hood" while Friston blankets can contribute to describing what is going on "in and around the hood." From an instrumentalist systems design perspective, both Pearl and Friston blankets are human ascriptions made in order to model real-world systems. Colloquial description using the "hood" analogy is not appropriate for natural science and social science, but is appropriate for action science within which the need for practical framings is emphasized (Friedman & Putman, 2014).

The limited practical usefulness of distinguishing between Pearl and Friston blankets can be illustrated with the following example related to functional disorders: that is, medical conditions without complete medical explanation that impair normal functioning of bodily processes (Stone, 2009). Better healthcare systems are needed to provide support for people suffering with functional disorders (Stone, 2016). Gait issues are involved in functional disorders (Espay et al., 2018). Gait is related to personality in ways that are not fully understood (Sun et al., 2018). This ascription problem is exacerbated by the difficulty of defining where one personality type ends and another begins (Haslam, 2019). Furthermore, gait is related to memory in ways that are not fully understood (Michalak, Rohde, & Troje, 2015). This ascription problem is exacerbated by the difficulty of defining what aspects of memory are in the mind and what aspects are in the body (Tozzi, 2014). Thus, distinguishing between what is "under the hood" and what is "in and around the hood" is very difficult. For example, are personality type and body memory "under the hood" while gait is "in and around the hood"? If so, how is the fascia's connection of bones and muscles in gait related to the fascia system holding body memories (Tozzi, 2014)? Should the same fascia be described with both Pearl and Friston blankets? This example illustrates that distinguishing between Pearl and Friston blankets does not necessarily make modelling complex systems any easier. Nor does it necessarily end potential argument that there should only be one type of Markov blanket or potential argument that Markov blankets could be superseded, for example, by causal blankets (Rosas, Mediano, Biehl, Chandaria, & Polani, 2020).

Although reading The Emperor's New Markov Blankets reveals limitations of Pearl and Friston blankets, this does not mean that nothing practical can be done. In particular, it has been argued that it may be useful to frame systems in terms of constructs such as Markov blankets, but without applying all technical details and associated mathematics (Fox, 2021). There is precedence for this in Kurt Lewin's force field analysis diagram being used widely, but many technical details and associated mathematics of his field theory not being used. This is a relevant example as force field analysis is applied to determine interactions between forces for change and forces against change during the evolution of organizations (Burnes & Cooke, 2013).

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

Apropos, Bruineberg, and colleagues provide another useful contribution with their Figure 5, which has potential as a boundary object: that is, as a means of providing meaningful information to different parties who have different backgrounds (Bowker & Star, 2020). In particular, their Figure 5 provides a succinct explanation of interactions between main constructs in the free energy principle. Importantly, it does not involve single line demarcations, which can facilitate the reification fallacy (Mishra & Mishra, 2010). Moreover, it has much potential to provide the basis for an interactive multimodal symbol system, which can facilitate effective communication among people who have different backgrounds (Fox, Moreno, & Vahala, 2019) during the design and implementation of biosocial–technical systems (Fox, Griffy-Brown, & Dabic, 2020). For example, facilitate effective communication among individuals suffering with functional disorders and healthcare practitioners who could provide them with support (Allen, 2009; Stone, 2016).

Financial support. Support was provided from European Commission grant number 952091.

Conflict of interest. None.

References

Allen, D. (2009). From boundary concept to boundary object: The practice and politics of care pathway development. *Social Science & Medicine*, *69*(3), 354–361.

Bowker, G. C., & Star, S. L. (2020). *Sorting things out: Classification and its consequences*. MIT Press.

Burnes, B., & Cooke, B. (2013). Kurt Lewin's field theory: A review and re-evaluation. *International Journal of Management Reviews*, *15*, 408–425.

Espay, A. J., Aybek, S., Carson, A., Edwards, M. J., Goldstein, L. H., Hallett, M., … Morgante, F. (2018). Current concepts in diagnosis and treatment of functional neurological disorders. *JAMA Neurology*, *75*(9), 1132–1141.

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.

Fox, S., Griffy-Brown, C., & Dabic, M. (2020). From socio-technical systems to biosocial technical systems: New themes and new guidance for the field of technology in society. *Technology in Society*, *62*, 101291.

Fox, S., Moreno, M., & Vahala, P. (2019). Innovation symbol system: Multimodal grammars and vocabularies for facilitating mutual innovation knowledge. *Journal of Innovation & Knowledge*, *4*, 12–22.

Friedman, V. J., & Putman, R. W. (2014). Action science. In D. Coghlan & M. Brydon-Miller (Eds.), *The Sage encyclopedia of action research* (pp. 15–18). SAGE.

Haslam, N. (2019). Unicorns, snarks, and personality types: A review of the first 102 taxometric studies of personality. *Australian Journal of Psychology*, *71*(1), 39–49.

Michalak, J., Rohde, K., & Troje, N. F. (2015). How we walk affects what we remember: Gait modifications through biofeedback change negative affective memory bias. *Journal of Behavior Therapy and Experimental Psychiatry*, *46*, 121–125.

Mishra, A., & Mishra, H. (2010). Border bias: The belief that state borders can protect against disasters. *Psychological Science*, *21*(11), 1582–1586.

Roelle, J., & Berthold, K. (2015). Effects of comparing contrasting cases on learning from subsequent explanations. *Cognition and Instruction*, *33*(3), 199–225.

Rosas, F. E., Mediano, P. A., Biehl, M., Chandaria, S., & Polani, D. (2020). Causal blankets: Theory and algorithmic framework. In T. Verbelen, P. Lanillos, C. L. Buckley, & C. De Boom (Eds.), *Active Inference. IWAI 2020. Communications in Computer and Information Science* (Vol. *1326*, pp. 187–198). Springer.

Stone, J. (2009). Functional symptoms in neurology: The bare essentials. *Practical Neurology*, *9*(3), 179–189.

Stone, J. (2016). Functional neurological disorders: The neurological assessment as treatment. *Practical Neurology*, *16*(1), 7–17.

Sun, J., Wu, P., Shen, Y., Yang, Z., Li, H., Liu, Y., … Chen, M. (2018). Relationship between personality and gait: Predicting personality with gait features. In *Proceedings of the 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Madrid, Spain, 3–6 December*, pp. 1227–1231.

Tozzi, P. (2014). Does fascia hold memories? *Journal of Bodywork and Movement Therapies*, *18*(2), 259–265.

Maps and territories, smoke, and mirrors

Karl Friston

Wellcome Centre for Human Neuroimaging, Queen Square Institute of Neurology, University College London, London WC1N 3AR, UK
k.friston@ucl.ac.uk

doi:10.1017/S0140525X22000073, e195

Abstract

It is a pleasure to comment on Bruineberg et al. – who raise some interesting questions of a philosophical and technical nature. I will try to answer three questions posed by the authors. Are Pearl and Friston blankets different things? Are Markov blankets used in an ontological sense? Is there a privileged Markov blanket?

Are Pearl and Friston blankets different things? Yes and no. Markov boundaries, blankets, chains, and fields are just ways of carving nature at its joints, in terms of conditional independencies. For example, the "present" constitutes a Markov blanket that separates the "past" from the "future." Pearl and Friston blankets are just Markov blankets applied in different settings. Pearl blankets (as defined here) are the minimal Markov boundaries in directed acyclic Bayesian networks. Friston blankets (as defined here) are exactly the same thing formulated for dynamic Bayesian networks. This distinction is important because Markov blankets are defined in terms of conditional independencies that require a well-defined probability density. In a dynamic setting, there are two candidates for this probability density: either the (nonequilibrium steady state) solution to density dynamics (e.g., obtained by the Fokker Planck formulation) or the probability density over paths through state space (e.g., obtained via the path integral formulation). It is probably the move to a dynamic setting that has led to puzzlement in the philosophical literature; especially, in understanding the link between sparse coupling in dynamical systems and the ensuing conditional independencies. (This puzzlement generally arises when focusing on linear edge cases; e.g., Biehl, Pollock, & Kanai, 2020.) However, Pearl and Friston blankets are just Markov blankets in the usual Markovian sense (Pearl, 2009).

Are Markov blankets used in an ontological sense under the free energy principle (FEP)? Yes. In the FEP, they are used in an ontologically robust sense, to model the actual boundaries of living systems (in other words, to model features of the territory) (Friston, 2013). One may be a realist or an instrumentalist about this usage (i.e., about the features of the map), but in either case, the aim is to model the actual properties – in this case, the real boundaries – of actual systems.

In relation to the distinction between maps and territories, the FEP says something quite radical, and philosophically significant: it says, heuristically, that to exist at all is to become a map of one's territory (i.e., to entail a generative model of one's environment). It might be helpful to think about map-making as the sense-making implicit in any internal states that are equipped with a Markov blanket. Internal states are effectively generating a coarse-grained map of the territory

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

beyond the Markov blanket. On this reading, the distinction between a (living) map and the accompanying territory only exists in virtue of a Markov blanket, which allows the map to mirror the territory, while allowing a separation of the territory from the map.

Is there a particular Markov blanket that is privileged for principled reasons? No. There is no unique Markov blanket or partition that is privileged under the FEP. However, this does not imply that particular partitions of a system are arbitrary. A failure to appreciate this precludes a proper treatment of the nature of things and, in particular, their scale invariance. Given a set of states at a particular scale, there are potentially many different partitions that have nontrivial Markov blankets. In other words, there are many ways of carving nature at its joints – as illustrated with the knee-jerk example in the target article.

The within-scale composition of Markov blankets is especially important in defining the architecture of generative models entailed by internal dynamics. Not only do Markov blankets define the structure of hierarchical generative models but also – within a hierarchical level – the factorisation afforded by Markov blankets can be read as modularity (Colas, Diard, & Bessiere, 2010; Parr, Sajid, & Friston, 2020b). Indeed, this aspect of Markov blankets speaks to their foundational role in the FEP; in the sense that variational free energy is a functional of a mean field approximation to posterior beliefs – and a mean field approximation entails a factorisation that just is a specification of Markov blankets (Dauwels, 2007; Winn & Bishop, 2005). This factorisation is ontological because it specifies the functional and computational architectures that are realised by (e.g., neuronal) message passing and implicit Bayesian belief updating. This means that the FEP rests on Markov blankets within (the internal states that are enclosed by) Markov blankets (Palacios, Razi, Parr, Kirchhoff, & Friston, 2020).

Finally, at the between-scale level, there is no privileged scale. The deeper question here is how one scale maps to the next and what variational principles are conserved over scales – and the implications for the top-down and bottom-up causation between scales (Kirchhoff, Parr, Palacios, Friston, & Kiverstein, 2018; Palacios et al., 2020; Parr, Da Costa, & Friston, 2020a). Formally, this is probably best dealt with using the apparatus of the renormalisation group. This apparatus has been applied in a realist fashion to the synthetic soup (see Figure 11 in Friston [2019]) and instrumentally in the modelling of neuronal dynamics in the brain (Friston et al., 2021). In brief, it's Markov blankets all the way down – and all the way up.

Financial support. This work was supported by funding for the Wellcome Centre for Human Neuroimaging (Ref: 205103/Z/16/Z); and a Canada–UK Artificial Intelligence Initiative (Ref: ES/T01279X/1).

Conflict of interest. None.

References

Biehl, M., Pollock, F. A., & Kanai, R. (2020). A technical critique of the free energy principle as presented in "Life as we know it." *arXiv e-prints*, arXiv:2001.06408.

Colas, F., Diard, J., & Bessiere, P. (2010). Common Bayesian models for common cognitive issues. *Acta Biotheoretica*, *58*, 191–216.

Dauwels, J. (2007). On variational message passing on factor graphs. *2007 IEEE International Symposium on Information Theory*, pp. 2546–2550.

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

Friston, K. (2019). A free energy principle for a particular physics. *arXiv eprint*, arXiv:1906.10184.

Friston, K. J., Fagerholm, E. D., Zarghami, T. S., Parr, T., Hipólito, I., Magrou, L., & Razi, A. (2021). Parcels and particles: Markov blankets in the brain. *Network Neuroscience*, *5*, 211–251.

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

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.

Parr, T., Da Costa, L., & Friston, K. (2020a). Markov blankets, information geometry and stochastic thermodynamics. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *378*, 20190159.

Parr, T., Sajid, N., & Friston, K. J. (2020b). Modules or mean-fields? *Entropy*, *22*, 552.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Winn, J., & Bishop, C. M. (2005). Variational message passing. *Journal of Machine Learning Research*, *6*, 661–694.

Making life and mind as clear as possible, but not clearer

Alex Gomez-Marin

Instituto de Neurociencias (CSIC-UMH), 03550 San Juan Alicante, Spain
agomezmarin@gmail.com
https://behavior-of-organisms.org/@behaviOrganisms

doi:10.1017/S0140525X22000127, e196

Abstract

Neuroscience needs theory. Ideas without data are blind, and yet mechanisms without concepts are empty. Friston's free energy principle paradigmatically illustrates the power and pitfalls of current theoretical biology. Mighty metaphors, turned into mathematical models, can become mindless metaphysics. Then, seeking to understand everything in principle, we may explain nothing in practice. Life can't live in a map.

In their brilliant but rather undiscovered book, *The Dialectical Biologist*, Richard Levins and Richard Lewontin write: "a theory that can explain everything explains nothing" (1985, p. 65). The paragraph is worth quoting in its entirety, but on another occasion. They refer to the ideological excesses of Marxism, Freudianism, and Darwinism. In the context of twenty-first century life and mind sciences, a similar cautionary note could be made about Fristonism.

Already well-known for his contributions to functional imaging, the British neuroscientist Karl Friston has distinguished himself over the last decade as a steady proponent of the so-called free energy principle (FEP) (Friston, 2010). Friston borrows foundational concepts from thermodynamics and combines them with two good old ideas: Hermann von Helmholtz's understanding of perception as inference (1867) and Claude Bernard's insight on the self-regulatory nature of vital processes (1878). The FEP is a normative theory, and its main commandment is the minimization of surprise. Framed in a modern Bayesian framework – whereby beliefs are updated by means of loops of prediction, error, and correction – the brain would be an active inference machine doing predictive processing.

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

But this is precisely what begs the question. It is not uncommon that a formal approach used by brain scientists to study other brains ends up surreptitiously postulated as what all brains actually do (Gomez-Marin, 2019): If *one can* make sense of brains as inference machines, then *brains must* be inference machines. Indeed, a pervasive temptation among scientists is to treat the conjunction "as if" as "is" (e.g., animals are [like] machines, brains are [like] computers). We know that maps are not terrains, but we still conflate them (Andrews, 2021).

The FEP illustrates what I call "the 3M fallacy," a sleight of mind whereby captivating metaphors are turned into mathematical models that then become covert or self-ignorant metaphysics. Bruineberg et al. thus offer a timely dissection of the notion of Markov blankets as interpreted by some FEP proponents, making a distinction between mathematical tools to study organisms and ontological pronouncements about those very organisms. The critique would not be so patently urgent had Friston and collaborators refrained from elevating their theoretical framework from a powerful heuristic into a sort of biological theory of everything.

Having started as an ambitious integrative theory of brain function, the FEP has been progressively expanded to behavior, the mind, and even consciousness! Perception, decision making and learning would all be explainable under the FEP too, in animals but also in plants and microorganisms. And yet one should be hesitant of any sweeping principle that accounts for nematode foraging and romantic love. Seeking an understanding of virtually everything in principle, the framework risks explaining actually nothing in practice.

Is Friston's FEP neurobiology's "string theory"? The latter has been deemed by prominent physicists as "not even wrong" (Woit, 2006). The FEP is no less ambitious. Sufficiently elaborated for professional acolytes to revere and pursue, it appears sufficiently vague to be immune to empirical data. Without indulging in falsification chauvinism, one should at least ask whether and how the FEP could be wrong.

Perhaps, like Darwin's, Friston's theory may not be disprovable by any one experiment, while still offering a powerful overall way to see things. Back to Freud and Marx, their hypotheses have also tremendously influenced modern thought, whether directly testable or not. Isn't it paradoxical that Darwin's account (1859), arguably the greatest of all biological theories (and firmly grounded in empirical observation), cannot be falsified by a single aberrant finding? Biology is the paradigmatic science of exceptions. Consistency is overrated; coherency is scarce. Overwhelmed by analytical facts, we desperately need synthetic views. Science accomplishes something remarkable when, apart from discovering causal mechanisms, it offers organizing principles. A virtuous middle may be found where castles in the air meet castles made of sand.

Should theoreticians rejoice at mass-producing prêt-à-porter "conceptual blankets," or strive to offer tailor-made suits to the phenomena of life? Concerned with clarity we lose touch with concrete reality, as the physicist's joke of the spherical cow infamously illustrates. Pushing the blanket analogy a bit further, let us ask: How closely fit should the cloth be? A loose one can fit almost everything, but properly fit nothing. A tight fit does a proper job, but just once. The "comfy blanket" may be a suitable garment to watch a movie at home, but a tremendously inappropriate one to attend a funeral. Paradoxically, as Iain McGilchrist remarks, "[k]nowledge of something that is by its nature not precise will *itself* have to be imprecise, if it is to be accurate" (2021, p. 583). Our duty as thinkers is to be as clear as possible, but not clearer.

In that respect, the FEP has remarkable ingredients. It puts action into perception, re-enacting Henri Bergson's pioneering theory of perception as virtual action (1896). It also emphasizes the life–mind continuity (Thompson, 2010), but see Bitbol and Gallagher (2018). The framework, however, leaves crucial biological aspects out. Based on statistical averages in steady states, it struggles to take historicity and individuality into account. Moreover, it overlooks the developmental origin of Bayesian priors (Ciaunica, Constant, Preissl, & Fotopoulou, 2021). Evolutionarily, the FEP focuses on important similarities across species, but dispenses with the relevant differences, say, between a shiitake mushroom and a beluga whale.

In sum, we can appreciate the most fecund aspects of the theory without mistaking stupendous abstractions for concrete reality. Such is the Herculean challenge faced when seeking knowledge of life as we know it versus as it is lived (Canguilhem, 1952). The objectivist stance disregards the perspective of each living organism. "We think in generalities, but we live in detail," wrote Alfred North Whitehead (1926, p. 192). Theories of everything can hardly be theories of *every* thing, since nothing is anything except in its context. Conversely, a theory of every *thing* cannot be a theory of everything since, to our dismay (and sometimes denial), physics does not account for the existence of lived experience. Theory strives for simplicity, but life is complicated. Living organisms can't live in a map.

Acknowledgments. I thank Michel Bitbol, Tim Elmo Feiten, Iain McGilchrist, and Rupert Sheldrake for insightful discussions.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Conflict of interest. None.

References

Andrews, M. (2021). The math is not the territory: Navigating the free energy principle. *Biology & Philosophy*, *36*, 30.

Bergson, H. (1896). *Matière et mémoire: Essai sur la relation du corps à l'esprit*.

Bernard, C. (1878). *Leçons sur les phénomènes de la vie communes aux animaux et aux végétaux*. Baillière.

Bitbol, M., & Gallagher, S. (2018). The free energy principle and autopoiesis. *Physics of Life Reviews*, *24*, 24–26.

Canguilhem, G. (1952). *La connaissance de la vie*. Vrin.

Ciaunica, A., Constant, A., Preissl, H, & Fotopoulou, K. (2021). The first prior: From co-embodiment to co-homeostasis in early life. *Consciousness and Cognition*, *91*, 103117.

Darwin, C. (1859). *On the origin of Species by means of natural selection, or the preservation of favoured races in the struggle for life*. John Murray.

Friston, K. J. (2010). The free energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138.

Gomez-Marin, A. (2019). A clash of Umwelts: Anthropomorphism in behavioral neuroscience. *Behavioral and Brain Sciences*, *42*, E229.

Levins, R., & Lewontin, R. (1985). *The dialectical biologist*. Harvard University Press.

McGilchrist, I. (2021). *The matter with things: Our brains, our delusions, and the unmaking of the world*. Perspectiva.

Thompson, E. (2010). *Mind in life: Biology, phenomenology and the sciences of mind*. Belknap Press, Harvard University Press.

von Helmholtz, H. (1867). *Handbuch der physiologischen optik*. Leopold Voss.

Whitehead, A. N. (1926). The education of an Englishman. *The Atlantic Monthly*, *138*, 192–198.

Woit, P. (2006). *Not even wrong: The failure of string theory and the continuing challenge to unify the laws of physics*. Basic Books.

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

Spatiotemporal constraints of causality: Blanket closure emerges from localized interactions between temporally separable subsystems

Casper Hesp[a,b,c,d]

[a]Department of Psychology, University of Amsterdam, 1098 XH Amsterdam, Netherlands; [b]Amsterdam Brain and Cognition Centre, University of Amsterdam, 1098 XH Amsterdam, Netherlands; [c]Institute for Advanced Study, University of Amsterdam, 1012 GC Amsterdam, Netherlands and [d]Wellcome Centre for Human Neuroimaging, University College London, London WC1N 3BG, UK
c.hesp@uva.nl
twitter: @casper_hesp

doi:10.1017/S0140525X22000620, e197

Abstract

In this commentary, I first acknowledge points of common ground with the target article by Bruineberg and colleagues. Then, I consider how certain ambiguities could be resolved by considering spatiotemporal constraints on causality. In particular I show how blanket closure emerges from localized interactions between temporally separable subsystems, and how this points to valuable directions of future research. Finally, I close with a process note discussing the allegorical implications of the authors' creative title.

In "The Emperor's New Markov Blankets," Bruineberg and colleagues present a stimulating treatment of developments in conceptually distinct uses of the concept of Markov blankets. The authors attribute the original definition of Markov blankets to Pearl and then explain how it differs from a more recent conception of Markov blankets, which they attribute to Friston. Their narrative is of the cautionary sort, warning readers against the conflation of these distinct uses. Bruineberg and I mentioned similar points in our co-authored commentary (Bruineberg & Hesp, 2018) calling researchers to move "beyond blanket terms." It was an early prelude to some of the concerns raised in the target article – namely that associating the physical boundaries of living systems with Markov blankets still leaves us with thorny issues and edge cases, requiring further theoretical, empirical, and computational work. The authors emphasize the following distinction between two "types" of Markov blankets:

(a) Pearl introduced the definition of a Markov blanket as the minimal set of nodes in a directed acyclic graph that render a given target node conditionally independent from all the other nodes in the network: parents, children, and co-parents.
(b) Friston's characterization of Markov blankets emphasizes circular causality by partitioning a given system of interest in terms of its external and internal states, whose recursive influences are mediated by "blanket states," which entail sensory and active states. Given a designated set of internal states, sensory states mediate inward influences (from external to internal states), while active states mediate outward influences (from internal to external states).

In the target article, the authors argue that Pearl's characterization of Markov blankets is more innocuous than that of Friston – as presented in "Life as we know it" (Friston, 2013). I will pay particular attention to the following technical points made by Bruineberg and colleagues:

(1) Friston's formulation focuses explicitly on circular causality and bi-directional connectivity, while Pearl's formulation focuses on directed acyclic graphs.
(2) There is an ambiguous mapping between Friston's formulation and Pearl's definition of a Markov blanket: If the internal states are designated as the target set, then sensory states are parent nodes and active states are child nodes, but this leaves co-parent nodes unaccounted for.
(3) The identification of internal states depends heavily on thresholding parameters and other modelling choices.

First, we consider the causal (in)dependency structure imposed on complex dynamical systems by temporal and physical constraints of interactions. Figure 1 illustrates that a combination of localized interactions combined with a separation of convergence time scales (as induced by the rate parameter in Friston's primordial soup) speaks to the first two points. Firstly, dynamical relationships are causally directed due to the arrow of time and exhibit recurrence when considering multiple time steps. Second, separation of time scales in localized interactions means that all co-parents of a target state are *also* its parents. As shown in Figure 1, these conditions allow for a correspondence between the Markov blanket of a target node – as originally defined by Pearl – and its causal blanket – as defined by Rosas, Mediano, Biehl, Chandaria, and Polani (2020). The significance of blanket leakage versus blanket closure is that unaccounted co-parents will act as confounding variables for any inferential process. Temporal separability minimizes such confounding relationships, affording some "probabilistic grip" to the target

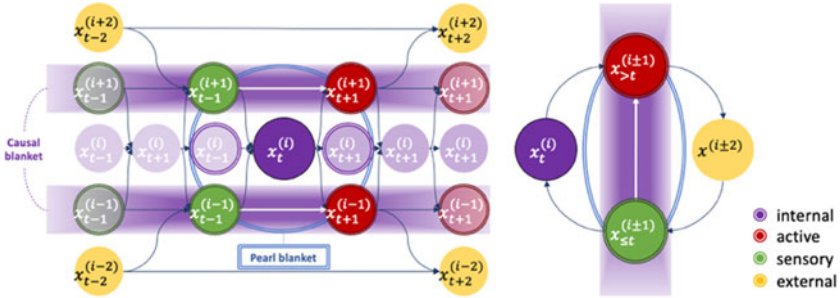


Figure 1 (Hesp). On the left, a directed acyclic graph describing a complex system consisting of five variables at five time points ($t-2$, $t-1$, t, $t+1$, $t+2$), with localized interactions and a separation of time scales where the target node i changes twice as fast as its neighbours ($i-1$, $i+1$), which in turn change twice as fast as their next neighbours ($i-2$, $i+2$). On the top right, the associated "Friston blanket," showing the resulting correspondence between the "Pearl blanket" (in blue) and the causal blanket in purple (as in Rosas et al., 2020), induced by the combination of localized interactions and separation of time scales.

node. As such, investigating the sufficient conditions for the stability of such "blanket closure" would be a valuable avenue for research into the emergence of life.

With respect to the dependence on modelling assumptions, I would echo George Box: "All models are wrong, but some are useful." Methods for partitioning systems will reveal different kinds of information about them, but for given variables of interest they can be evaluated against each other. The authors rightly noted the additional complexity given the fact that predictive accounts of cognition tend to involve "models within models." These nuances do not detract from the utility of such formalisms as modelling heuristics. For any given living system, any model being modelled is – by logical necessity – epistemologically bounded by influences crossing its causal blanket. Furthermore, such models should be biased towards those aspects of the environment that are relevant to organismic integrity and function. Because their capacity to maintain such a probabilistic grip would depend heavily on the stability of blanket closure, this approach naturally emphasizes the functional relevance of autopoiesis and – in extension – self-modelling (Ramstead et al., 2021; Sandved-Smith et al., 2021). At this point, we can consider "models of models within models" to characterize the heterarchical structure of cognition. Perhaps to the frustration of those who prefer philosophical clarity, I would argue that, when territories are devoted to mapping (sub)sections of themselves recursively on different levels of description, maps and territories can mingle – blurring their conceptual boundaries.

Allegorical implications of "The Emperor's New Clothes"

The authors have selected a pithy title that fits with the theme of publicly calling into question a common belief. However, Anderson's original story suggests a much darker allegorical message. Intentional deceit was attributed explicitly to every single character in this story except the "little child," who heroically disrupted the echo chamber. The echoes were started by the weavers, who falsely claimed that "a simpleton, or one unfit for his job would be unable to see the cloth." While everyone was taken hostage by their own social insecurities, only the little child dared to speak out loud.

Transposed to our context, this allegory appears to suggest that researchers have formed an echo chamber – driven by reliance on hearsay and intellectual dishonesty – for fear of being seen as a "simpleton." Bruineberg and colleagues associate their own message with the little child – exposing an obvious lie. The implied accusation appears to run counter to the principle of charity, which is essential for effective academic discourse.[1] Presumably this was not intended, but the authors could have steered clear of such ambiguities by explaining their choice of title in-text – as is common practice when academics use popular references. Hopefully, if nothing else, my commentary could elicit such clarification from the authors.

Financial support. The work of this author is supported by funding from a NWO Research Talent Grant of the Dutch Government (no. 406.18.535).

Conflict of interest. None.

Note

1. The target article is not an isolated case of academics invoking a theme of trickery, even when we constrain ourselves to recent writing on the topic of Markov blankets. For example, Raja et al. (2021; cited in the target article) aimed to expose what they called "the Markov blanket trick."

References

Bruineberg, J., & Hesp, C. (2018). Beyond blanket terms: Challenges for the explanatory value of variational (neuro-)ethology. *Physics of Life Reviews*, *24*, 37–39. https://doi.org/10.1016/j.plrev.2017.11.015

Friston, K. (2013). Life as we know it. *Journal of the Royal Society, Interface*, *10*(86), 20130475. https://doi.org/10.1098/rsif.2013.0475

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. *Physics of Life Reviews*, *39*(2), 49–72. doi:10.1016/j.plrev.2021.09.001

Ramstead, M. J. D., Hesp, C., Tschantz, A., Smith, R., Constant, A., & Friston, K. (2021). Neural and phenotypic representation under the free-energy principle. *Neuroscience & Biobehavioral Reviews*, *120*, 109–122. https://doi.org/10.1016/J.NEUBIOREV.2020.11.024

Rosas, F. E., Mediano, P. A. M., Biehl, M., Chandaria, S., & Polani, D. (2020). Causal blankets: Theory and algorithmic framework. *Communications in Computer and Information Science*, *1326*, 187–198. https://doi.org/10.1007/978-3-030-64919-7_19

Sandved-Smith, L., Hesp, C., Mattout, J., Friston, K., Lutz, A., & Ramstead, M. J. D. (2021). Towards a computational phenomenology of mental action: Modelling meta-awareness and attentional control with deep parametric active inference. *Neuroscience of Consciousness*, *2021*(2). https://doi.org/10.1093/NC/NIAB018m

Free-energy pragmatics: Markov blankets don't prescribe objective ontology, and that's okay

Inês Hipólito[a] and Thomas van Es[b]

[a]Berlin School of Mind and Brain, Humboldt-Universität zu Berlin, 10117 Berlin, Germany and [b]Centre for Philosophical Psychology, Universiteit Antwerpen, 2000 Antwerpen, Belgium
ines.hipolito@hu-berlin.de; thomas.vanes@uantwerpen.be
http://www.ineshipolito.org/

doi:10.1017/S0140525X22000322, e198

Abstract

We target the ontological and epistemological ramifications of the proposed distinction between Friston and Pearl blankets. We emphasize the need for empirical testing next to computational modeling. A peculiar aspect of the free energy principle (FEP) is its purported support of radically opposed ontologies of the mind. In our view, the objective ontological aspiration itself should be rejected for a pragmatic instrumentalist view.

In their impressive paper, Bruineberg et al. make a significant contribution to the free energy principle (FEP) literature by distinguishing between "Pearl blankets" and "Friston blankets," identifying the former as an epistemic tool for Bayesian inference and the latter in terms of its "novel metaphysical use in the free energy framework to demarcate the physical boundary between an agent and its environment" (abstract). Yet the authors have another aspiration. They call out the presupposed legitimacy of extracting ontological predictors from mathematical formalisms, which we applaud.

One thing that is fascinating about the Markov blanket is that this tool allows us to make greater sense of a nested world. Every scale is seen as part of a multiscale network of reciprocal influences interactively shaped by the history of interactions into a common environment. Computational models and simulations can then be viewed as the folk ontology of constructing "imaginary biological populations, imaginary neural networks" (Godfrey-Smith, 2009) to explore the viability of conceiving of cognitive life as active inference under the FEP.

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

But a theory does not reduce to the tools constructed to explore its viability. Tools such as Markov blankets under Bayesian statistics or simulation models are deprived of truth value in themselves outside of the context of the theory. Markov blankets and computational models are built to explore the predictive power of the FEP as a theory of cognition. A theory precedes all the mathematics and computational models in the world. It arises by noticing a pattern, sometimes by what Karl Friston called a "Gerald Durrell" moment (Friston, Fortier, & Friedman, 2018); the FEP, interestingly, first arose to him while preoccupied with some woodlice's antics who were frantically scurrying around trying to find some shade. Just like this, a theory unfolds as discernment of correlations between events or processes of change under philosophical contours and commitments. After all, as Dennett well says, "[t]here is no such thing as philosophy-free science; there is only science whose philosophical baggage is taken on board without examination" (Dennett, 1995, p. 21).

The FEP as a theory of cognition too must answer the empirical test to see if it lives up to its promises. The FEP theory and its models might be mistaken, thus, they must be tested empirically to see whether their predictions are borne out. While the FEP theory may seem plausible, establishing its applications in, say, neurocognitive activity, is not a trivial matter of translating it into models and proclaiming the truth. The FEP, as a theory of cognition, needs to answer to the tedious process of hypothesis and experimental verification. If, for example, a human being acts like an ideal (active) inference machine, this is an experimental and not a computational model fact. It must be tested under a wide variety of experimental situations.

Yet what is it that should be tested in the first place? A peculiar aspect of the FEP is its use in support of *radically opposed* ontologies of the mind. Using the FEP's formal framework, different groups of theorists have come to a wide range of solutions, such as Hohwy's (2016, p. 274) neurocentric representational view, Bruineberg et al.'s (2018) embodied dynamic view, or Kirchhoff and Kiverstein's (2019) view of an activity-dependent, gear-switching fluid boundary. These ontologies identify what the respective workers deem the appropriate boundaries of a study of the mind. Yet, Bruineberg et al. argue, such ontologies are the result not of inherent features of the formalism, but instead of "additional philosophical premises." We suggest that this is not a fault. In our view, the objective ontological aspiration itself should be rejected; we propose a more thoroughly pragmatic instrumentalist view.

The relevant scale of investigation is relative to pragmatic research considerations. An example may help. Say that we want to understand an outfielder's flyball catching activity. We could investigate the outfielder in relation to the flyball and the field they are running on. Yet if we want to understand the outfielder's baseball play their catching is part of, we need to consider the larger-scale dynamics, including the relation of the outfielder to the other players, the current score, and so on. This can explode if we are instead interested in, say, the outfielder's weekly leisure routine. As such, the boundaries of the relevant system of study when studying the mind can change drastically depending on our focus.

Our view can be seen as an instrumentalist take on Kirchhoff and Kiverstein's (2019) realist view. Their view takes the boundary of the mind to swing within the spectrum ranging from environmentally extensive to skull-bound depending on the organism's activity. Yet the determination of the relevant processes for each activity rests, as Bruineberg et al. show, not upon fundamental mathematics, but, as we have described here in brief, on pragmatic considerations. As such, activity-centrism bottoms out into pragmatic research interest-dependence, and does not ground an objective ontology. We thus agree with Bruineberg et al. that the FEP in itself will not adjudicate ontological questions. Yet we argue that, under pragmatic instrumentalism, this is superfluous anyway. After all, to demand an ontology over and above what is relevant to our research interests is to demand an ontology that is epiphenomenal to our investigations. Our view thus provides a pragmatic way forward for an instrumentalist FEP.

Financial support. This work has been funded by the Postdoctoral Fellowship by Humboldt University (I. H.).

Conflict of interest. The authors (Thomas van Es, Inês Hipolito) certify that they have no affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or nonfinancial interest (such as personal or professional relationships, affiliations, knowledge, or beliefs) in the subject matter or materials discussed in this manuscript.

References

Bruineberg, J., Kiverstein, J., & Rietveld, E. (2018). The anticipating brain is not a scientist: The free-energy principle from an ecological-enactive perspective. *Synthese*, *195* (6), 2417–2444.

Dennett, D. C. (1995). Darwin's dangerous idea. *The Sciences*, *35*(3), 34–40.

Friston, K., Fortier, M., & Friedman, D. A. (2018). Of woodlice and men: A Bayesian account of cognition, life and consciousness. An interview with Karl Friston. *ALIUS Bulletin*, *2*, 17–43.

Godfrey-Smith, P. (2009). Models and fictions in science. *Philosophical Studies*, *143*(1), 101–116.

Hohwy, J. (2016). The self-evidencing brain. *Noûs*, *50*(2), 259–285.

Kirchhoff, M. D., & Kiverstein, J. (2019). *Extended consciousness and predictive processing: A third-wave view*. Routledge.

Bayesian realism and structural representation

Alex Kiefer[a] and Jakob Hohwy[b]

[a]Cognition & Philosophy Lab, Monash University, Melbourne, VIC 3800, Australia and [b]Monash Center for Consciousness & Contemplative Studies, Cognition & Philosophy Lab, Monash University, Melbourne, VIC 3800, Australia
Alex.Kiefer@monash.edu
Jakob.Hohwy@monash.edu
alexbkiefer.net
https://research.monash.edu/en/persons/jakob-hohwy

doi:10.1017/S0140525X22000231, e199

Abstract

We challenge the authors' view that Markov blankets are illicitly reified when used to describe organismic boundaries. We do this both on general methodological grounds, where we appeal to a form of structural realism derived from Bayesian cognitive science to dissolve the problem, and by rebutting specific arguments in the target article.

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

In this commentary we argue that, from the point of view of Bayesian cognitive science, concerns about the reification of Markov blankets are misplaced. We assume that scientific theories represent reality in the same way in which we have argued (Kiefer & Hohwy, 2017, 2019) that organisms' internal generative models represent their external worlds: by way of (exploitable) structural similarity. Mathematically articulated theories or models set up relations among their variables, which, when successful, mimic the relations that obtain among elements of the target system (Cummins, 1991). In asserting such theories, we impute these very structural properties to the world.

This view is fundamentally a structural-realist one (Ladyman, 1998; Ladyman & Ross, 2007), and is in a sense nothing new, but whether novel or not it is adequate to the task of addressing the present criticism. To start, this perspective on the semantics of scientific theories completely defuses the concern that projection of the content of a mathematically expressed theory (such as the free energy principle [FEP]) onto the physical world involves a category-mistake or fallacy. The "substance" of the representational medium (whether collections of abstract mathematical symbols or bits of pasta) makes no difference apart from its expressive adequacy, since it is the form and not the substance of the target domain that the theory attempts to capture in virtue of *its* form.

A traditional argument against instrumentalism (Putnam, 1975) has it that theories are predictive only to the extent that they are true. Structuralist representationalism allows us to nuance this argument: We may expect observational adequacy *to the degree* that the structure of the theory matches that of the generator of observations. A Bayesian might point out that the confidence we (ought to) place in a theory also scales with its predictive accuracy (as well as its prior plausibility). By adopting a Bayesian attitude towards scientific theories, we can then dispense with a categorical distinction between realism and instrumentalism, while taking on board the epistemic humility that makes the latter appealing.

The preceding remarks do not on their own, of course, justify a realist attitude towards Markov blankets in living systems or elsewhere – this would depend on the empirical adequacy of such Markovian descriptions. They do however suggest that there is nothing *methodologically* flawed in the practice of imputing formally characterized structures (such as structures of conditional independence in Bayesian networks) to the real world.

In the remainder we rebut three arguments given in the target article that purport to establish the contrary.

The first argument poses an analogy between Markov blankets and contour lines on cartographic maps. I expect to observe rivers and mountains when I navigate by a map, but *pace* the authors, I also expect to experience elevation and other aspects of the terrain represented by contour lines. To mistake contour lines as such for features of the terrain would indeed be a radical mistake, but so would expecting the river to be ink-blue. We do not make these mistakes in practice any more than proponents of the FEP suppose on the basis of their diagrams that the sensory epithelia of organisms consist in labelled circles with black outlines.

The moral is that contour lines contribute to the same fundamentally spatial (indeed structural-similarity-based) representation as other elements of the map, though in a slightly more abstract and conventional way. There may be more reasonable concerns about reification, for example that graphical models cut the world artificially into discrete, repeatable event-types, but this worry would impugn the use of such abstractions to represent causal structure (such as smoking's causing cancer) quite generally, and so is not properly directed against the FEP literature.

In fact, the authors *do* argue for a blanket instrumentalism with respect to Bayesian networks, which brings us to the second argument: That the choice of model for a given system depends in part on extrinsic circumstances like data availability or the interests of the scientist. But the intrusion of this pragmatic element means merely that there exist *many* sets of conditional dependencies, some more interesting than others, which does not impugn a realist attitude towards any one of those sets.

The authors further suggest that the Markov blanket formalism does little work in delineating organismic boundaries if we must already have selected a model in order to consider its blanket. It is unclear to us how this epistemic point could count against realism, but in any case the plurality of Markov blankets present in any system of interest has long been admitted by FEP theorists (cf. Friston et al., 2021; Hohwy, 2016), and in practice the FEP is interesting not as a tool for distinguishing organisms from their environments, which we can do well enough without it, but for its formal account of how systems act to *maintain* the integrity of their boundaries, however initially identified.

The final argument we consider suggests that the literature around the FEP stratifies into two distinct projects, the respectable empirical one of using Markov blankets to characterize aspects of organisms' cognitive models of their environments ("inference with a model"), and the metaphysically ambitious one of using Markov blankets to characterize organisms themselves and their boundaries with respect to their environments ("inference within a model"). Here, it is sufficient to point out that these projects are not, actually, fundamentally distinct in kind. The explanatory targets in both cases are real features of organisms (their cognitive models, in the first case, and their sensorimotor boundaries, in the second). What may be controversial in the case of the FEP is the idea that the entire organism (as opposed to some construct in its brain) may be regarded as a "model," but this is not the ground on which the authors stake their claim.

In conclusion, we have seen no reason, either on the basis of its general mathematical character or on the basis of its particular modes of application, to suspect that the Markov blanket formalism has been used by FEP theorists to commit fallacies of reification.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Conflict of interest. None.

References

Cummins, R. (1991). *Meaning and mental representation*. MIT Press.

Friston, K. J., Fagerholm, E. D., Zarghami, T. S., Parr, T., Hipólito, I., Magrou, L., & Razi, A. (2021). Parcels and particles: Markov blankets in the brain. *Network Neuroscience*, *5*(1), 211–251. https://doi.org/10.1162/netn_a_00175

Hohwy, J. (2016). The self-evidencing brain. *Noûs*, *50*(2), 259–285. https://doi.org/10.1111/nous.12062

Kiefer, A., & Hohwy, J. (2017) Content and misrepresentation in hierarchical generative models. *Synthese*, *195*(Special issue on predictive brains (M. Kirchhoff, ed.)), 2387–2415.

Kiefer, A. B., & Hohwy, J. (2019). Representation in the prediction error minimization framework. In S. Robins, J. Symons, & P. Calvo (Eds.), *The Routledge companion to the philosophy of psychology* (2nd ed., pp. 384–409). Routledge.

Ladyman, J. (1998) What is structural realism? *Studies in History and Philosophy of Science*, *29*(3), 409–424. https://doi.org/10.1016/S0039-3681(98)80129-5

Ladyman, J., & Ross, D. (2007). *Every thing must go: Metaphysics naturalised*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199276196.001.0001

Putnam, H. (1975). *Mathematics, matter and method*. Cambridge University Press.

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

Scientific realism about Friston blankets without literalism

Julian Kiverstein[a] and Michael Kirchhoff[b]

[a]Department of Psychiatry, Amsterdam University Medical Centre, 1105 AZ, Amsterdam, The Netherlands and [b]Faculty of Art, Social Sciences and Humanities, School of Liberal Arts, University of Wollongong, Wollongong, NSW 2522, Australia
j.d.kiverstein@amsterdamumc.nl
kirchhoff@uow.edu.au

doi:10.1017/S0140525X22000267, e200

Abstract

Bruineberg and colleagues' critique of Friston blankets relies on what we call the "literalist fallacy": the assumption that in order for Friston blankets to represent real boundaries, biological systems must literally possess or instantiate Markov blankets. We argue that it is important to distinguish a realist view of Friston blankets from the literalist view of Bruineberg and colleagues' critique.

In our commentary we set out to offer a defence of scientific realism about Markov blankets. Bruineberg and colleagues are right to highlight the choices that go into constructing a scientific model, such as a causal Bayesian network. However, it doesn't follow that scientific models thereby do not, or cannot, indirectly represent – even if only in an approximate fashion – target systems. Second, we argue that Bruineberg and colleagues' critique of the free energy principle (FEP) relies on a fallacy we call the "literalist fallacy." This is the fallacy of assuming that in order for Markov blankets to identify real boundaries, biological systems must literally possess or instantiate Markov blankets. Scientific realism should be distinguished from literalism, as Bruineberg and colleagues acknowledge. In sum, we conclude that while the target article asks many important questions about the use of the Markov blanket construct in the context of the FEP, it falls short of making the case against scientific realism.

Bruineberg and colleagues, in their lucid presentation of the FEP, make a helpful distinction between three meanings that attach to the use of the word "model." The first use of the term refers to the scientist's explanatory target: The neurobiological systems that are argued, by proponents of the FEP, to literally *be* models of their environments (Friston, 2013). Second, are the models the scientist makes of neurobiological systems by applying the mathematics of the FEP. Finally, following a significant amount of idealisation and abstraction on the part of the scientist, the scientist arrives at an explanatory model that purports to represent something of interest about a target system. It is at this stage in the modelling process that we arrive at the Friston blanket: A mathematical construct that purports to describe the autopoietic processes that produce a boundary separating the agent from its environment (Kirchhoff, Parr, Palacios, Friston, & Kiverstein, 2018).

Bruineberg and colleagues correctly argue that to identify a Friston blanket certain nodes in a Bayesian causal network have to be labelled as internal, external, active, and sensory states. They go on to argue (in sect. 5) that the location of the Markov blanket within a model is largely dependent on the "arbitrary" modelling choices of the scientist. The Markov blanket construct should therefore be understood as a property of the model the scientist is constructing. To think otherwise is to "reify" the Markov blanket, mistaking a construct that is the outcome of modelling statistical relationships of conditional independence, for the cause of this conditional independence.

We fully agree with Bruineberg et al., that the scientist has to make choices about where to locate the Markov blanket within a model. It does not follow, as Bruineberg and colleagues claim, that such choices are arbitrary. The scientist's decisions about how to interpret a model are a part of the process of model-building. The Markov blanket is a formal or mathematical construct. To model anything this mathematical construct has to be given an interpretation by the scientist. This interpretation does however purport to represent (i.e., describe and explain), the unobservable causes of the behaviour of real-world target systems. In the case of Markov blankets, the unobservable causes are the autopoietic processes that produce and maintain a boundary distinguishing the individual agent from its environment. The choices the scientist makes about how to interpret the Markov blanket formalism are therefore not arbitrary. They are guided by the explanatory interests of the scientist, which in this case concern the process of autopoiesis.

Bruineberg and colleagues devote a good deal of attention to uncovering hidden assumptions that are required to apply Friston blankets to biological systems. Scientific models are however very often idealised models that allow for highly complex and intractable problems to be solved – for example, placing a free energy bound on entropy. One might think that idealisation rules out models from providing accurate representations of their target systems. Idealisation introduces distortion into a model, rendering the resulting model inaccurate. Such an objection rests on a short-term view of what scientific modelling can contribute. As Weisberg (2007) notes, scientific idealisation is best understood in the context of a longer-term scientific programme to provide an accurate representation of a target system. A scientific model can represent a target system partially, approximately or probably. The descriptions of the system the scientist provides need not be literally true of the system to approximately describe the behaviour of a target system.

Consider again, with these helpful reminders from the philosophy of science in place, the philosophical mistake Bruineberg and colleagues claim to have uncovered in the very idea of Friston blankets. We have argued that Friston blankets are interpretations given of the Markov blanket formalism in the context of the FEP that purport to describe autopoietic processes. Bruineberg and colleagues claim that to take Friston blankets to represent the processes that cause the conditional independencies differentiating the agent from its environment, one must take the biological agent to literally instantiate or possess a Markov blanket. However, such a claim relies on a fallacious assumption: that scientific realism implies literalism. The realist claims that the Markov blanket formalism can, as an idealised interpretation of a model, nevertheless purport to represent an unobservable causal property. The literalist claims that for a model to represent an unobservable cause, the model must literally be true of a system. Bruineberg and colleagues' argument against Friston blankets trades on a confusion of realism and literalism we call the "literalist fallacy."

The target article offers other arguments against Friston blankets. They suggest for instance that for a Friston blanket to mark

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

the boundary of a biological system one must already know where to place the boundary. However, this neglects the application of Friston blankets to extended cognitive systems whose boundaries are negotiable (Clark, 2017; Kirchhoff & Kiverstein, 2021). The question of where to place the boundary of the system is what is at stake in applying the Markov blanket formalism to such systems (Kirchhoff & Kiverstein, 2021). Bruineberg and colleagues have not shown that Friston blankets cannot help to settle such a question.

Financial support. Kiverstein is supported by an Amsterdam Brain and Cognition Project Grant; The European Research Council (Grant number 679190) and the Netherlands Scientific Organisation (NWO, Vici, awarded to Erik Rietveld). Kirchhoff is supported by Australian Research Council Discovery Project Minds in Skilled Performance (Grant number DP170102987).

Conflict of interest. None.

References

Clark, A. (2017). How to knit your own Markov blanket: Resisting the second law with metamorphic minds. In T. Metzinger & W. Wiese (Eds.), *Philosophy and predictive processing: 3* (pp. 1–19). MIND Group. https://doi.org/10.15502/9783958573031

Friston, K. (2013). Life as we know it. *Journal of the Royal Society Interface*, *10*(86), 20130475.

Kirchhoff, M. D., & Kiverstein, J. (2021). How to determine the boundaries of the mind: A Markov blanket proposal. *Synthese*, *198*, 4791–4810. https://doi.org/10.1007/s11229-019-02370-y

Kirchhoff, M. D., Parr, T., Palacios, E., Friston, K. J., & 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.

Weisberg, M. (2007). Three kinds of idealization. *The Journal of Philosophy*, *104*(102), 639–659.

Markov blankets do not demarcate the boundaries of the mind

Richard Menary and Alexander J. Gillett

Macquarie University, Sydney, NSW 2109, Australia
richard.menary@mq.edu.au
alexander.gillett@mq.edu.au

doi:10.1017/S0140525X22000371, e201

Abstract

We agree with Bruineberg and colleagues' main claims. However, we urge for a more forceful critique by focusing on the extended mind debate. We argue that even once the Pearl and Friston versions of the Markov blanket have been untangled, that neither is sufficient for tackling and resolving the question of demarcating the boundaries of the mind.

When demarcating the boundaries of the mind the demarcation conditions should meet the following criteria:

(1) naturalistically motivated,
(2) non-question begging, and
(3) carve nature at the joints.

Markov blankets have been proposed as providing the conditions for the demarcation of the boundary of mind and world (Clark, 2016; Hohwy, 2016; Kirchhoff & Kiverstein, 2021; Kirchhoff, Parr, Palacios, Friston, & Kiverstein, 2018; Ramstead, Kirchhoff, Constant, & Friston, 2021). We argue that Bruineberg and colleagues' distinction between Friston and Pearl blankets exposes why Markov blankets cannot meet the criteria required to demarcate the boundaries of mind.

Pearl blankets are naturalistically motivated and non-question begging but do not carve nature at the joints. This is because Pearl blankets are based upon practices in a range of scientific fields which do not have the purpose of carving nature at its joints, but are methodologically salient.

Friston blankets are not naturalistically motivated. There is an "explanatory leap," as Bruineberg and colleagues put it, whereby the scientific practices and mathematical rigour of Pearl blankets are erroneously transposed onto Friston blankets by free energy principle (FEP) proponents despite them being disanalogous. Principled reasons and evidence need to be presented to go beyond Pearl blankets to metaphysical claims about reality. For example, in a survey article by Mann, Pain, and Kirchhoff (forthcoming), they effortlessly move between Pearl's restricted sense of a Markov blanket as the total information provided about a node in a Bayesian network by all of the nodes to which it is locally connected, to the "special usage" of Markov blanket as a Friston blanket: A set of nodes that screen off an agent from a set of nodes that are external to the agent. They do so without any explanation of how we get from Pearl's conception to Friston's, this is a very clear example of the explanatory leap that Bruineberg and colleagues have identified.

Furthermore, it is debatable whether the Friston blanket is carving nature at the joints, and it is question begging insofar that it modifies the debate rather than tackles it – transforming the debate from one about the extended mind (Clark & Chalmers, 1998) to one about the applicability of mathematics (Bangu, 2012). Friston blankets carry radical metaphysical baggage that is, at present, unjustified and unacknowledged by FEP proponents. For instance, Ramstead et al. (2021) posit a formal ontology in which "traditional" metaphysical enquiry is replaced by mathematics. Rather than eliminating metaphysics, this *is a metaphysical view* – with a long lineage in Western philosophy. As we have argued previously (Menary & Gillett, 2021), the inherent Platonism and Pythagoreanism here is controversial. There are a number of issues that must be tackled in order to make such a view legitimate. For example, if one contends that nature is mathematical, then there are a range of concerns about the relationship between mathematical and physical structures, and whether one is claiming that mathematical entities have causal powers. If Friston blankets are to function as demarcating the actual boundaries of the mind, then the inference is that they have causal powers. But, as mentioned above, FEP proponents often move between methodological and ontological usages of the concept. One cannot simply forgo philosophical argumentation by "doing the math," as nicely noted by Bruineberg and colleagues. Instead, serious principled reasons need to be given as to why this metaphysical picture is justified and warranted. To be clear, our claim is not that a Friston blanket style-approach to demarcating the boundaries of the mind is untenable, rather it is currently undefended.

Pearl blankets are insufficient

One option is to rely on Pearl blankets as a heuristic for demarcation. Some FEP proponents (e.g., Ramstead, Friston, and

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

Hipólito, 2020a; Ramstead, Kirchhoff, & Friston, 2020b) have dabbled with instrumentalism. However, even a deflationary position does not help. Pearl blankets by themselves are insufficient to adjudicate between internalist (e.g., Hohwy, 2013) and extended (e.g., Clark, 2016) accounts of the boundaries of cognition that draw upon the Markov blanket conception. The issue is that a Pearl blanket – as the evidentiary boundary beyond which are hidden variables – can be articulated as a shifting boundary based on an action-orientated engagement with the world (a la Clark) or as a fixed boundary of the skull-bound brain (a la Hohwy). Not only is a Pearl blanket consistent with these opposing positions but it is also insufficient by itself to differentiate between them.

Friston blankets are insufficient

To clearly demarcate the boundaries of the mind one must determine the ontological conditions under which such boundaries are drawn, and this is not a matter of heuristics or instrumental thinking; it is a matter of carving nature at its joints. We note that this scientific realist construal of Friston blankets allows that the scientific models can be approximately true, partial, and probabilistic. However, they nevertheless make concrete ontological claims about the existence of causal phenomena in the world – that is, Friston blankets are real entities. As such, we are forced to return to the challenges discussed above and Friston blankets fail the first and second criteria. We would like those utilising Friston blankets to demarcate the boundaries of mind to be more careful in their shift between Pearl blankets and Friston blankets and to begin to give an account of how these entities function in nature in terms of their causal powers.

In summary, the conundrum for FEP proponents is that both conceptions fail to fully meet the criteria for demarcating the boundaries of the mind. If we adopt a Pearl blanket conception, then we are incapable of differentiating between internalist and extended positions because a methodological heuristic cannot decide an ontological matter. Therefore, we must turn to the more ontologically robust Friston blanket conception. This move involves an explanatory leap that is not currently justified. This either retreats the position back to the unworkable but justified Pearl blanket conception, or requires a commitment to the causal powers of Friston blankets – but this only replaces one set of problems with another (from the extended mind to the applicability of mathematics) rather than actually helping to tackle the original problem.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Conflict of interest. None.

References

Bangu, S. (2012). *The applicability of mathematics in science: Indispensability and ontology*. Palgrave Macmillan.

Clark, A. (2016). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, *58*(1), 7–19.

Hohwy, J. (2013). *The predictive mind*. Oxford University Press.

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

Kirchhoff, M., & Kiverstein, J. (2021) How to determine the boundaries of the mind: A Markov blanket proposal. *Synthese*, *198*, 4791–4810.

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

Mann, S. F., Pain, R., & Kirchhoff, M. (forthcoming). Free energy: A user's guide.

Menary, R., & Gillett, A. J. (2021). Are Markov blankets real and does it matter? In D. Mendonça, M. Curado & S. S. Gouveia (Eds.), *The philosophy and science of predictive processing* (pp. 39–58). Bloomsbury Academic.

Ramstead, M., Kirchhoff, M. D., Constant, A., & Friston, K. (2021). Multiscale integration: Beyond internalism and externalism. *Synthese*, *198*(Suppl 1), S41–S70.

Ramstead, M. J., Friston, K. J., & Hipólito, I. (2020a). Is the free energy principle a formal theory of semantics? From variational density dynamics to neural and phenotypic representations. *Entropy*, *22*(8), 889.

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

Boundaries and borders gone! But life goes on

Kathryn Nave

Department of Philosophy, School of Philosophy, Psychology and Language Sciences, The University of Edinburgh, Edinburgh EH8 9AD, UK
Kathryn.nave1@gmail.com

doi:10.1017/S0140525X22000152, e202

Abstract

Unlike machines, living systems are distinguished by the continual destruction and regeneration of their boundaries and other components. Stable Markov blankets may be a real feature of the world, or they may be merely a construction of particular models, but they are neither a feature of organisms nor of any model that can capture the necessary conditions of their existence.

Suppose we took the view that fundamental reality is one great big Bayes net, decomposable into distinct units whose state is governed by local interactions that respect the causal Markov condition. Fristonite metaphysicians would not be the first to advance such a reduction – see Weslake (2006) for a critical review. Then, even if our partial models and their particular blankets are pragmatic constructions, there would still be a fact of the matter about where the *real* Markov blankets lie.

Unfortunately, as the authors note, we would end up with far more overlapping "real Markov blankets" than we'd know what to do with. This seems problematic if they are to be treated as distinctive of living systems, and if we wish, as Kirchhoff, Parr, Palacios, Friston, and Kiverstein (2019) apparently do, to "think about any system that possesses a Markov blanket as some rudimentary (or possibly sophisticated) 'agent'" (p. 2).

There is, however, a more significant problem than the promiscuous vitalism this entails. This is the fact, oddly neglected in the free-energy literature, that living systems are specifically distinguished by their rare ability to persist through, and constitutive dependence upon, the continual destruction and regeneration of their boundaries.

In the free-energy literature, the cellular membrane is taken as the canonical example of a biological agent's Markov blanket. Friston (2013, 2019) repeatedly contrasts this to the candle flame that cannot possess a Markov blanket "because its constituent particles are in constant flux" (2019, p. 50).

Yet the membrane's stability is illusory, its constituent parts being relentlessly exchanged through endo- and exocytosis for

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

regeneration, growth, and particle transport. In the slime mould *Dictyostelium*, membrane turnover enables locomotion, with estimated times for complete turnover in the order of 4–10 min (Aguado-Velasco & Bretscher, 1999). The same goes for the interior of the cell, where we find turnover times for its enzymes, that are far shorter than the lifespan of the cell itself (Toyama & Hetzer, 2013).

A "literalist" who treats the fundamental relata of causality as the states of particular token particles will find that the components of any "real" Markov blanket they identify around an organism will dissipate on timescales shorter than that lifespan of the organism whose "very existence," they claim, "depends" on that boundary's preservation (Allen & Friston, 2018).

We don't have to take the states of specific particles as the nodes of reality's network. One's metaphysics of causality could be statistical reductionism, but where the relata are higher-level macrophysical variables (Papineau, 1992). In an organism, as in a machine, we could suggest that what must be fixed are not particular material components but the formal parts making up its functional organization. While it would seem less plausible to regard a set of formal parts as constituting a physical boundary *in the world*, this would be compatible with the realist position that the Markov blanket describes something objective *about the world*.

Still, the realist needs to explain how we identify this organization amid the constant turnover of the stuff that realizes it, in order to then evaluate any statistical or causal relationships that might hold between its parts. Once we have done so, it's not clear what further the Markov blanket formalism offers beyond establishing a beachhead for the deployment of the free energy principle.

The instrumentalist is in a slightly easier position, needing only to offer a pragmatic justification for dividing the world up in a particular way, and being free to admit that the represented stability of some network (and resulting Markov blanket) is a modelling distortion that abstracts away from material turnover, in order to focus on other features of the organism's dynamics.

This is fine if the purpose of our model is only to describe a specific behaviour, such as the regulation of body temperature. But if our model is supposed to provide the basis for a general theory of life, as Friston (2013, 2019) presents the free energy framework, then to acknowledge that it, like all models, is partial and distorted is not sufficient. The task of a model of "life in general" is to highlight the *right* things, and neglect only those contingent features of the particular instances we happened to have encountered.

To abstract away from metabolic turnover is not merely to neglect some capacities common to many organisms, it is to fundamentally misconceive what an organism is, and what differentiates it from a mechanism. Unlike in machines the structures that constrain an organism's dynamics, its membrane, its enzymes, and so on are inherently unstable and recursively dependent upon those dynamics for continued repair and reproduction (Bickhard, 2009; Montévil & Mossio, 2015). The flow of matter is not just channelled *through* fixed constraints like fuel in an engine, it constitutes those constraints. In organisms, as Nicholson (2018) puts it, "everything flows."

A statistical network and its attendant Markov blanket describe how a system's structure constrains its dynamics, they do not address any reciprocal dependence of this structure upon those dynamics. Once Huygen's coupled pendulums wind down, the connecting beam remains as a constraint on possible interactions should they be perturbed again.

Organisms are not just homoeostatic mechanisms, "acting" only in response to perturbation. They are intrinsically unstable structures – stabilized only via their own ceaseless activity and dependent upon the environment as a resource for such self-production. As Jonas (1953) criticized of the free-energy framework's cybernetic precursor, "A feedback mechanism may be going, or may be at rest: in either state the machine exists. The organism has to keep going, because to be going is its very existence" (p. 191).

Markov blankets may be useful for modelling coupled feedback mechanisms. Such mechanisms may even *literally* have Markov blankets. But for a theory of living systems, the principal issue is not whether Markov blankets are features of reality, or just of our models. It's that cells are much more like candle flames than they are like pendulums. While it may sometimes be convenient to treat an organism like a machine, this fiction obscures why the cell is alive, and the pendulum is not.

Financial support. This study was supported by the Royal Institute of Philosophy, The Aristotelean Society and ERC advanced grant X-SPECT – DLV-692739.

Conflict of interest. None.

References

Aguado-Velasco, C., & Bretscher, M. S. (1999). Circulation of the plasma membrane in *Dictyostelium*. *Molecular Biology of the Cell*, *10*(12), 4419–4427.
Allen, M., & Friston, K. J. (2018). From cognitivism to autopoiesis: Towards a computational framework for the embodied mind. *Synthese*, *195*(6), 2459–2482.
Bickhard, M. H. (2009). The interactivist model. *Synthese*, *166*(3), 547–591.
Friston, K. (2013). Life as we know it. *Journal of the Royal Society Interface*, *10*(86), 20130475.
Friston, K. (2019). Beyond the desert landscape. In M. Colombo, E. Irvine, & M. Stapleton (Eds.), *Andy Clark and his critics* (pp. 174–190). Oxford University Press.
Jonas, H. (1953). A critique of cybernetics. *Social Research*, *20*, 172–192.
Kirchhoff, M., Parr, T., Palacios, E., Friston, K., & Kiverstein, J. (2018). The Markov blankets of life: Autonomy, active inference and the free energy principle. *Journal of the Royal Society Interface*, *15*(138), 20170792.
Montévil, M., & Mossio, M. (2015). Biological organisation as closure of constraints. *Journal of Theoretical Biology*, *372*, 179–191.
Nicholson, D. J. (2018). Reconceptualizing the organism: From complex machine to flowing stream. In D. J. Nicholson & J. Dupré (Eds.), *Everything flows: Towards a processual philosophy of biology* (pp. 139–166). Oxford University Press.
Papineau, D. (1992). Can we reduce causal direction to probabilities? In A. Woody (Ed.), *PSA: Proceedings of the biennial meeting of the philosophy of science association* (Vol. *1992*, No. 2, pp. 238–252). Philosophy of Science Association.
Toyama, B. H., & Hetzer, M. W. (2013). Protein homeostasis: Live long, won't prosper. *Nature Reviews Molecular Cell Biology*, *14*(1), 55–61.
Weslake, B. (2006). Common causes and the direction of causation. *Minds and Machines*, *16*(3), 239–257.

What's special about space?

Thomas Parr

Wellcome Centre for Human Neuroimaging, Queen Square Institute of Neurology, University College London, London WC1N 3AR, UK
thomas.parr.12@ucl.ac.uk
https://tejparr.github.io/

doi:10.1017/S0140525X2200019X, e203

Abstract

This commentary suggests that, although Markov blankets may have different interpretations in different systems, these distinctions rest not upon the type of blanket, but upon the model that

determines the blanket. As an example, the conditions for a model in which the Markov blanket may be interpretable as a physical (spatial) boundary are considered.

I enjoyed reading the target article by Bruineberg et al. and fully endorse the authors' agenda of ambiguity resolution. In this spirit, it is worth beginning this article by carefully considering what we mean by the words we choose, and specifically by the word "model." The definition of a model is a subject that deserves much more space than is available for this commentary. However, the use of the term by the authors of the target article suggests a very simple definition of the sort of model we are interested in. For the purposes of this commentary, a model is another word for a joint probability distribution. The graphical models of the target article can be derived simply from this starting point (Wainwright & Jordan, 2008).

Under this definition of a model, it follows that a Markov blanket, itself defined in terms of conditional probabilities (Pearl, 1988), must always be defined in relation to a model. This raises the question of where models come from. One source of the probability distributions that make up a model is the potential function, Hamiltonian, or steady-state density of some dynamical system (Friston & Ao, 2012). This represents a model of a specific sort of (time-evolving) system. Bruineberg et al. suggest that this means blankets in such systems are different in nature to those in other kinds of systems. However, the identification of a Markov blanket in such a model is no different to its identification in any other model.

As such, the key distinction is not between different kinds of Markov blankets. A taxonomy of models, from which conditional independencies can be identified, may be a more meaningful way of addressing the implicit distinctions. Bruineberg et al. make some steps towards this, subcategorising models in several different ways. For instance, they highlight the distinction between our model of some system versus some (sentient) system's model of its world (Friston, Wiese, & Hobson, 2020). While many of these were interesting, the distinction that I found most intriguing was that between a model in which the Markov blanket plays the role of a "physical" boundary and one in which it offers only a "statistical" boundary.

So, what is it that makes a boundary physical? The remainder of this commentary explores this question, assuming that physical is here synonymous with spatial, with the aid of an example system depicted in Figure 1. Here, we have three beads on a string, separated by springs. If two beads become too close, the compressed spring pushes them away from one another. If too far apart, the extended spring pulls them back together.

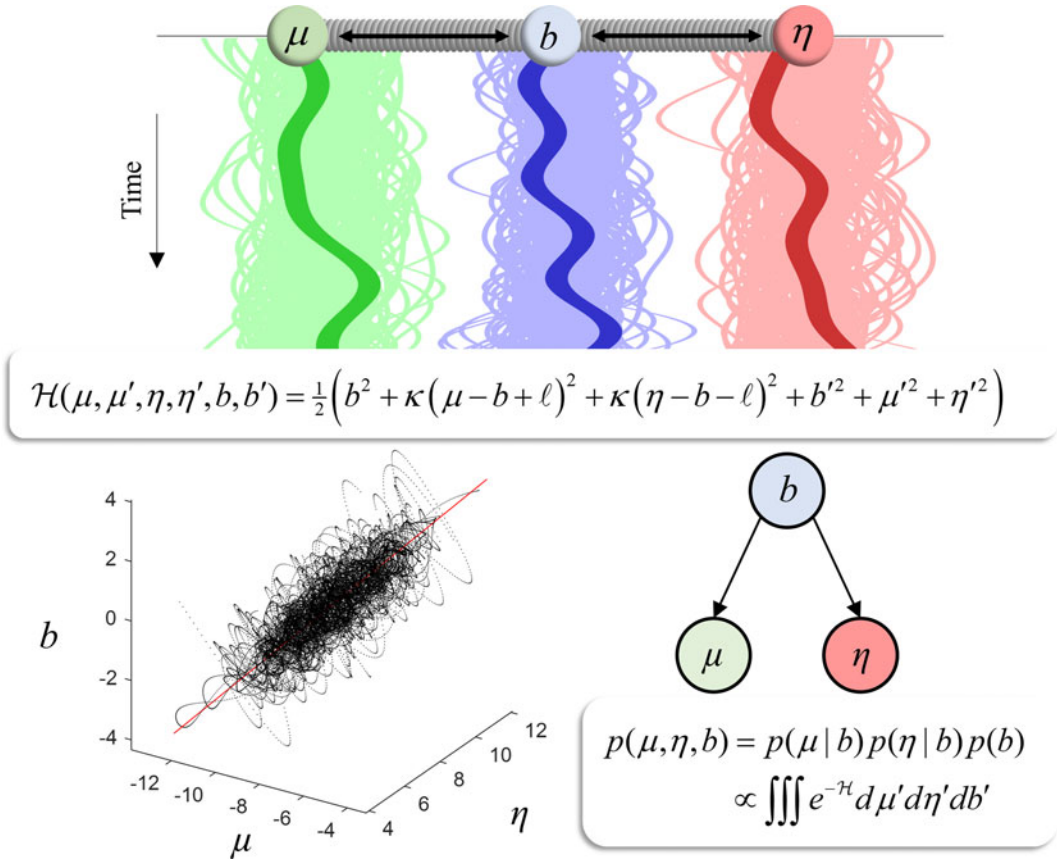


Figure 1 (Parr). Spatial boundaries. This figure depicts an example system that can be interpreted as having a "physical" boundary. It comprises three beads (with positions μ, η, and b, and velocities indicated by prime notation) separated by two springs with equilibrium lengths ℓ. The Hamiltonian ($\mathcal{H}$) incorporates the potential energies associated with the springs, and the kinetic energies of each bead. The decomposition of the Hamiltonian into a sum of terms ensures the associated steady-state density is consistent with the conditional independencies required for a Markov blanket. Interpreting this steady state as a model, we can express it as a Bayesian network (lower right). The trajectories of each bead are shown at the top of the figure, showing many trajectories with different initial conditions (sampled from the steady-state density). An example trajectory is superimposed upon each. The lower left plot shows all simulated trajectories plotted in three dimensions, with the direction of the first principal component shown in red.

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

In this example, the positions of each bead collectively constitute a three-dimensional system (as plotted in the lower left). What is it that licenses us to embed each of these along a single spatial dimension, such that it is meaningful to describe the positions of objects relative to one another? Only by doing so can we talk of spatial boundaries for which it would be surprising to observe one element of the system cross from one side of this boundary to the other.

The steady-state density implicit in this system's Hamiltonian factorises to reveal a Markov blanket. Specifically, the positions of the left and right beads are conditionally independent of one another given the central bead. This means the middle bead's position is a Markov blanket for the positions of the other two beads. Another way of putting this is that all covariance shared between the left and right beads is dependent upon the middle bead. The lower left plot of Figure 1 illustrates this heuristically by plotting multiple trajectories with different initialisations to give a sense of the shape of the joint probability density. Note that the axis that accounts for most of the variance (shown in red) is a linear combination of the original three axes.

The use of prepositions (left, right, and middle) in describing the beads is crucial in interpreting the Markov blanket implied by this model as a spatial boundary. Clearly it would be meaningless to plot the three positions on the same axis (as in the upper plot of Fig. 1) if their positions relative to one another played no role in their behaviour. The model of the beads ensures pairs of adjacent beads constrain one another's positions such that (for example) it would be very surprising to find two beads in the same spatial location.

In short, a model that lends itself towards an interpretation of its Markov blankets as spatial boundaries must exhibit the following features:

(1) The probability density associated with the model must have a non-spherical covariance structure. This is guaranteed by the presence of a non-trivial Markov blanket.
(2) The pair of conditional densities describing the positions of the variables partitioned by the blanket, given the blanket, must assign very low probability to positions immediately proximate to the position of the blanket. This is achieved in our example via the repulsive forces when the springs were compressed.
(3) The model must be interpretable as a steady-state density. This implies a system that evolves in time but whose density dynamics, when initialised at the steady state, are static. The time evolution is important in that it provides an explicit link between the model and the Lagrangians and Hamiltonians found in physics.

Presumably, one of the reasons Bruineberg et al. highlighted the special case in which Markov blankets take on the flavour of spatial boundaries is that they are sometimes called upon (Ramstead, Badcock, & Friston, 2018) in an attempt to address Schrödinger's famous question about the physics of life (Schrödinger, 1944), formulated explicitly in terms of a spatial boundary. This commentary was written to question whether such boundaries require special kinds of Markov blankets and suggests that, instead, they require special kinds of model.

Financial support. The Wellcome Centre for Human Neuroimaging is supported by core funding (203147/Z/16/Z).

Conflict of interest. None.

References

Friston, K. J., & Ao, P. (2012). Free energy, value, and attractors. *Computational and Mathematical Methods in Medicine*, *2012*, 937860–937860. doi: 10.1155/2012/937860

Friston, K. J., Wiese, W., & Hobson, J. A. (2020). Sentience and the origins of consciousness: From Cartesian duality to Markovian monism. *Entropy*, *22*(5), 516. doi: 10.3390/e22050516

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

Ramstead, M. J. D., Badcock, P. B., & Friston, K. J. (2018). Answering Schrödinger's question: A free-energy formulation. *Physics of Life Reviews*, *24*, 1–16. doi: 10.1016/j.plrev.2017.09.001

Schrödinger, E. (1944). *What Is Life?: The Physical Aspect of the Living Cell*. Retrieved from Dublin.

Wainwright, M. J., & Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, *1*(1–2), 1–305. doi: 10.1561/2200000001

The emperor has no blanket!

Vicente Raja[a], Edward Baggs[b], Anthony Chemero[c] and Michael L. Anderson[d]

[a]Rotman Institute of Philosophy, Western University, London, Ontario N6A 5B7, Canada; [b]Department of Language and Communication, Danish Institute for Advanced Study, University of Southern Denmark, 5230 Odense, Denmark; [c]Departments of Philosophy and Psychology, University of Cincinnati, Cincinnati, OH 45221, USA and [d]Department of Philosophy, Brain and Mind Institute, Rotman Institute of Philosophy, Western University, London, Ontario N6A 5B7, Canada
vgalian@uwo.ca
ebag@sdu.dk
chemeray@ucmail.uc.edu
mande54@uwo.ca
http://www.emrglab.org
https://danish-ias.dk/people/edward-baggs-danish-ias/
https://uc.academia.edu/TonyChemero
http://www.emrglab.org

doi:10.1017/S0140525X22000243, e204

Abstract

While we applaud Bruineberg et al.'s analysis of the differences between Markov blankets and Friston blankets, we think it is not carried out to its ultimate consequences. There are reasons to think that, once Friston blankets are accepted as a theoretical construct, they do not do the work proponents of free energy principle (FEP) attribute to them. The emperor is indeed naked.

The title of the target article gestures toward H. C. Andersen's *The Emperor's New Clothes* short story. The core of the story is a child saying what all the neighbors already knew but were afraid to say: that the emperor had no clothes! After this suggestive title, we expected a paper that not only would analyze the differences between Markov and Friston blankets but that, following the child's lead, would tell us whether the theoretical construct of "Friston blankets" is doing the job many free energy principle (FEP) proponents attribute to it. Namely, it would tell us whether Friston blankets are indeed a principled way to find the ontological boundaries between biological/cognitive systems and environments.

Against our expectations, the target article concludes in an ecumenical way taking Markov and Friston blankets as two different

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

constructs for two different projects. In the case of Friston blankets the project is an ontological one that is focused on finding boundaries in a principled way – for example, organismal boundaries in the case of the cell membrane or cognitive boundaries in the case of sensory receptors and motor effectors. We think, however, this ecumenical solution is supported neither by the arguments in the target article nor in the current literature on FEP. Indeed, all things considered, a more radical consequence is preferable: Friston blankets do not provide a principled way to find ontological boundaries between biological/cognitive systems and their environments.

The reasons to endorse such a radical consequence are both technical and theoretical. Technically speaking, recent work has consistently shown problems with both the formalism of Friston blankets (Biehl, Pollock, & Kanai, 2021) and its scope (Aguilera, Millidge, Tschantz, & Buckley, 2021). We think these problems are substantive and point to a mismatch between the promises of FEP as a framework and its theoretical and mathematical development to date. The typical response to these technical problems is that the FEP formalism is a work in progress, so these problems will eventually be solved. This is perfectly fine, but a work in progress is not sufficient to support the grandiose claims FEP proponents make about its current relevance for theoretical biology, cognitive science, or even physics (e.g., Friston, 2019; see also the Introduction of the target article). However interesting and important these technical issues are, we think the problems with Friston blankets go beyond them. Even granting that the formalism is right and fairly complete, Friston blankets do not do the principled, ontological work they are claimed to do. We discussed this in depth in Raja, Valluri, Baggs, Chemero, and Anderson (2021). One of our arguments parallels the target article's argument of ambiguous boundaries exemplified by various Friston blanket models of the knee-jerk reflex, so we will not repeat it here.

Another argument has to do with the fact that the states of the system partitioned by Friston blankets must be decided before finding the blanket. Consider an organism that, moving around in its environment, encounters a gap between obstacles. The organism must know whether the gap is big enough to fit its body through. In this situation, the environmental states might be of at least two kinds: (1) the position of each of the obstacles or (2) the relative position of the obstacles (i.e., the gap). Depending on describing this environment in terms of (1) or (2), the internal states of the organism will be inferring relationships from non-relative states or detecting relative states, respectively. Friston blankets do not help with this decision. A different set of resources to decide ontological questions such as *what* environmental states are and *how* these states relate to the internal ones is needed. These resources are the ones making the ontological heavy lifting, so the principled nature of the Friston blankets seems to be challenged. More generally, this argument relates to the inability of Friston blankets to deal with relational properties. This is a big problem for FEP's ambition to provide a theory of cognition because relational properties are ubiquitous within organism–environment systems. For instance, if affordances are organism–environment relationships, they seem to cut across any partition of the systemic states with Friston blankets.

A further issue, pointed out by Di Paolo, Thompson, and Beer (2021) and Raja et al. (2021), is that Friston blankets are unable to account for autopoietic self-organization. The paradigmatic example of autopoietic self-organization is the cell. The membrane of a cell is the *product* of the internal mechanisms of the cell itself. FEP proponents have suggested that the cell membrane can be understood as a Friston blanket. However, while Friston blankets are used to model the input–output relations of the cell through its membrane boundary, they say nothing about how the blanket itself comes to existence as the product of cellular activities. This is what autopoietic self-organization seeks to explain but within the Friston blanket framework it is merely presupposed. Additionally, in the case of cognitive systems, Friston blankets always appear in the context of inferential frameworks (therefore active inference), and inferential frameworks have their own issues (see Raja, 2020) that are neither dependent on nor solved by the use of Friston blankets.

In summary, Friston blankets need many other assumptions, and these other assumptions are the ones doing the ontological work (e.g., deciding what the states are, what the system does, how the system self-organizes, etc.). Friston blankets cannot be the arbiters of ontological debates. They might be just tools for modeling a previously decided ontology, but that claim requires further work that is not found in the FEP literature so far. In this context, we are in general agreement with the target article that Friston blankets are not just Markov blankets. However, we think the authors do not fully embrace the consequences of their own conceptual move. Friston blankets are not good resources for finding ontological boundaries. Maybe it is more sensible to follow William James in understanding the boundary-line of the mental – and, we add, of life – as something paradigmatically vague and, therefore, to be more pluralistic in our attempts to model it.

Financial support. This research was supported in part by a Canada Research Chairs Program award to MA, grant number SSHRC 950-231929. AC was supported by the Charles Phelps Taft Research Center at University of Cincinnati (USA).

Conflict of interest. None.

References

Aguilera, M., Millidge, B., Tschantz, A., & Buckley, C. L. (2021). How particular is the physics of the free energy principle? *Physics of Life Reviews*, *40*, 24–50. https://doi.org/10.1016/j.plrev.2021.11.001

Biehl, M., Pollock, F. A., & Kanai, R. (2021). A technical critique of some parts of the free energy principle. *Entropy*, *23*(3), 293.

Di Paolo, E., Thompson, E., & Beer, R. D. (2021). Laying down a forking path: Incompatibilities between enaction and the free energy principle. https://doi.org/10.31234/osf.io/d9v8f

Friston, K. J. (2019). *A free energy principle for a particular physics*. [preprint] arXiv:1906.10184.

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

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. *Physics of Life Reviews*, *39*, 49–72. https://doi.org/10.1016/j.plrev.2021.09.001

The empire strikes back: Some responses to Bruineberg and colleagues

Maxwell J. D. Ramstead[a,b]

[a]Wellcome Centre for Human Neuroimaging, University College London, London WC1N 3AR, UK and [b]VERSES Research Lab and Spatial Web Foundation, Los Angeles, CA 90016, USA
maxwell.d.ramstead@gmail.com

doi:10.1017/S0140525X22000139, e205

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

Abstract

In their target paper, Bruineberg and colleagues provide us with a timely opportunity to discuss the formal constructs and philosophical implications of the free energy principle. I critically discuss their proposed distinction between Pearl and Friston blankets. I then critically assess the distinction between inference with a model and inference within a model in light of instrumentalist approaches to science.

Bruineberg and colleagues provide us with a timely opportunity to discuss the core mathematical and philosophical aspects of the variational free energy principle (FEP) and the Bayesian mechanics that follows from it. They focus on the construct of Markov blankets (MBs), which they claim has been deployed in two different – but largely conflated – ways in the literature. In their view, this conflation has led some to unduly project the epistemic virtues of one use of the construct – to formalize conditional independence in the context of inference in Bayesian networks, what they call "Pearl blankets" (PBs) – to another use; namely, to demarcate the boundaries of things that exist – what they call "Friston blankets" (FBs). Furthermore, the authors argue that these constructs are employed to pursue quite distinct research projects. These are "inference with a model," where the features of the model (here, the MBs) are assumed to be part of the scientist's model of the world, and "inference within a model," where these features are assumed to be present in the modelled system itself (i.e., the MB is cast as an actually existing boundary between the system modelled and its embedding environment).

Here, I respond critically to these claims. First, I argue that PBs and FBs are nothing more than articulations of MBs in different mathematical contexts (i.e., in static statistical inference vs. stochastic dynamics). Second, I claim that an instrumentalist reading of the FEP is available and informative, and that it is not given enough consideration in the account by Bruineberg and colleagues.

Friston and Pearl blankets are just Markov blankets

The target paper claims that FBs are a novel mathematical construct that do not inherit their epistemic virtues from the use of MBs in statistical inference (i.e., from the use of PBs). Mathematically speaking, however, there is no justification for this hardline distinction.

This is because FBs and PBs *just are* MBs – in different mathematical contexts. Generally speaking, as the authors note, MBs formalize conditional (in)dependence between variables, where the MB itself is a set of variables that renders two other sets of ("internal" and "external") variables conditionally independent of each other. The fundamental distinction is that PBs are the kind of MB that arises in the context of static statistical inference within a Bayes network, while FBs arise when considering the interdependencies between dynamics, which is crucial, for instance, when defining the self in relation to the non-self. However, both PBs and FBs are MBs.

A same mathematical object may have different mathematical properties in different mathematical contexts. Consider an analogy. Let us list the numbers that are generated by the Peano axioms. We start by defining a first number, and we call it "0." Next, we define a successor function S, such that for ever number n, $S(n) = n + 1$. Now, when considered as objects of the *category of natural numbers*, these Peano numbers have specific properties. For instance, there exist no numbers between any numbers and its successor, and there exists no number smaller than zero. Next, consider the same sequence, now interpreted in the *category of real numbers*. Although the objects have not changed (since, after all, we are considering the same sequence), their properties are remarkably different. For instance, there are no longer zero, but now infinitely many numbers between any number in the sequence and its successor, and there are infinitely many numbers smaller than zero.

This example illustrates that mathematical context matters in determining the properties of mathematical objects. Just as the number 1 is the same object in both categorical interpretations (albeit with different properties when considered as a natural vs. as a real number), so, too, are FBs and PBs – just MBs, albeit defined in different mathematical contexts. In category-theoretic terms, we have changed the category, so the object instance has new properties, but it is still the same object.

The Bayesian mechanics is physics, not metaphysics

I believe that the core issue with the target paper is that it treats the FEP as if it were a metaphysical statement, when in reality, it is better understood as a new chapter of physics – a Bayesian mechanics – and is compatible with instrumentalism about scientific theories (Friston, Heins, Ueltzhoffer, Da Costa, & Parr, 2021). Although the authors discuss this possibility in passing, they downplay its significance.

There is a longstanding tradition in the philosophy of physics that is *instrumentalist* about physical theories (Giere, 1999, 2010; Van Fraassen, 1980). On the instrumentalist view, all scientific models, such as the ones used in physics, are literally false and play the role of useful fictions that help us to understand the world.

Now, there is nothing about the FEP that commits us to realism about scientific models. In fact, we have argued that precisely the opposite is the case; see Ramstead, Friston, and Hipólito (2020).

In section 3.2 of their paper, the authors provide a great outline of the modelling strategy at play in the computational neuroscience literature: what they call "models of models." We agree with this way of putting things but would suggest to extend the logic at play to FBs. They point out that in computational neuroscience, there are two "levels" of modelling at play, which are explored in Ramstead et al. (2020). The first level is that of the scientist that is modelling some target phenomenon, for example, constructing scientific models of living systems. The second level is that of the system being modelled. In computational neuroscience, scientists are constructing *scientific models of the inferential models and processes* that are assumed to be used by cognitive systems themselves. Crucially, this "second level" of modelling is just a special case of the first level – it just happens to be that the physical system being modelled, is *modelled as if* it were inferring the causes of its sensory states. As Alex Kiefer put it (personal communication), according to the FEP, the best scientific model of the organism is a statistical model of its world.

From an instrumentalist perspective, there is no robust philosophical difference between inference with a model and inference within a model. There is no reason why we cannot use MBs as a modelling tool, to carve out the boundaries of systems, when these are modelled as random dynamical systems (i.e., as sets of random

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

variables over paths – i.e., stochastic processes – with dependence relations).

The authors appeal to the notion of a *formal ontology* that was introduced by Ramstead, Constant, Badcock, and Friston (2019), but misapprehend it. The formal ontology that flows from the Bayesian mechanics is not an *a priori* attempt to find or draw boundaries in nature. Rather, it is an attempt to construct *scientific models of these boundaries*, in an *instrumentalist* fashion. The formal ontology only entails that we create empirically evaluable, formal models of organism boundaries. Interestingly, this model of boundaries is itself evaluable via Bayesian model evidence, leading to a nice, reflexive aspect to the framework: Our scientific model of the boundaries of living systems that can be improved iteratively via free energy or prediction error minimization.

Acknowledgements. I wish to express my gratitude to Lancelot Da Costa, Karl Friston, Conor Heins, Alex Kiefer, Dalton Sakthivadivel, and Thomas Parr for helpful discussions that were of great assistance in writing this paper.

Financial support. No funding to report.

Conflict of interest. None.

References

Friston, K., Heins, C., Ueltzhoffer, K., Da Costa, L., & Parr, T. (2021). Stochastic chaos and Markov blankets. *Entropy (Basel)*, *23*(9), 1220. doi:10.3390/e23091220

Giere, R. N. (1999). *Science without laws*. University of Chicago Press.

Giere, R. N. (2010). *Scientific perspectivism*. University of Chicago Press.

Ramstead, M. J., Constant, A., Badcock, P. B., & Friston, K. J. (2019). Variational ecology and the physics of sentient systems. *Physics of Life Reviews*, *31*, 188–205.

Ramstead, M. J., Friston, K. J., & 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.

Van Fraassen, B. C. (1980). *The scientific image*. Oxford University Press.

Enough blanket metaphysics, time for data-driven heuristics

Wiktor Rorot[a,b], Tomasz Korbak[c], Piotr Litwin[b] and Marcin Miłkowski[d]

[a]Faculty of Philosophy, University of Warsaw, 00-927 Warszawa, Poland; [b]Faculty of Psychology, University of Warsaw, 00-183 Warszawa, Poland; [c]Department of Informatics, University of Sussex, Brighton BN1 9RH, UK and [d]Institute of Philosophy and Sociology, Polish Academy of Sciences, 00-330 Warszawa, Poland

w.rorot@uw.edu.pl, tomasz.korbak@gmail.com, piotr.litwin@psych.uw.edu.pl, mmilkows@ifispan.edu.pl

https://wiktor.rorot.pl, https://tomekkorbak.com, http://marcinmilkowski.pl/

doi:10.1017/S0140525X22000280, e206

Abstract

Bruineberg and colleagues criticisms' have been received but downplayed in the free energy principle (FEP) literature. We strengthen their points, arguing that Friston blanket discovery, even if tractable, requires a full formal description of the system of interest at the outset. Hence, blanket metaphysics is futile, and we postulate that researchers should turn back to heuristic uses of Pearl blankets.

Bruineberg and colleagues point out an important, yet hitherto overlooked flaw in the free energy principle (FEP) literature: The term "Markov blanket" has unnoticeably evolved into a more ontologically involved concept of "Friston blanket." However, the gravity of this problem has been underplayed by some of the proponents of the FEP (e.g., Wiese & Friston, 2021, p. 4) who ignore the trouble that the reification of formal concepts leads to. We want to highlight one particular issue for the proponents of the FEP, especially of an associated metaphysical programme of "Markovian monism" (Friston, Wiese, & Hobson, 2020; Wiese & Friston, 2021), concerned with the procedures for identification of Friston blankets in the world.

The problem stems from an important tension: Most other fields of computational modelling use Markov blankets as approximations or optimization tools (e.g., in machine learning for the purpose of dimensionality reduction and variable selection, see Aliferis, Tsamardinos, and Statnikov, 2003; Peña, Nilsson, Björkegren, and Tegnér, 2007; Tsamardinos, Aliferis, and Statnikov, 2003; or for causal search, see Bai et al., 2004; Pellet & Elisseeff, 2008). However, the FEP requires an (in principle) exact identification of a unique Markov blanket for each system of interest, what Friston, Heins, Ueltzhöffer, Da Costa, and Parr (2021a) call a "particular partition." This is necessary because, as Friston argues (2019; Friston et al., 2020), the existence of a Markov (Friston) blanket in a (non-equilibrium) steady-state system is sufficient to prove that the (autonomous, i.e., internal and active) states of the system will "look as if they are trying to minimise (…) the surprisal of states that constitute the thing, particle, or creature. (…) This means that anything that exists must, in some sense, be self-evidencing" (Friston et al., 2020, p. 6). Hence, for Friston, the existence of a particular partition secures that the system will conform to the FEP and allows for deducing it from first principles.

For this reason, in the recent FEP literature, there has been a quickly growing number of attempts to provide solutions to the problem of identifying Markov blankets (e.g., Da Costa, Friston, Heins, and Pavliotis, 2021; Friston et al., 2021a, 2021b). All those attempts focus on providing sufficiently strong approximations, as developing an exact analytical solution to this problem would require solving difficult open problems in partial differential equations. Additionally, researchers in this research community overlook an even more important issue, namely that both strong approximations of Markov blankets, and hypothetical methods for exact solutions to this problem require a full formal description of the system of interest (i.e., the equation describing its dynamics) at the outset. This defeats the practical purpose of finding Markov blankets.

Hence, the paradoxical tension between Markov and Friston blankets we want to highlight is that the pursuit of the metaphysical programme associated with the identification of Friston blankets under the FEP entails intractable mathematical problems that depend on our prior knowledge of the system's dynamics. But if we had a formal description of the system's behaviour, what new knowledge would Friston blankets provide? They certainly would not allow us to find the boundaries of entities of interest in the wild, since those must be assumed for the purpose of description of the system (even if it takes the general form of a Langevin equation, it still requires the assumption that the system is sufficiently stationary). And, if we assumed the whole causal structure of the system beforehand, there would be no need to refer to Pearl or Friston blankets to show that the system will behave in accordance with the FEP, as this would entirely follow from the description of

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

the dynamics. As a consequence, neither this result nor blankets themselves would follow from first principles, but rather from a fallible heuristic analysis of the system of interest.

On the other hand, if we eschew precision and accept approximate optimization methods for finding Pearl blankets such as those widespread in machine learning and causal search (e.g., Pellet & Elisseeff, 2008), we can use them as tools of discovery to identify the boundaries of entities (e.g., nodes in neural networks for the purpose of systems neuroscience). Furthermore, showing that a system delineated in this way conforms to the FEP might provide much more insight into the nature of the process, as it would require less knowledge at the outset. However, approximate methods do not allow for the use of the concept of Friston blanket and effectively preclude the viability of the metaphysical programme of the FEP as a naturalist ontology for life sciences.

Perhaps it is too quick to throw the blankets entirely at this point. Nonetheless, we believe that the use of the Markov blanket construct should enable us to solve pressing issues in computational modelling in the sciences of brain and behaviour. While Markovian monism metaphysics is not such a pressing issue, studying the causal and functional dynamics of cognitive systems is. In this context, we need various fallible heuristics for delineating Pearl blankets; that is, many stupid (Smaldino, 2017), approximate, and tractable models, and we need more of them to be able to make use of the error diversity inherent in any heuristic enterprise (Wimsatt, 2007). While stronger analytical methods for finding Markov (and Friston) blankets are not necessarily dead ends, the FEP theorists' focus on those difficult methods makes them overlook a lot of lower hanging fruits.

Acknowledgements. We want to thank Mel Andrews, Conor Heins, Dalton Sakthivadivel, and the Active Inference Lab for helpful clarifications.

Financial support. This work was supported by the Ministry of Education and Science (Poland) research grant (W. R., DI2018 010448), as part of the "Diamentowy Grant" programme; National Science Centre (Poland) research grant (M. M., P. L., grant number 2014/14/E/HS1/00803) and Leverhulme Doctoral Scholarship (T. K.).

Conflict of interest. None.

References

Aliferis, C. F., Tsamardinos, I., & Statnikov, A. (2003). HITON: a novel Markov blanket algorithm for optimal variable selection. *AMIA 2003 Annual Symposium Proceedings. AMIA Symposium*, 2003, 21–25.

Bai, X., Glymour, C., Padman, R., Ramsey, J., Spirtes, P. L., & Wimberly, F. C. (2004). *PCX: Markov blanket classification for large data sets with few cases*. Center for Automated Learning and Discovery. Retrieved from http://reports-archive.adm.cs.cmu.edu/anon/cald/CMU-CALD-04-102.pdf

Da Costa, L., Friston, K., Heins, C., & Pavliotis, G. A. (2021). Bayesian mechanics for stationary processes. Proceedings of the Royal Society A: Mathematical, *Physical and Engineering Sciences*, *477*(2256), 20210518. https://doi.org/10.1098/rspa.2021.0518

Friston, K., Heins, C., Ueltzhöffer, K., Da Costa, L., & Parr, T. (2021a). Stochastic chaos and Markov blankets. *Entropy*, *23*(9), 1220. https://doi.org/10.3390/e23091220

Friston, K. J. (2019). A free energy principle for a particular physics. *arXiv:1906.10184 [q-bio]*. http://arxiv.org/abs/1906.10184

Friston, K. J., Fagerholm, E. D., Zarghami, T. S., Parr, T., Hipólito, I., Magrou, L., & Razi, A. (2021b). Parcels and particles: Markov blankets in the brain. *Network Neuroscience*, *5*(1), 211–251. https://doi.org/10.1162/netn_a_00175

Friston, K. J., Wiese, W., & Hobson, J. A. (2020). Sentience and the origins of consciousness: From Cartesian duality to Markovian monism. *Entropy*, *22*(5), 516. https://doi.org/10.3390/e22050516

Pellet, J.-P., & Elisseeff, A. (2008). Using Markov blankets for causal structure learning. *Journal of Machine Learning Research*, *9*(43), 1295–1342.

Peña, J. M., Nilsson, R., Björkegren, J., & Tegnér, J. (2007). Towards scalable and data efficient learning of Markov boundaries. *Eighth European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU 2005)*, 45(2), 211–232. Retrieved from https://doi.org/10.1016/j.ijar.2006.06.008

Smaldino, P. E. (2017). Models are stupid, and we need more of them. In R. R. Vallacher, S. J. Read, & A. Nowak (Eds.), *Computational social psychology* (pp. 311–331). Routledge. https://doi.org/10.4324/9781315173726-14

Tsamardinos, I., Aliferis, C. F., & Statnikov, A. (2003). Time and sample efficient discovery of Markov blankets and direct causal relations. *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 673–678. https://doi.org/10.1145/956750.956838

Wiese, W., & Friston, K. J. (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

Wimsatt, W. C. (2007). *Re-engineering philosophy for limited beings: Piecewise approximations to reality*. Harvard University Press.

Markov blankets as boundary conditions: Sweeping dirt under the rug still cleans the house

Javier Sánchez-Cañizares

Mind-Brain Group at the Institute for Culture and Society (ICS), University of Navarra, 31009 Pamplona, Spain
js.canizares@unav.es
https://www.issr.org.uk/fellows/user/496/

doi:10.1017/S0140525X22000097, e207

Abstract

Bruineberg et al. underestimate the ontological weight of Markov blankets as actual boundaries of systems and lean toward an instrumentalist understanding thereof. Yet Markov blankets need not be deemed mere tools. Determining their reality depends on the fundamental problem of distinguishing between system and environment in physics, which, in turn, demands a metaphysical bedrock backed by a realist stance on science.

Do Markov blankets have any ontological weight? Most of the literature on the free energy principle (FEP) equates Markov blankets with actual system boundaries without further ado. However, as the authors rightly point out, there is a crucial difference between using blankets as an epistemic tool in Bayesian networks to identify independencies among random variables (Pearl blankets) and using them as actual boundaries between a system and its environment (Friston blankets).

I agree with the authors on the insufficiency of Friston blankets to demarcate the physical boundary between an agent and its environment, and the need for additional philosophical assumptions "to do such heavy metaphysical lifting." FEP theorists seem aware of this since they recently gave their framework a freshly minted metaphysical interpretation, dubbed Markovian monism (MM) (Friston, Wiese, & Hobson, 2020). According to MM, the very fact that one may demarcate a system from its environment through a Markov blanket induces a dual aspect information geometry in the system's internal states that enables it to represent its surroundings. MM thus reveals the troublesome transition between Markov blankets as epistemic model-bound

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

tools and their alleged ontological consequences, which are beset by circular reasoning (Sánchez-Cañizares, 2021).

However, within FEP formalism, the ontology proper to Markov blankets need not take the brunt of the blow because it inherits the more general problem of distinguishing between a system and its environment in physics. In the absence of a theory of everything, conventional physics must accept initial conditions to start computing the world's behavior. "You need a starting point!" (Wilczek, 2015). Remarkably, the existence of systems seems to require the universe to have exceptional initial conditions and dynamics (Tegmark, 2015) since the theory that is currently most fundamental, that is, quantum mechanics, deems the distinction between system and environment as relative (Lombardi, Fortín, & Castagnino, 2012). Physics teaches us that one must make an *ansatz* to progress in the scientific description of nature. Whether a chosen *ansatz* holds at a specific process only becomes evident *a posteriori*. Markov blankets are undoubtedly the most basic *ansatz* for FEP formalism to work.

Things being so, the authors' critiques of Markov blankets as Friston blankets become less weighty in two interrelated respects:

(1) Markov blankets define the dominion – via the partition of variables – for which it makes sense to minimize free energy. Their precise definition must change whenever the model becomes unsuitable for describing unexpected dynamics, that is, dynamics that cannot be fully grasped within the assumptions of a particular model. But this does not differ from the usual procedure of changing boundary conditions for the distinct models that are compatible with a principle theory. The authors themselves recognize this when quoting Andy Clark, "boundaries are malleable (over time) and multiple." FEP is a principle and needs reinterpretation for each model. One such reinterpretation must state what the system and its environment are, even though such a distinction does not strictly stem from FEP formalism. As a consequence, one should not consider FEP as wholly explanatory of living beings' natural history (Longo & Montévil, 2014), if only because Friston blankets also change throughout the system's history.

(2) Even though Friston blankets are metaphysically wanting, they should not be judged as a mere instrumental tool. The authors, however, ultimately lean toward "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." Instrumentalism undoubtedly looms large in philosophical interpretations of scientific research, but, in the end, this seems to be a self-defeating strategy for several reasons:

    (2.1) The instrumentalist that deems Markov blankets as Pearl blankets forgoes endowing the former with any ontological weight. Yet such a mindset could easily lead to denying the very existence of systems – as sheer constructs of human perception. Nevertheless, if one does not wish to embrace such a radical position and, on the contrary, accepts the existence of systems in the universe, something quite similar to Friston blankets must also exist in each system in order to sieve through the many environmental influences that foster or threaten the system's identity.

    (2.2) It seems paradoxical to emphasize the insufficient justification for transforming Pearl blankets into Friston blankets whereas, ultimately, glossing over what additional philosophical assumptions might look like for such a move. A scientific realist, for instance, could call on formal causation as a valid metaphysical framework that allows for the use of ad hoc boundary conditions to individuate systems that enjoy specific dynamics in nature (Owen, 2020, 2021; Sánchez-Cañizares, 2022a, 2022b). In other words, Markov blankets reflect the emergence of boundary conditions for living systems. If boundary conditions exist for some relevant time scale, Markov blankets are Friston blankets. In addition, such a philosophical commitment can adequately frame the emergence of complex dynamical systems, which one cannot just deduce from their underlying dynamics (Juarrero, 2002; Sánchez-Cañizares, 2016).

    (2.3) The methodological issue at play refers to whether other kinds of knowledge – for example, knowing living systems as wholes – that influence scientific research should be accepted within the overall explanatory picture. Certain pre-scientific knowledge is necessary for guiding scientific methodology. If one assumes said knowledge, there is no fundamental reason to deny that some Markov blankets are also Friston blankets, even if for a limited period, or that Friston blankets are not fixed and may transition in a variety of ways toward different instantiations. In doing so, the inevitable, closed circularity of scientific instrumentalism turns into the open circularity of scientific realism, which admits a hierarchical variety of assumptions and hypotheses about the world, as well as the possibility of cognitive progress based on constant confrontation with observation.

Financial support. This work has been funded with the help of Fundación Ciudadanía y Valores (FUNCIVA) and Proeduca Summa S.L.

Conflict of interest. None.

References

Friston, K. J., Wiese, W., & Hobson, J. A. (2020). Sentience and the origins of consciousness: From Cartesian duality to Markovian monism. *Entropy*, *22*(5), 516. https://doi.org/10.3390/E22050516

Juarrero, A. (2002). Complex dynamical systems and the problem of identity. *Emergence*, *4*(1), 94–104.

Lombardi, O., Fortín, S., & Castagnino, M. (2012). The problem of identifying the system and the environment in the phenomenon of decoherence. In H. W. de Regt, S. Hartmann & S. Okasha (Eds.), *The European philosophy of science association proceedings: Amsterdam 2009* (pp. 161–174). Springer. https://doi.org/10.1007/978-94-007-2404-4_15

Longo, G., & Montévil, M. (2014). *Perspectives on organisms: Biological time, symmetries and singularities*. Springer.

Owen, M. (2020). Aristotelian causation and neural correlates of consciousness. *Topoi*, *39* (5), 1113–1124. https://doi.org/10.1007/s11245-018-9606-9

Owen, M. (2021). Circumnavigating the causal pairing problem with hylomorphism and the integrated information theory of consciousness. *Synthese*, *198*, 2829–2851. https://doi.org/10.1007/s11229-019-02403-6

Sánchez-Cañizares, J. (2016). Entropy, quantum mechanics, and information in complex systems: A plea for ontological pluralism. *European Journal of Science and Theology*, *12*(1), 17–37.

Sánchez-Cañizares, J. (2021). The free energy principle: Good science and questionable philosophy in a grand unifying theory. *Entropy*, *23*(2), 238. https://doi.org/10.3390/e23020238

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

Sánchez-Cañizares, J. (2022a). Formal causation in integrated information theory: An answer to the intrinsicality problem. *Foundations of Science*, *27*, 77–94. https://doi.org/10.1007/s10699-020-09775-w
Sánchez-Cañizares, J. (2022b). Integrated information theory as testing ground for causation: Why nested hylomorphism overcomes physicalism and panpsychism. *Journal of Consciousness Studies*, *29*(1–2), 56–78. https://doi.org/10.53765/20512201.29.1.056
Tegmark, M. (2015). Consciousness as a state of matter. *Chaos, Solitons and Fractals*, *76*, 238–270. https://doi.org/10.1016/j.chaos.2015.03.014
Wilczek, F. (2015). *A beautiful question. Finding nature's deep design*. Penguin Press.

A continuity of Markov blanket interpretations under the free-energy principle

Anil Seth, Tomasz Korbak and Alexander Tschantz

Department of Informatics, School of Engineering and Informatics, University of Sussex, Brighton BN1 9QJ, UK
a.k.seth@sussex.ac.uk
tomasz.korbak@gmail.com
tschantz.alec@gmail.com

doi:10.1017/S0140525X2200036X, e208

Abstract

Bruineberg and colleagues helpfully distinguish between instrumental and ontological interpretations of Markov blankets, exposing the dangers of using the former to make claims about the latter. However, proposing a sharp distinction neglects the value of recognising a continuum spanning from instrumental to ontological. This value extends to the related distinction between "being" and "having" a model.

"We should not confuse the foundations of the real world with the intellectual props that serve to evoke that world on the stage of our thoughts." This quote from Ernst Mach (Mach [2012], p. 531, translated in Sigmund [2017], p. 19), surfacing from the origins of the philosophy of science, connects directly to the target article, in which Bruineberg and colleagues discuss how Markov blankets (MBs) should be understood within the wider literature of the free energy principle (FEP, Friston, 2010), as well as how "models" and "modelling" should be interpreted within the cognitive and brain sciences more generally.

MBs are statistical descriptions that partition systems into internal, external, and blanket variables – where the internal variables are conditionally independent of the external variables, given the blanket variables. Bruineberg et al. provide a valuable service by distinguishing two interpretations of MBs: "Pearl blankets" (PBs) and "Friston blankets" (FBs). PBs embody an instrumental approach, in which MBs are used as tools to aid the analysis of complex systems, for example by identifying sets of variables suitable for further investigation. In contrast, FBs adopt an ontological stance in which they are assumed to either *be* (a literalist reading) – or *usefully approximate* (a realist reading) actually existing boundaries in the world, such as the boundary between a cell and its milieu, or between an organism or agent and its environment. Bruineberg et al. reveal the dangers of conflating these two interpretations, in particular when an instrumental (PB) application is implicitly or explicitly taken to justify ontological (FB) conclusions. Their arguments should be borne in mind by those inclined to help themselves to the FEP to explain their favourite grand mystery, or to take it as gospel.

Having said this, making a sharp distinction is often a useful prelude to recognising a spectrum of positions, each of which may be useful when assessed on its own merits. We suggest this is the case here. For example, one may begin with an instrumentalist approach and progressively refine and extend the corresponding model so as to make increasingly specific claims about the causal mechanisms at play in the system under study – in this way, gradually moving towards a more ontological or realist stance. What does "refine and extend" mean? It could mean equipping the model with additional features that represent potentially important and context-specific aspects of the relevant boundaries, such as autopoietic (self-producing) processes for biological boundaries (Kirchhoff, Parr, Palacios, Friston, & Kiverstein, 2018; Maturana & Varela, 1980), and embodied and embedded interactions for cognitive boundaries (Clark & Chalmers, 1998; Kirchhoff & Kiverstein, 2021), as well as a recognition of the limited degree to which statistical identification of an MB might generalise to nonequilibrium systems (Aguilera, Millidge, Tschantz, & Buckley, 2022; Biehl, Pollock, & Kanai, 2021).

Bruineberg et al. mention these possibilities, but downplay their significance by drawing a contrast between "additional philosophical assumptions" and "additional technical assumptions," where the latter implicitly subsumes everything just mentioned. But these modelling strategies and mathematical constraints are more than just additional assumptions, they can often be part-and-parcel of the explanatory model itself. And rather than "additional philosophical assumptions," what seems to be required is a *recognition* of the model's philosophical context and the claims made on its behalf, so as to avoid the sort of conflation helpfully identified by Bruineberg.

Bruineberg et al. worry that, given such additional aspects, whether the MB formalism itself can still be doing any work? The answer is yes, to the extent that it helps specify those parts of a model that are focused on boundaries. By casting the distinction between PBs and FBs as sharp, rather than as extrema on a continuum, Bruineberg et al. underestimate the explanatory work that MBs may uniquely be able to do.

Digging a little deeper, one reason we might be tempted to invoke a bright-line distinction between PBs and FBs is because of the dramatic claims made for literalist readings of FBs, in which MBs are seen as really existing ontological boundaries in physical systems. But – as Mach reminds us – models are always models, whatever their granularity. Once we discount the relevance of an overly literalist reading, the value of a continuity between instrumental and ontological stances becomes easier to appreciate. (Here, it is worth separating Mach's valuable scepticism about literalism from his ultimately doomed project to ground physics solely in phenomenology; it is not likely that Mach would have had much time for FBs, even of a realist flavour.)

The same reasoning can be applied to the distinction between "being a model" and "having a model" – a distinction that Bruineberg et al. mention, but only in passing (see also Seth and Tsakiris, 2018). Under the FEP, and following the spirit of the cybernetic pioneers (Conant & Ashby, 1970), many systems can be interpreted as "being" a model of their environment. In an example briefly discussed by Bruineberg et al., even a simple Watt governor can be described as performing inference – however it is best thought of not as *having* a model that is used to

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

perform inference, but as *being* a model of its environment, from the perspective of an external observer (see Van Gelder [1995] for the original and still instructive version of this argument, in the context of computational theories of mind). By contrast, neurocognitive systems that are modelled as implementing generative models of their sensorium, in order to perform inference through prediction error minimization, are better described as *having* models, rather than merely *being* models.

This distinction is important, because the status of having (rather than being) a model may speak to a variety of interesting phenomena, such as the potential for counterfactual cognition, imagination and imagery, volitional action of various kinds, and perhaps even the difference between conscious and unconscious perception. Methodologically, the hypothesis that a system *has* a model can be warranted if having that hypothesis leads to novel testable predictions that would not have been made without that hypothesis (see Chemero [2000] for a related argument). Again, it is beneficial to recognise that this distinction comes in degrees, and that even the (realist, ontological) claim that a system *has* a model should not confuse the map with the territory.

The broader lesson from Bruineberg et al. is the need for a healthy interaction across disciplinary boundaries, and especially among philosophy, physics, biology, and cognitive science, in order to avoid the pitfalls of explanatory overreach, and to take advantage of the many opportunities that arise at disciplinary boundaries. Ernst Mach – a physicist who eventually took a Chair in the Department of Philosophy at the University of Vienna, making lasting contributions to psychology and physiology along the way – exemplifies these virtues.

Acknowledgements. The authors are grateful to Chris Buckley and Miguel Aguilera for helpful comments.

Financial support. This work was supported by the European Research Council (AKS, Advanced Investigator grant number 101019254), by the Canadian Institute for Advanced Research (AKS and AT, CIFAR Program on Brain, Mind, and Consciousness), by the Dr. Mortimer and Theresa Sackler Foundation (The Sackler Centre for Consciousness Science, AKS and AT), and by the Leverhulme Trust (AKS and TK, Doctoral Scholarship Programme grant number DS-2017-011).

Conflict of interest. None.

References

Aguilera, M., Millidge, B., Tschantz, A., Buckley, C. L. (2022). How particular is the physics of the free energy principle? *Physics of Life Reviews*, *40*, 24–50.

Biehl, M., Pollock, F. A., & Kanai, R. (2021) A technical critique of some parts of the free energy principle. *Entropy (Basel)*, *23*(3), 293.

Chemero, A. (2000). Anti-representationalism and the dynamical stance. *Philosophy of Science*, *67*(4). https://doi.org/10.1086/392858

Clark, A., & Chalmers, D. J. (1998) The extended mind. *Analysis*, *58*, 10–23.

Conant, R, & Ashby, W. R. (1970) Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, *1*(2), 89–97.

Friston, K. J. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138.

Kirchhoff, M., Parr, T., Palacios, E., Friston, K. J., & Kiverstein, J. (2018). The Markov blankets of life: Autonomy, active inference and the free energy principle. *Journal of the Royal Society, Interface*, *15*(138). https://doi.org/10.1098/rsif.2017.0792

Kirchhoff, M. D., & Kiverstein, J. (2021). How to determine the boundaries of the mind: A Markov blanket proposal. *Synthese*, *198*, 4791–4810.

Mach, E. (2012). Die mechanik in ihrer entwicklung. In *Ernst Mach studienausgabe*. Xenomoi.

Maturana, H., & Varela, F. (1980). *Autopoiesis and cognition: The realization of the living*. Boston Studies in the Philosophy of Science, Vol. *42*. D. Reidel.

Seth, A. K., & Tsakiris, M. (2018). Being a beast machine: The somatic basis of selfhood. *Trends in Cognitive Sciences 22*(11), 969–981.

Sigmund, K. (2017). *Exact thinking in demented times*. Basic Books.

Van Gelder, T. (1995) What might cognition be if not computation? *Journal of Philosophy*, *92*(7), 345–381.

Blankets, heat, and why free energy has not illuminated the workings of the brain

Donald Spector[a] and Daniel Graham[b]

[a]Department of Physics, Hobart and William Smith Colleges, Geneva, NY 14456, USA and [b]Department of Psychological Sciences, Hobart and William Smith Colleges, Geneva, NY 14456, USA
spector@hws.edu
graham@hws.edu
http://people.hws.edu/spector/
http://people.hws.edu/graham/

doi:10.1017/S0140525X22000188, e209

Abstract

What can we hope to learn about brains from the free energy principle? In adopting the "primordial soup" physical model, Bruineberg et al. perpetuate the unsupported notion that the free-energy principle has a meaningful physical – and neuronal – interpretation. We examine how minimization of free energy arises in physical contexts, and what this can and cannot tell us about brains.

To determine the implications of applying free-energy principles to the study of the brain, it is worth examining how free energy arises in physics in the first place, and then considering the implications for studies of the brain. We focus on two questions: What is the functional content of applying a free-energy principle to the brain? If the free-energy principle does work phenomenologically, can it tell us about the underlying workings of the brain?

Free energy arises in thermodynamics, the field that describes the bulk behavior of large systems. Statistical mechanics, in turn, is the field that derives thermodynamics from more fundamental principles. Via the ergodic hypothesis, statistical mechanics says that the bulk properties of a system (macrostates) can be found by ignoring the detailed dynamics of the intractably large number of microstates, and instead performing ensemble averages over the possible microstates (with equal likelihoods in isolated systems, which implies Boltzmann weightings at finite temperature). The bulk properties found by ensemble averages in statistical mechanics can alternatively be found thermodynamically, by minimizing the quantity known as the free energy.

The power of free energy is thus not that it is optimized at equilibrium; after all, there are non-thermodynamic optimization problems. It is that there is a language of macrovariables which can characterize a system, while the underlying microvariables evolve in a way functionally indistinguishable from randomly. But as invoked in the target article, the free-energy principle does not point to any macro- or microvariables, which are needed

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

for either a high- or low-level understanding of the workings of the brain.

If the free energy principle in the study of the brain is to be useful, we should hope that the process of deriving thermodynamics from statistical mechanics can be run in reverse: That establishing the efficacy of a free-energy principle to describe the behavior and representational strategies of agents with a brain can reveal the fundamental dynamics of the brain. Even if we posit that there are microstates, all that is required for thermodynamics to arise is that the dynamics cause those microstates to be sampled over time with the correct weightings to allow the ensemble average to mimic the dynamics. Alas, this does not uniquely determine the underlying microstate dynamics.

Imagine thermodynamics had been invented before Newtonian mechanics. Could one deduce Newton's laws of motion from this formalism? The answer is no. For example, a gas at finite temperature can be modeled using kinetic theory or using a Metropolis algorithm. These provide different dynamical rules on microstates that produce the same thermodynamics; thermodynamics alone cannot reveal the fundamental dynamics. While broad results like entropy maximization arise in a general framework, to use statistical mechanics to obtain the thermodynamics of specific systems relies on knowledge about those systems extrinsic to thermodynamics, already obtained in other contexts.

Furthermore, even if one has posited micro-level dynamics for the brain, producing a thermodynamic language still requires identifying suitable macrostate variables. When tossing 1 million coins, if, instead of focusing on which particular coins are heads or tails (the microstates), we label states just by their total numbers of heads and tails, we can perform a free-energy style analysis to get the average behavior (and show fluctuations from this are negligible). This methodology hinges on choosing appropriate macrostate variables (e.g., the number of heads, not, say, the number of heads squared). Without a suitable analogous connection between microstates and macrostates, the promise of a free-energy principle for the brain remains unfulfilled.

Of course, if the brain does achieve certain equilibrated behavioral states, one could by construction create a free-energy function that said states minimize. Leaving aside the potential tautology of this philosophy, the question remains, what are those states? What macrovariables are static in equilibrium? Perhaps more importantly, how are they connected to the microstates of the brain? Should we focus on neuronal states or their interactions? Should we describe the brain in terms of synaptic events, spikes, spike timing, oscillations, local potentials, voxel-wise patterns, or some combination of these? What microstates can be lumped together into useful macrostates, and by what rules?

Although the brain is complicated, accepting ignorance of its workings is untenable (imagine if thermodynamics itself had stopped with Carnot's generation and we never developed statistical mechanics and all that ensued). Still the free-energy principle could be used to solve real-world problems with a set of well-understood affectors and effectors, that is, in situations like neurorobotics where we do not necessarily want to model the brain but do want "intelligent" solutions to environmental challenges.

As we think about thermodynamics and brains, let us imagine how mysterious heat must have seemed at first. But heat, it turned out, was not a new form of energy, simply familiar forms of energy carried by degrees of freedom whose details were no longer being tracked. In studies of the brain, what plays the role of heat (or any other thermodynamic quantity), not literally, but as a seemingly distinct macro feature that embodies hidden micro behavior?

The free-energy principle for brains is couched in the language of statistical mechanics but not justified by it. However, we would welcome attempts to work from brain microstates to a thermodynamic approach (and see what variables or principles are useful). Whatever the differences between the principles that prevail in brains and those relevant to physics, we still stand a better chance of understanding both the brain and behavior through the analogous study of principles in the brain as opposed to ensemble properties with unknown relationships to microstates and microstate dynamics. This is essentially the "inside-out" approach to systems neuroscience (Buzsáki, 2019). For example, what rules govern and how does the brain manage flexible, brain-wide communication flow on a neuronal network with short paths between essentially any populations of neurons (Graham, 2021)? If elucidating principles of brain function proves successful, we could interrogate the entire system of many physically networked elements and their interaction with the environment directly, and potentially dispense with blankets altogether.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Conflict of interest. None.

References

Buzsáki, G. (2019). *The brain from inside out*. Oxford University Press.
Graham, D. (2021). *An internet in your head*. Columbia University Press.

Good theoretical debate, but insufficient proof of concept

Rainer Spiegel

Internal Medicine Section, Department of Acute Medicine, Basel University Hospital, 4031 Basel, Switzerland
rainer_spiegel@hotmail.com
rainer.spiegel@usb.ch
https://www.researchgate.net/profile/Rainer-Spiegel

doi:10.1017/S0140525X22000140, e210

Abstract

Bruineberg and colleagues argue that the patellar reflex cannot be modeled sufficiently with a Friston blanket due to counterintuitive sensorimotor boundaries. Although I agree with their theoretical discussion, their model of the patellar reflex is insufficiently based on clinical knowledge. Consequently, this example should not be applied to challenge Friston blankets. I will provide an alternative example.

After explaining Markov and Friston blankets in particular, Bruineberg et al. demonstrate how difficult it is for these to adequately enclose real-world examples. One reason is the assumption of conditional independence, which they are based on. To

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

underline their point of view, they choose an example from clinical medicine: To model the patellar reflex, they construe a simple Bayesian network and a Friston blanket.

I agree with Bruineberg and colleagues that the assumption of conditional independence in Friston blankets is problematic for real-world examples. Consequently, their theoretical debate on Friston blankets makes sense. However, their example of the patellar reflex is insufficiently based on knowledge from neurophysiology or clinical medicine. This makes the example in this context problematic. As a result, I argue that the authors should refrain from applying this example as a proof of concept for their theoretical arguments. To make it clear what is problematic about their choice of example, let me first describe how they try to simulate the patellar reflex.

They propose a Bayesian network where different nodes are dedicated to different purposes: The patient's intention to move the leg, the doctor's intention to move the patient's leg, the spinal neurons, the intervention of striking the patellar tendon with a hammer, the motor command from the central nervous system sent to the spinal neurons, and finally, a third way (e.g., someone else) moving the leg. An example for their Bayesian network can be found in Figure 7a of their target article. Subsequently, they present an elaboration of the Bayesian network, where the nodes are partitioned into external states, internal states, sensory states, and active states. This partitioning is done in what they call a "Friston blanket," which is a transfer of the Markov blanket idea from statistics (Pearl, 1988) to the life sciences (e.g., Friston, 2013).

The Friston blanket from this example is problematic because of the following reasons:

Bruineberg et al. use nodes that incorporate the patient's intention. However, the patient's intention as well as the central motor commands are not part of a monosynaptic reflex arc. Consequently, they are not part of the patellar reflex, which is a monosynaptic reflex arc. Rather, the incorporation of intention would involve several neurons, interneurons and therefore several synapses between neurons. A reflex arc with several synapses, however, would be termed a polysynaptic reflex arc. Therefore, Bruineberg et al. have construed a Friston blanket for a polysynaptic reflex arc, although their aim was actually to model a monosynaptic reflex arc. As a result, they cannot argue that their Friston blanket does not adequately enclose the patellar reflex. To understand why their example cannot be applied as an argument against Friston blankets, let us consider the patellar reflex, which should not incorporate the patient's intention, as the synaptic transmission happens at the level of the spinal cord: After striking the patellar tendon with a hammer, the muscle spindle in the quadriceps femoris muscle is activated, followed by an afferent signal traveling along the sensory neuron to the dorsal root of the spinal cord. In the spinal cord, a monosynaptic transmission to an alpha-motor neuron takes place, which produces an efferent signal traveling along this alpha-motoneuron to the quadriceps femoris muscle, eliciting the movement (e.g., Ginanneschi, Mondelli, Piu, and Rossi, 2015). Although some interaction with interneurons at the level of the spinal cord is possible (Ginanneschi et al., 2015), it is obvious that the Friston blankets construed by Bruineberg et al., Figures 7b and 7c, involve intentional leg movements, which should not be the case.

If Bruineberg et al. had built a Friston blanket for a truly monosynaptic reflex, they could easily avoid the counterintuitive sensorimotor boundaries that they consider problematic for Friston blankets. I will attempt to do this by using the same terms and the same states as Bruineberg et al. in Figures 7b and 7c. I will only get rid of those parts that do not belong to a monosynaptic reflex arc: The patient's intention *ip* is not part of the monosynaptic patellar reflex, nor is the motor command sent from the central nervous system *c*, nor someone else kicking the patient's leg *k*. Rather, there would be nodes for the doctor's intention *id – external state*, the hammer *h – sensory state*, the spinal neuron *s – internal state*, and the motor command *m active state*. Given that all nodes represent different states and each state can lie on a different Friston blanket, there would be no counterintuitive sensorimotor boundaries in this example. Because Bruineberg et al. have not modeled a truly monosynaptic reflex arc with their choice of nodes, they see a problem. This problem vanishes when omitting the nodes that are not part of a monosynaptic reflex arc. Does this imply that their critique on counterintuitive sensorimotor boundaries with Friston blankets is not justified? I would argue that their critique is still justified and here is the reason why: There are several conditions in clinical medicine that would challenge Friston blankets. For example, the patellar reflex would require additional nodes for the muscle spindle / the quadriceps muscle. If these were added, as well as other internal states that change the extent of reflexes, for example, endocrinopathies (Rodriguez-Beato & De Jesus, 2021), electrolyte derangements (Espay, 2014; Hensle & Lambert, 2010), there would be several internal states. Consequently, there would be several Friston blankets with counterintuitive boundaries all impacting the extent of the patellar reflex, which would challenge Friston blankets.

To conclude, Bruineberg et al. contribute substantially to the theoretical debate on Friston blankets, but their idea to challenge Friston blankets with the patellar reflex example does not work due to the aforementioned shortcomings of their chosen model. If their example is modified and other relevant nodes for the patellar reflex are added, one can easily find counterintuitive sensorimotor boundaries, which would be a challenge for Friston blankets.

Financial support. No funding received in relation to this manuscript.

Conflict of interest. No conflict of interest in terms of this commentary. Considering all possible conflicts of interest, Rainer Spiegel owns a small number of stocks from the respirator/ventilator company Draeger in his private portfolio.

References

Espay, A. J. (2014). Chapter 23 – Neurologic complications of electrolyte disturbances and acid-base. In J. Biller & J. M. Ferro (Eds.), *Handbook of clinical neurology* (pp. 365–382). Elsevier. https://doi.org/10.1016/B978-0-7020-4086-3.00023-0

Friston, K. J. (2013). Life as we know it. *Journal of the Royal Society Interface*, *10*(86), 20130475. https://doi.org/10.1098/rsif.2013.0475

Ginanneschi, F., Mondelli, M., Piu, P., & Rossi, A. (2015). Pathophysiology of knee jerk reflex abnormalities in L5 root injury. *Functional Neurology*, *30*(3), 187–191. https://doi.org/10.11138/fneur/2015.30.3.187

Hensle, T. W., & Lambert, E. H. (2010). Chapter 3 – Renal function, fluids, electrolytes, and nutrition from birth to adulthood. In J. P. Gearhart, R. C. Rink & P. D. E. Mouriquand (Eds.), *Pediatric urology* (2nd ed., pp. 23–30). Saunders. https://doi.org/10.1016/B978-1-4160-3204-5.00003-7

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible Inference* (Chapter 3.2, pp. 96–115). Morgan Kaufmann.

Rodriguez-Beato, F. Y., & De Jesus, O. (2021). *Physiology, deep tendon reflexes* (Updated 26 July 2021). StatPearls. https://www.ncbi.nlm.nih.gov/books/NBK562238/

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

What realism about agents requires

Mark Sprevak

School of Philosophy, Psychology and Language Sciences, University of Edinburgh, Edinburgh, UK
mark.sprevak@ed.ac.uk
https://marksprevak.com/

doi:10.1017/S0140525X22000164, e211

Abstract

Bruineberg et al. argue that the formal notion of a Markov blanket fails to provide a single principled boundary between an agent and its environment. I argue that one should not expect a general theory of agenthood to provide a single boundary; and the reliance on auxiliary assumptions is neither arbitrary nor reason to suspect instrumentalism.

Bruineberg et al. distinguish a metaphysically robust use from a merely formal use of the concept of a Markov blanket (Friston vs. Pearl blankets). They argue that Friston blankets are only able to do the work required of them *if they yield a single principled boundary* between the agent and world. They argue that Friston blankets cannot do this (sect. 5). Reasons include that a Friston blanket depends on a number of non-trivial assumptions that don't flow purely from the formalism, including the choice of which Bayesian network one uses to model the system. They conclude that Friston blankets cannot do the work required of them to demarcate agents from world. They suggest an alternative role for Friston blankets as merely instrumental constructs rather than as real boundaries in the world.

Bruineberg et al. present a stark divide: either a Friston blanket provides a *single*, *objective*, *principled boundary* or it is merely an *instrumental construct*. While Bruineberg et al. are correct on many points about the limitations of Friston blankets, this central dilemma mischaracterises the intention and potential future prospects of that notion.

First, it is unclear whether Friston blankets were intended to meet, or even should meet the exacting standard of yielding a *single principled boundary*. The idea that there is a single, objectively correct way to divide the world up into states that are "inside" and "outside" agents is deeply suspect (Craver, 2009). Agents are nested inside each other and their boundaries crosscut. From various perspectives, individual humans, groups of humans, nations, brain regions, individual cells, and sub-cellular assemblies count as agents (Dennett, 2017; Huebner, 2014; Kingma, 2019). When attempting to distinguish an agent from the world, one's first question should be "What *sort* of agent is one talking about?." Attempting to identify agential boundaries without making assumptions about the specific physical differences and similarities that matter to that kind of agent's identity and integrity – that is, that determine one's subject matter – does not make sense. One should not expect the way one partitions the world into agents to be indifferent to the type of agent and agenthood one is interested in (e.g., planetary-scale agents vs. cellular agents).

Second, the authors rightly emphasise the role of auxiliary assumptions in applying the notion of a Friston blanket. The auxiliary assumptions are needed to link the formal notion of a Markov blanket to the physical world – to determine what are the principal variables of the target system, the kinds of stability one is interested in (and over what timescale and set of possible interventions), and which Bayesian network should model the physical system. However, with less justification, they suggest that these auxiliary assumptions are arbitrary, pragmatic, or merely instrumental. There is little reason to think this. The assumptions appear to be necessary, motivated, and unavoidable even to a realist. Before partitioning the world into agents, one has to decide the type of agent one is talking about. This explains why Friston's example (sect. 4) has to make non-trivial assumptions about which forces should be considered as relevant in the target system (electrochemical) and which threshold to apply to interactions between particles (how much is required for a connection). It also explains why the agential boundary is relative to which Bayesian network one chooses to model the system – this specifies the sort of invariances, dependencies, and physical variations one wishes to consider. These are not merely pragmatic issues, concerned with convenience or the personal preferences of the modeller. They are necessary to settle the subject matter. If one is interested in certain forms of stability and manipulation, then the world divides into certain sorts of agents. If one is interested in other forms of stability, then the world divides into a different set of agents. Reliance on these assumptions does not entail that agenthood is conventional or pragmatic. It is needed because one must decide what kind of agent one is talking about before asking the question of where its boundaries lie.

Regarding the "reification fallacy," it is worth bearing in mind that liberal talk here is relatively commonplace in the applied sciences and it is not necessarily indicative of a confusion regarding map and territory. Consider a simpler formal notion: the arithmetic mean of a set of numbers. In the language of the authors, this counts as a feature of the map as it is defined over numbers, not over any concrete physical features. Yet we regularly ascribe arithmetic means to the territory: We may refer to my *mean coffee consumption*, my *mean income*, or my *mean bodyweight*. What permits this slippage from map to territory? Is it an illicit reification? No. In each case, the ascription presupposes a range of assumptions that connect select aspects of the physical territory with abstract numbers over which an arithmetic mean is defined and may be calculated. Different schemes for representing my coffee consumption with numbers may result in different numerical means being attributed to the territory. Similarly, when proponents of active inference use Markov blankets to demarcate agents, *by necessity* they must employ a background of auxiliary assumptions about which physical features in the physical system matter and how they should be formally represented in the Markov framework.

Bruineberg et al. are right that proponents of active inference should be more explicit about these assumptions. But they give no reason to think that those assumptions are unprincipled or instrumental conceits. The intention of Friston's proposal – which has arguably been obscured by loose talk about "just applying the maths" – is that it identifies a formal pattern that is characteristic of agenthood and that may be manifest in different ways in different contexts given different auxiliary assumptions. This yields multiple crosscutting agential boundaries, but that outcome should be expected on any theory of agenthood. In light of what Bruineberg et al. say, there is no reason to think that the notion of a Friston blanket could not serve as the *formal part* of a version of realism about agents worth wanting.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

Conflict of interest. None.

References

Craver, C. F. (2009). Mechanisms and natural kinds. *Philosophical Psychology*, *22*, 575–594.
Dennett, D. C. (2017). *From bacteria to Bach and back: The evolution of minds*. W. W. Norton.
Huebner, B. (2014). *Macrocognition*. Oxford University Press.
Kingma, E. (2019). Were you a part of your mother? *Mind; A Quarterly Review of Psychology and Philosophy*, *128*, 609–646.

Against free energy, for direct perception

Thomas A. Stoffregen[a] and Robert Heath[b]

[a]The School of Kinesiology, University of Minnesota, Minneapolis, MN 55455, USA and [b]Hiawatha Valley Education District, Winona, MN 55987, USA
tas@umn.edu
ptbob55987@yahoo.com
https://apal.umn.edu/

doi:10.1017/S0140525X22000103, e212

Abstract

We question the free energy principle (FEP) as it is used in contemporary physics. If the FEP is incorrect in physics, then it cannot ground the authors' arguments. We also question the assumption that perception requires inference. We argue that perception (including perception of social affordances) can be direct, in which case inference is not required.

We raise an issue relating to the physics that undergird the free energy principle (FEP), and one relating to whether the FEP actually is relevant to perception, and to phenomena of animacy, in general.

The authors argue the merits of various formulations of the FEP in relation to animacy (e.g., Markov blankets, Friston blankets, Pearl blankets). They acknowledge that "the core of the FEP rests upon an intertwined web of mathematical constructs borrowed from physics" (sect. 1, para. 4). They do not question the validity of the underlying physics of the FEP. However, a consistent thread of scholarship raises doubts about the validity of the FEP as a description of physical reality (rather than as a mathematical abstraction that is not meant to be taken as a claim about reality; Schrodinger, 1952a, 1952b).

All versions of the FEP assume that time is discrete. That is, the mathematical equations of the FEP are defined only if we assume that time is discrete. In most physics, models and theories are structured in ways that assume that time exists as discrete temporal units. This assumption is accepted even by scholars who have criticized other aspects of the FEP (e.g., Colombo and Palacios, 2021; Raja, Valluri, Baggs, Chemero, and Anderson, 2021; Unnikrishnan, 2020). Yet not everyone accepts this assumption. Bergson (1922/1999) claimed that time does not exist in discrete units but, rather, exists as a continuum that cannot be sectioned into discrete units (Robbins, 2014). At minimum, Bergson's alternative conception of time alerts us to the fact that mainstream views of discrete time are assumptions or descriptions, rather than established facts (Schrodinger, 1952a, 1952b). Claims that are based on this assumption, such as Friston's FEP and the current authors' treatment, should more explicitly acknowledge their reliance on these contingent assumptions. It is also important to carefully evaluate the validity of Bergson's alternative perspective and the implications it may have for our understanding of both physics and animacy.

Bergson's (1922/1999) conception of time is consistent with Gibson's conception of physics, including time. Gibson argued that traditional physics, including electromagnetism, thermodynamics, quantum mechanics, and abstract, discrete time, cannot account for the phenomena of animacy (Gibson, 1975, 1979). More broadly, Rosen (1991) argued that living things rely on physical principles (what he referred to as "new physics") that are primary to the physics of inanimacy. Put plainly, each of these scholars raised deep questions, not only about the presumed primacy of traditional physics, but also about its literal accuracy as a description of reality. Friston's FEP is part of an ancient tradition by which the physics of inanimacy are assumed to be basic, with the physics of animacy being derivative. Bergson, Gibson, Rosen, and others argue just the opposite: That the physics of animacy are primary, and the physics of inanimacy derivative.

The above considerations relate intimately to our second issue, which concerns the authors' assumption that perception is inferential. They offer as options only inference with a model, or inference within a model. But other options exist. The ecological approach to perception and action claims that the animal–environment interaction lawfully structures patterns in ambient energy such that reality is *specified* (e.g., Gibson, 1966; Turvey, 2019). If reality is specified, then perception can be direct and, consequently, there is no requirement for inference. Bruineberg, Chemero, and Rietveld (2019) accepted this logic, but argued that social affordances cannot be specified and that, therefore, perception of social affordances cannot be direct. It would follow that knowledge of social affordances must depend upon inference. The sole basis for their argument was the fact that social affordances emerge from social conventions, such as linguistic grammar and syntax, or highway speed limits. However, they offered no evidence, either logical or empirical, that social conventions or social affordances actually cannot be specified.

The fact that social affordances emerge from social conventions does not imply that they are free of physical law, such that they cannot be specified. Social conventions are constrained by physical law. For example, all phonetic systems must conform to the acoustic capabilities of the speech organs. Similarly, grammar and syntax, which vary widely across languages, nevertheless exhibit consistencies, and cannot operate outside physical law. Language is used to communicate about physical reality, such that grammar and syntax may be constrained by the physical laws that constrain the events that are the principal subject of linguistic interaction (e.g., Anthony, 2007). Even metaphor is grounded in embodied experience (Gibbs, Lima, & Francozo, 2004). In short, the claim that social affordances cannot be specified and that, therefore, perception must be inferential, is a claim, rather than a self-evident fact. It may be that social affordances are specified in conformity with physical law, such that all perception can be direct (e.g., Stoffregen and Bardy, 2001; Stoffregen, Mantel, and Bardy, 2017).

Empirical research can help to address the continuity or unity of perception. Empirical research is consistent with the idea that perception of social affordances may be direct. As one example, human observers can transition easily between perception of personal and interpersonal affordances (e.g., Richardson, Marsh, and

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

Baron, 2007), suggesting that perception of personal and interpersonal affordances may have a similar basis. Perception of social affordances may be grounded in the perception and control of affordances for the individual. For example, locomotor experience (typically, learning to crawl) causally drives the infant's developing understanding of referential communication (e.g., Campos et al., 2000), while the physical experience of interpersonal synchrony has causal influence on the development of prosocial behavior (Cirelli, 2018), and social conventions are taught through guided interactions (Nonaka & Stoffregen, 2020; Reed, 1996). These findings suggest that perception of social affordances may emerge from the kinds of physical interactions that Bruineberg et al. (2019) accepted as being amenable to direct perception.

Acceptance of Friston's FEP mandates rejection of any form of direct perception (e.g., Friston, 2013). This stark requirement may explain the uncritical nature of the authors' views on specification (Bruineberg et al., 2019). The alternative is equally stark: If perception is direct, then Friston's FEP cannot be a factual description of animate systems.

Financial support. Thomas A. Stoffregen was supported by NSF-1901423, CHS: Medium: Prediction, Early Detection, and Mitigation of Virtual Reality Simulator Sickness.

Conflict of interest. None.

References

Anthony, D. W. (2007). *The horse, the wheel, and language*. Princeton University Press. https://doi.org/10.1515/9781400831104

Bergson, H. (1999). *Duration and simultaneity*. Trans. Leon Jacobson. Clinaman Press. (Original work published in 1922).

Bruineberg, J., Chemero, A., & Rietveld, E. (2019). General ecological information supports engagement with affordances for "higher" cognition. *Synthese*, *196*, 5231–5251. doi:10.1007/s11229-018-1716-9

Campos, J. J., Anderson, D. I., Barbu-Roth, M., Hubbard, E. M., Hertenstein, M. J., & Witherington, D. (2000). Travel broadens the mind. *Infancy*, *1*, 149–219.

Cirelli, L. K. (2018). How interpersonal synchrony facilitates early prosocial behavior. *Current Opinion in Psychology*, *20*, 35–39. http://dx.doi.org/10.1016/j.copsyc.2017.08.009

Colombo, M., & Palacios, P. (2021). Non-equilibrium thermodynamics and the free energy principle in biology. *Biology & Philosophy*, *36*, 41. https://doi.org/10.1007/s10539-021-09818-x

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

Gibbs, R. W., Lima, P. L. C., & Francozo, E. (2004). Metaphor is grounded in embodied experience. *Journal of Pragmatics*, *36*, 1189–1210.

Gibson, J. J. (1966). *The senses considered as perceptual systems*. Houghton-Mifflin.

Gibson, J. J. (1975). Events are perceivable but time is not. In J. T. Fraser & N. Lawrence (Eds.), The study of time II (pp. 295–301). Springer-Verlag.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Houghton-Mifflin.

Nonaka, T., & Stoffregen, T. A. (2020). Social interaction in the emergence of toddler's mealtime spoon use. *Developmental Psychobiology*, *62*, 1124–1133. doi:10.1002/dev.21978

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. *Physics of Life Reviews*, *39*, 49–72. https://doi.org/10.1016/j.plrev.2021.09.001

Reed, E. S. (1996). *Encountering the world: Toward an ecological psychology*. Oxford University Press.

Richardson, M. J., Marsh, K. L., & Baron, R. M. (2007). Judging and actualizing intrapersonal and interpersonal affordances. *Journal of Experimental Psychology: Human Perception & Performance*, *33*, 845–859. doi:10.1037/0096-1523.33.4.845

Robbins, S. E. (2014). *The mists of special relativity: Time, consciousness, and a deep illusion in physics*. CreateSpace.

Rosen, R. (1991). *Life itself. A comprehensive inquiry into the nature, origin, and fabrication of life*. Columbia University Press.

Schrodinger, E. (1952a). Are there quantum jumps? Part I. *British Journal for the Philosophy of Science*, *3*, 109–123.

Schrodinger, E. (1952b). Are there quantum jumps? Part II. *British Journal for the Philosophy of Science*, *3*, 233–242. doi:10.1093/bjps/III.11.233

Stoffregen, T. A., & Bardy, B. G. (2001). On specification and the senses. *Behavioral and Brain Sciences*, *24*, 195–261.

Stoffregen, T. A., Mantel, B., & Bardy, B. G. (2017). The senses considered as one perceptual system. *Ecological Psychology*, *29*, 165–197.

Turvey, M. T. (2019). *Lectures on perception: An ecological perspective*. Routledge.

Unnikrishnan, C. S. (2020). A new gravitational paradigm for relativity and dynamics, and its philosophical scope. *Journal of Physics, Conference Series*, *1466*, 012007. doi:10.1088/1742-6596/1466/1/012007

Who tailors the blanket?

Keisuke Suzuki[a], Katsunori Miyahara[a] and Kengo Miyazono[a,b]

[a]Center for Human Nature, Artificial Intelligence, and Neuroscience (CHAIN), Hokkaido University, Sapporo 060-0812, Japan and [b]Department of Philosophy and Religious Studies, Hokkaido University, Sapporo 060-0812, Japan
ksk@chain.hokudai.ac.jp
kmiyahara@chain.hokudai.ac.jp
miyazono@let.hokudai.ac.jp
https://sites.google.com/view/keisukesuzuki/
https://kmiyahara.weebly.com/
http://kengomiyazono.weebly.com/

doi:10.1017/S0140525X22000206, e213

Abstract

The gap between the Markov blanket and ontological boundaries arises from the former's inability to capture the dynamic process through which biological and cognitive agents actively generate their own boundaries with the environment. Active inference in the free-energy principle (FEP) framework presupposes the existence of a Markov blanket, but it is not a process that actively generates the latter.

We endorse the authors' claim that there is a gap between the Markov blanket qua statistical tool and biological and cognitive boundaries qua ontological structures in the world. We will offer an explanation for the gap's existence: It arises from the Markov blanket's inability to capture the dynamic process through which biological and cognitive agents create their own boundaries with the environment over time. Active inference presupposes the existence of a Markov blanket, but it is not envisioned as a process that actively generates the latter.

Biological systems actively produce their boundaries with the environment through autopoietic processes (Varela, 1979). Autopoiesis refers to a network of processes that continually regenerates its components and constructs their own physical boundaries, which Varela and Maturana proposed as the essence of life and its autonomy (Varela, 1979; Varela, Maturana, & Uribe, 1974). For instance, a biological cell maintains its own identity distinct from the environmental medium with a membrane system constructed by a network of metabolic processes. Moreover, living beings maintain their bounded identity not only by exchanging energy and material through metabolism, but also by actively interacting with the environment over space and time. A good illustration is that of a single cell in a nutrient-poor environment that climbs up a glucose gradient to maintain its physical boundary with the environment, keeping its internal states within viable ranges (Egbert & Di Paolo, 2009; Ikegami & Suzuki, 2008; Suzuki &

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

Ikegami, 2009). A biological boundary, then, is not a given, but is actively defined by the system itself. The same principle applies across a wide variety of living systems, from single cell creatures to more complex, multicellular animals, which live sensorimotor lives (Thompson, 2007).

Cognitive systems likewise actively produce their boundaries through their interaction with the environment. We can see this in the case of extended cognition and mind (Clark, 2008; Clark & Chalmers, 1998), where cognitive boundaries extend beyond the biological body by incorporating environmental items as their constitutive parts. Cognitive extension is not a state upon which we stumble by chance; rather, it is a process we actively bring about, or "enact," based on skills and habits cultivated over time (Miyahara & Robertson, 2021; Miyahara, Ransom, & Gallagher, 2020). To illustrate, consider Otto from Clark and Chalmers' (1998) famous thought experiment. Otto suffers a mild case of Alzheimer's disease and uses a notebook to compensate for his memory deficit. According to Clark (2010), Otto and his notebook exhibit a tight functional coupling with each other to constitute a unified cognitive system (Miyazono, 2017) to the extent that they satisfy the following "trust and glue" conditions: (1) the resource (viz., the notebook) is reliably available and typically invoked; (2) any information thus retrieved is more or less automatically endorsed; and (3) information contained in the resource is easily accessible as and when required. Obviously, Otto will not meet these conditions merely by developing a memory problem. Rather, he would have to learn to use notebooks to complement his compromised cognitive capacities and continue to do so repeatedly until it became a habit for him to always carry around a notebook and use it for constant notetaking. The functional coupling is a product of Otto's active engagement with the notebook and his development of relevant skills and habits over time (which is why Otto's Markov blanket is malleable [Clark, 2017] or negotiable [Kirchhoff and Kiverstein, 2021]).

The main shortcoming of the Friston blanket approach concerns the relationship between action (i.e., active inference) and identity (i.e., the Markov blanket). In this approach, active inference depends upon the Markov blanket, but not the other way round. Biological and cognitive systems are defined by Markov blankets as boundaries with the external environment. These systems perform active inferences through looping interactions between sensory states, internal states, and active states defined by the Markov blanket to keep their internal parameters within viable bounds (Friston, 2013). On the other hand, as we saw above, both biological and cognitive systems actively create and maintain their bounded identity by interacting with the environment. As Clark puts it: "Creatures like us […] are Nature's experts at knitting their own Markov blankets" (Clark, 2017, p. 14). To accommodate this within the free-energy principle (FEP) framework, we must conceive of active inference as playing an essential role in autopoiesis, that is, in creating and maintaining the system's bounded identity (cf. Kirchhoff, 2018). In fact, Friston (2010) describes living systems as performing active inference to reduce sensory surprisal and consequently maintain its homeostasis. Nevertheless, on the FEP, active inference does not explicitly participate in the autopoietic formation of the boundary between the system and its environment, which defines the identity of living beings, that is, the Markov blanket. That is, the dynamic relationship between action and identity is missing in the Friston blanket approach that depicts Markov blankets not as a product, but only as a precondition of active inference (Friston, 2013).

In short, the Friston blanket approach fails to identify the tailor who creates the boundaries. At most, Markov blankets coincide with the outcome of the boundary-making processes carried out by biological and cognitive agents. Markov blankets are tailored by statistical patterns but living agents do not outsource boundary-making: We actively weave our own boundaries with the world.

Acknowledgments. We are grateful to Masatoshi Yoshida for his inputs during the preparation stage of this commentary.

Financial support. This work was supported by the JSPS KAKENHI Grant Number 20K00001.

Conflict of interest. None.

References

Clark, A. (2008). *Supersizing the mind*. Oxford University Press.
Clark, A. (2010). Memento's revenge: The extended mind, extended. In R. Menary (Ed.), *The extended mind* (pp. 43–66). MIT Press.
Clark, A. (2017). How to knit your own Markov blanket: Resisting the second law with metamorphic minds. In T. Metzinger & W. Wiese (Eds.), *Philosophy and predictive processing*: 3. MIND Group. https://doi.org/10.15502/9783958573031
Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, *58*(1), 7–19.
Egbert, M. D., & Di Paolo, E. (2009) Integrating autopoiesis and behavior: An exploration in computational chemo-ethology. *Adaptive Behavior*, *17*(5), 387–401.
Friston, K. (2010) The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*, 127–138.
Friston, K. (2013). Life as we know it. *Journal of The Royal Society Interface*, *10*(86), 20130475. http://dx.doi.org/10.1098/rsif.2013.0475
Ikegami, T., & Suzuki, K. (2008). From a homeostatic to a homeodynamic self. *BioSystems*, *91*(2), 388–400.
Kirchhoff, M. D. (2018). Autopoiesis, free energy, and the life–mind continuity thesis. *Synthese*, *195*(6), 2519–2540.
Kirchhoff, M. D., & Kiverstein, J. (2021). How to determine the boundaries of the mind: A Markov blanket proposal. *Synthese*, *198*(5), 4791–4810.
Miyahara, K., Ransom, T. G., & Gallagher, S. (2020). What the situation affords: Habit and heedful interrelations in skilled performance. In F. Caruana & I. Testa (Eds.), *Habit: Pragmatist approaches from cognitive neurosciences to social sciences* (pp. 120–136). Cambridge University Press.
Miyahara, K., & Robertson, I. (2021) The pragmatic intelligence of habits. *Topoi*, *40*, 597–608.
Miyazono, K. (2017). Does functionalism entail extended mind?. *Synthese*, *194*(9), 3523–3541.
Suzuki, K., & Ikegami, T. (2009). Shapes and self-movement in protocell systems. *Artificial Life*, *15*(1), 59–70.
Thompson, E. (2007). *Mind in life: Biology, phenomenology, and the sciences of mind*. Harvard University Press.
Varela, F. R. (1979). *Principles of biological autonomy*. North Holland.
Varela, F. R., Maturana, H. R., & Uribe, R. (1974). Autopoiesis: The organization of living systems, its characterization and a model. *BioSystems*, *5*, 187.

Life, mind, agency: Why Markov blankets fail the test of evolution

Walter Veit[a] and Heather Browning[b]

[a]School of History and Philosophy of Science, The University of Sydney, Sydney, NSW 2006, Australia and [b]London School of Economics and Political Science, Centre for Philosophy of Natural and Social Science, London WC2A 2AE, UK
wrwveit@gmail.com
DrHeatherBrowning@gmail.com
https://walterveit.com/
https://www.heatherbrowning.net/

doi:10.1017/S0140525X22000115, e214

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

Abstract

There has been much criticism of the idea that Friston's free-energy principle can unite the life and mind sciences. Here, we argue that perhaps the greatest problem for the totalizing ambitions of its proponents is a failure to recognize the importance of evolutionary dynamics and to provide a convincing adaptive story relating free-energy minimization to organismal fitness.

In the recent explosion of literature on the free-energy principle, many authors have become increasingly frustrated with the grand ambitions toward using it as a general and unified theory of life, mind, and agency. While many have noted the gulf between the mathematical framework of the free-energy principle and its application to real target systems, in their target article Bruineberg et al. offer what is perhaps the most detailed and sustained criticism of the use of Markov blankets in the biological and cognitive sciences. They argue against what they consider an imprecise use in these sciences for defining entities such as organisms, agents, and minds, differentiating between the theoretical "Pearl blankets" and the more metaphysically laden "Friston blankets." As these two interpretations are often confused and those making metaphysical claims often retreat to an instrumentalist view once pushed, Bruineberg et al. have provided us with a useful tool to distinguish inferences within the model from inferences with a model, which ought not to be done based on the usefulness of the mathematical framework alone. We welcome the challenge to a perceived conflation between the in-principle applicability of the mathematical framework to any self-organizing system and to the conviction that Markov blankets are able to revolutionize our understanding of the living world (Friston, 2013).

The authors note that a realist reading of the application of Friston blankets requires not just the mathematical frameworks established for the use of Pearl blankets, but also independent metaphysical assumptions that, they argue, have not yet been provided. Here, we wish to build on this point by emphasizing the need for these assumptions to align with a plausible Darwinian story. We argue that one of the major problems in recent attempts to use Markov blankets to define the boundaries of organisms and their environments is that they fail to pass the bottleneck of evolutionary theory and give us a misleading picture of living agents and what they are *for*.

Bruineberg et al. show that one cannot just "read off" the boundary between agent and environment from the mathematical formalism provided in the theoretical models. Instead, these are ambiguous and depend on additional assumptions by the modeler, thus requiring quite substantive metaphysical supplementation for Markov blankets to do their work. Here they note that one of the ways of picking out the "right" model for identifying the ontologically significant Friston blanket is through use of the free-energy principle – relying on the assumption that living systems aim at minimizing free energy. It is this basic assumption of the free-energy principle that we wish to challenge. This framework fails to demarcate the organismal boundary that *matters*, from an evolutionary point of view.

As philosophers such as Ruth Millikan and Dan Dennett have long argued, it is only by paying attention to the theoretical bottleneck of evolutionary theory that we can distinguish important properties, boundaries, and processes of living systems between those that *matter* to the organism from those that do not. Markov blankets are said to be able to identify the boundaries of any agent in the sense of a self-organizing system (Ramstead, Kirchhoff, Constant, & Friston, 2019), but they fail to distinguish the right boundaries to understand the evolution of living systems. It has been an oversight within Friston's framework to fail to engage with evolutionary theory and the question of what the organism is *for*. It is only in this *teleonomic* context, that we can make sense of the functional boundaries of life, mind, and agency as properties of biological systems. As the framework fails to answer the hard question of why it is the properties picked out by attempts to apply Markov blankets to biological systems, it cannot succeed in both its explanatory and metaphysical ambitions.

The question that this framework would need to answer in order to be successful in this biological context, is what is the adaptive function of minimizing free energy? That is, how does this process contribute to the survival and reproduction of the organism? One response may be to simply assert that adaptive fitness and negative free energy are "the same thing" (Friston, Thornton, & Clark, 2012, p. 2). However, it is not clear why one should take this to be true – predictive expectations and fitness values do not on their surface appear to constitute anything like the same thing. Another path may be to argue instead that minimization of free energy, while not *constituting* fitness, is still a strong *contributor* to it, in that organisms that act in this way will typically have higher survival and reproduction. However, again, it is not immediately clear why one should believe this. As an example of why this is not particularly plausible, take the *Dark Room Problem*, which offers the challenge that prediction error would be best minimized through sitting still in a dark room, but organisms clearly did not evolve this way (Clark, 2013; Mumford, 1992). If we treat all of the cognitive activities of organisms as a form of prediction or surprise minimization, there will inevitably be "a wedge between what is typical and what is good" (Klein, 2018, p. 2548); we should instead allow that there may be other functions that will not always align with prediction minimization. We then need a more detailed description of the fitness benefits, and how they might be weighted or traded off against other adaptive functions of an organism.

As well as the problems described by the authors of mistaking the useful abstraction Markov blankets provide for the purposes of Bayesian modeling with the idea that free-energy minimization is all that goes on in living systems, we add what is perhaps the greatest problem in the biological context: That it forces us to idealize away from the most important features of living organisms and thus will provide a false and diminished picture of the world. Without the recognition of the importance of evolutionary dynamics, the totalizing ambitions of the free-energy principle to unite the mind and life sciences must fail.

Financial support. W. V.'s research was supported under Australian Research Council's Discovery Projects funding scheme (project number FL170100160).

Conflict of interest. None.

References

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–253.

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

Friston, K. J. (2013). Life as we know it. *Journal of the Royal Society Interface*, *10*(86), 20130475.
Friston, K., Thornton, C., & Clark, A. (2012). Free-energy minimization and the dark-room problem. *Frontiers in Psychology*, *3*, 1–7.
Klein, C. (2018). What do predictive coders want? *Synthese*, *195*(6), 2541–2557.
Mumford, D. (1992). On the computational architecture of the neocortex. *Biological Cybernetics*, *66*(3), 241–251.
Ramstead, M. J., Kirchhoff, M. D., Constant, A., & Friston, K. J. (2019). Multiscale integration: Beyond internalism and externalism. *Synthese*, *198*, 41–70.

Embracing sensorimotor history: Time-synchronous and time-unrolled Markov blankets in the free-energy principle

Nathaniel Virgo[a], Fernando E. Rosas[b,c,d,e] and Martin Biehl[f]

[a]Earth-Life Science Institute (ELSI), Tokyo Institute of Technology, Tokyo 152-8550, Japan; [b]Department of Brain Science, Centre for Psychedelic Research, Imperial College London, London W12 0NN, UK; [c]Data Science Institute, Imperial College London, London W7 2AZ, UK; [d]Centre for Complexity Science, Imperial College London, London W7 2AZ, UK; [e]Department of Informatics, University of Sussex, Brighton BN1 9QJ, UK and [f]Cross Labs, Cross Compass, Tokyo 104-0045, Japan
nathanielvirgo@elsi.jp
f.rosas@imperial.ac.uk
martin.biehl@cross-compass.com
http://www.elsi.jp/en/members/researchers/nvirgo
https://www.imperial.ac.uk/people/f.rosas
https://twitter.com/36zimmer

doi:10.1017/S0140525X22000334, e215

Abstract

The free-energy principle (FEP) builds on an assumption that sensor–motor loops exhibit Markov blankets in stationary state. We argue that there is rarely reason to assume a system's internal and external states are conditionally independent given the sensorimotor states, and often reason to assume otherwise. However, under mild assumptions internal and external states are conditionally independent given the sensorimotor *history*.

Bruineberg and colleagues provide a thorough review of Markov blankets and their limitations in the context of the free-energy principle (FEP). We wish to complement this by drawing attention to two additional issues that we believe have important consequences for the FEP.

Firstly, contrary to what one might expect, the condition known as "Markov blanket" in the FEP literature is generally not guaranteed by a sensor–motor loop structure. Secondly, the Markov blanket condition needed for the FEP is far stronger than it appears to be. These issues severely limit the scope of applicability of current formulations of the FEP. Fortunately, we believe they can be solved, and give some hints towards a resolution.

As Bruineberg et al. explain, the notion of a Markov blanket arises in the context of graphical models, and in particular, Bayesian networks. In a Bayesian network each node represents a random variable, and their joint distribution factors in a particular way that depends on the topology of the graph (Pearl, 1988).

The literature on FEP is also concerned with graphs that are not Bayesian networks. Each node in these graphs represents a dynamical variable of a system and an edge represents the possibility that one dynamical variable can influence another. These include the adjacency matrix described in Bruineberg et al.'s section 4.2, and also the sensor–motor loop as illustrated in their Figure 2. Typically, the edges in such graphs correspond to non-zero terms in a Jacobian matrix. We will call such graphs *influence graphs*.

A stationary state defines a joint distribution over the nodes of an influence graph. There is then some resemblance between the influence graphs and Bayesian networks, since both contain nodes that represent random variables and edges that represent influences of some kind.

However, these two types of graph are fundamentally different. Influence graphs are not necessarily acyclic, but more importantly, the theorems in Pearl's formalism do not apply to influence graphs. In particular, one might expect that the sensor–motor loop (Bruineberg et al.'s Fig. 2) would imply the *time-synchronous Markov blanket condition*

$$\mu_t \perp\!\!\!\perp \phi_t \mid s_t,\ a_t. \tag{1}$$

However, this is not the case in general – and this is important because (1) is used in deriving the FEP. This issue has been recently pointed out (Aguilera, Millidge, Tschantz, & Buckley, 2021; Biehl, Pollock, & Kanai, 2021), and while it has been acknowledged in some of the most recent FEP literature it is not as widely known as it should be. We sketch the underlying reason for it in Figure 1.

Recent works (e.g., Friston, Heins, Ueltzhöffer, Da Costa, and Parr, 2021a; Friston, Da Costa, and Parr, 2021b) have sought to address this by seeking additional conditions or conjectures under which the needed relationship holds. However, the fact that these conditions are highly non-trivial suggests that the scope of the FEP may be much more limited than previously thought.

Furthermore, (1) itself puts a very strong constraint on a system's dynamics. One way to see this is via the *data processing inequality* (Cover & Thomas, 2006, p. 34), which imposes that if (1) holds then all information that μ_t and ϕ_t share needs to be present in $(s_t,\ a_t)$. This would mean that the internal and external states could share no more information than is contained in the sensor and motor states *at the current time*.

But cases where information is stored in the environment and the agent but not in the blanket are ubiquitous. Imagine a friend gives you a phone number written on a piece of paper, which you memorise and then store in a box. The statistical independence between internal and external variables conditioned on active and sensory ones is broken as soon as the piece of paper is away from your sensory input. Once it's out of sight the phone number cannot be stored simultaneously in your internal state and on the piece of paper. As Parr, Da Costa, Heins, Ramstead, and Friston (2021) discuss, this need not be true in transients even if it holds in stationary state. Nevertheless it puts an unrealistic constraint on the stationary dynamics, which we don't expect to be applicable to living organisms.

A possible resolution of this limitation follows from Figure 1. Although (1) cannot be assumed for a general sensor–motor loop,

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

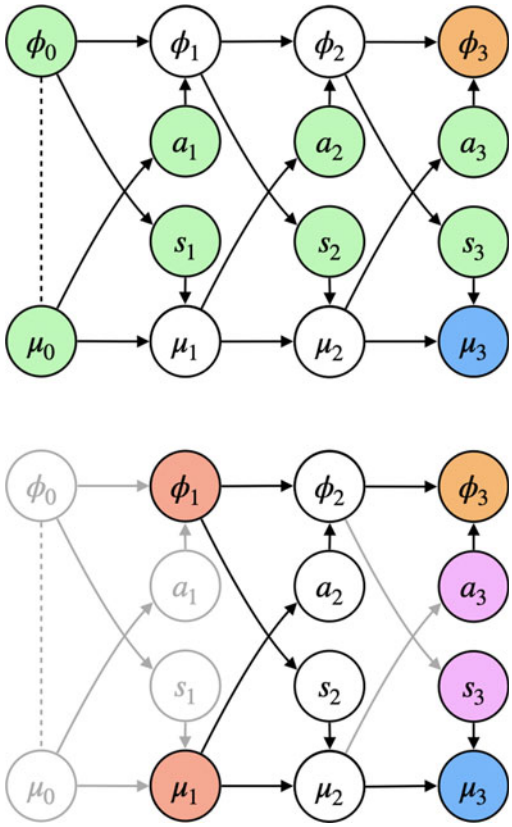


Figure 1 (Virgo et al.). Top: A "time-unrolled" sensor–motor loop in discrete time. The time-unrolled graph is a Bayesian network, and hence the Pearl framework can be applied. It follows that the current internal and external states, μ_3 and ϕ_3 (blue, orange) are conditionally independent given the nodes in light green, which consist of the past *histories* of sensor and actuator states, s_1, s_2, s_3 and a_1, a_2, a_3, as well as the initial states μ_0 and ϕ_0. (The dashed line indicates that the initial states might be correlated.) Bottom: The current internal and external states are in general *not* conditionally independent given only the current sensor and actuator states, s_3 and a_3, because μ_3 and ϕ_3 have common ancestors that are not screened off by these nodes, for example, μ_1 and ϕ_1 (red). This is true regardless of stationarity. If the system is ergodic then the dependence on the initial states will disappear in the infinite limit, so that we can effectively say that μ_t and ϕ_t are conditionally independent given the infinite past history of the sensorimotor states.

we *do* have the relationship

$$\mu_t \perp\!\!\!\perp \phi_t \mid s_t,\ a_t,\ s_{t-1},\ a_{t-1},\ s_{t-2},\ a_{t-2},\ \ldots. \tag{2}$$

We expect an analogous result in continuous time. The internal and external states are not conditionally independent given the *current* sensorimotor states, but, under only mild assumptions, they are conditionally independent given the sensorimotor *history*. Alternative constructions of blankets that follow these principles are currently being investigated (e.g., Rosas, Mediano, Biehl, Chandaria, and Polani, 2020).

This makes intuitive sense: Your knowledge of the world is not limited by what you can sense at the current moment, but it is limited by what you have been able to sense over your whole lifetime. If a new version of the FEP can be constructed based on this alternative conditional independence relation then it will be more encompassing and will have something close to the broad applicability that was originally intended.

Financial support. M. B. and N. V. acknowledge the support of Grant 62229 from the John Templeton Foundation. The opinions expressed in this publication are those of the author(s) and do not necessarily reflect the views of the John Templeton Foundation. F. R. is supported by the Ad Astra Chandaria Foundation.

Conflict of interest. None.

References

Aguilera, M., Millidge, B., Tschantz, A., & Buckley, C. L. (2021). How particular is the physics of the free energy principle? *Physics of Life Reviews*, *40*, 24–40.

Biehl, M., Pollock, F. A., & Kanai, R. (2021). A technical critique of some parts of the free energy principle. *Entropy*, *23*(3), 293.

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory* (2nd ed.). Wiley.

Friston, K., Heins, C., Ueltzhöffer, K., Da Costa, L., & Parr, T. (2021a). Stochastic chaos and Markov blankets. *Entropy*, *23*(9), 1220.

Friston, K. J., Da Costa, L., & Parr, T. (2021b). Some interesting observations on the free energy principle. *Entropy*, *23*(8), 1076.

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.

Rosas, F. E., Mediano, P. A., Biehl, M., Chandaria, S., & Polani, D. (2020). Causal blankets: Theory and algorithmic framework. In T. Verbelen, P. Lanillos, C. L. Buckley, & C. De Boom (Eds.), *International workshop on active inference* (pp. 187–198). Springer.

Does the metaphysical dog wag its formal tail? The free-energy principle and philosophical debates about life, mind, and matter

Wanja Wiese

Institute of Philosophy II, Ruhr University Bochum, D-44780 Bochum, Germany
wanja.wiese@rub.de
https://homepage.ruhr-uni-bochum.de/wanja.wiese/

doi:10.1017/S0140525X22000292, e216

Abstract

According to Bruineberg and colleagues, philosophical arguments on life, mind, and matter that are based on the free-energy principle (FEP) (1) essentially draw on the Markov blanket construct and (2) tend to assume that strong metaphysical claims can be justified on the basis of metaphysically innocuous formal assumptions provided by FEP. I argue against both (1) and (2).

Bruineberg et al. distinguish between Markov blankets (MBs) as properties of models (epistemic "Pearl blankets," PBs) and MBs as properties of physical systems (metaphysical "Friston blankets," FBs). Unlike PBs, FBs are not exclusively defined in terms of probabilistic relations between random variables. Instead, FBs are also defined in terms of causal relations to a system's internal and external states: FBs comprise sensory and active states that mediate causal influences between internal and external states. In perceiving agents, this causal interplay between internal, external, and blanket states is supposed to capture perception-action loops (Friston, 2012).

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

As Bruineberg et al.'s lucid analysis shows, a realist interpretation of the notion of an FB is metaphysically demanding: Positing the existence of an FB entails a statement about the physical system, that is, it entails the existence of a boundary between internal and external states. Furthermore, the authors argue that the formalism that is used to identify internal, active, sensory, and external states (as in Friston, 2013) presupposes specific, non-arbitrary assumptions about the underlying model (target paper, sect. 4.1). This means that FBs are not *theoretically neutral* descriptions of a model's features. Furthermore, many descriptions of FBs in the literature are not *metaphysically neutral*, either, because they describe these blankets not just as properties of a model, but as boundaries of physical systems. As the authors point out, this is problematic to the extent that such metaphysical interpretations are assumed to "follow from the formal details" (target paper, sect. 4.2).

In the context of the free-energy principle (FEP), the notion of an MB is not just used to describe perception-action loops in certain self-organising systems. The distinction between internal, blanket, and external states is also invoked to describe the system's internal states as representations of probability distributions (of external states, given blanket states). According to FEP, this enables a dual description of the system's dynamics: On the one hand, in terms of the probabilistic evolution of internal states; on the other hand, in terms of the evolution of probabilistic beliefs about external states, parameterised by internal states (see Friston, 2019; Friston, Wiese, & Hobson, 2020; see also Kiefer, 2020; Sprevak, 2020).

Apart from claims about the boundaries of minds, most philosophical claims about the relationships between life, mind, and matter do not essentially draw on the MB construct provided by FEP. Contrary to what Bruineberg et al. suggest, the main philosophically fruitful idea afforded by FEP is that internal states can be regarded as probability distributions over external states, given sensory and active states (see the examples given below).

One could object that this presupposes the existence of a boundary between a system and its environment – which, as the authors argue, cannot be derived from FEP, without presupposing strong metaphysical assumptions. We can grant this for the sake of argument, because the assumption that there is a boundary between a system and its environment is shared by many metaphysical theories of the mind (e.g., functionalism and, arguably, most versions of property physicalism), and none of these theories shows how to derive this partition from metaphysically innocuous assumptions. As long as FEP-based philosophical accounts avoid the mistake of assuming that FEP provides a metaphysically neutral way of determining boundaries, it is therefore not particularly problematic to posit the existence of a boundary between internal and external states (i.e., if it is problematic, it is problematic independently of FEP).

Bruineberg et al. argue that claims about the relationships between life, mind, and matter must make further metaphysical assumptions that do not follow from FEP. The authors worry that these assumptions "may in the end be doing all of the interesting work themselves" (target paper, sect. 7). I disagree with this statement, but here I will focus on the concern, expressed by the authors, that researchers mistakenly believe the FEP can be used to "settle fundamental metaphysical questions" (target paper, sect. 1).

In contrast to what Bruineberg et al. suggest, the fact that additional assumptions are needed to contribute to metaphysical debates is acknowledged in FEP-based arguments for philosophical claims. Let me illustrate this with three examples (involving my own work).

In Friston et al. (2020), we argue that the most parsimonious interpretation of FEP's dual description of internal states is a form of property physicalism (reductive materialism) that we call *Markovian monism*. Crucially, we do not assume that this interpretation follows from the formalism itself: We consider and evaluate different metaphysical interpretations of the formalism (Friston et al., 2020, pp. 17–21).

Wiese and Friston (2021a) argue that FEP is compatible with a strong continuity between life and mind. Again, this is not assumed to follow directly from the formalism. Rather, we explicitly point out that we presuppose a mechanistic account of physical computation and a representationalist interpretation of predictive processing/active inference (Wiese & Friston, 2021a, pp. 10–13).

In Wiese and Friston (2021b), we draw on the duality between the probabilistic evolution of internal states and the evolution of probabilistic beliefs (entailed by FEP) to argue: A self-organising system is a conscious system (as opposed to a mere simulation of a conscious system) only if computational correlates of consciousness (CCCs) are instantiated by the physical processes that help the system sustain its existence (for details, see Wiese & Friston, 2021b, pp. 22–24). FEP is here primarily invoked to provide a precise description of how a conscious system differs from a mere simulation: The dynamics of CCCs (in terms of probabilistic beliefs encoded by internal states) must be equivalent to the dynamics of internal states in terms of the system's non-equilibrium steady-state density.

In sum, I disagree with the authors' claim that FEP-based philosophical debates about the relationships between life, mind, and matter presuppose that FEP "can be used to settle fundamental metaphysical questions" (target article, sect. 1). Does this mean that the metaphysical heavy lifting is not done by FEP itself, but by additional metaphysical assumptions? I don't think so: Philosophical arguments require clear concepts, and this is exactly what FEP's dual description of system dynamics affords.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Conflict of interest. None.

References

Friston, K. (2012). A free energy principle for biological systems. *Entropy*, *14*(11), 2100–2121. https://doi.org/10.3390/e14112100

Friston, K. (2013). Life as we know it. *Journal of the Royal Society Interface*, *10*(86), 1–12. https://doi.org/10.1098/rsif.2013.0475

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

Friston, K. J., Wiese, W., & Hobson, J. A. (2020). Sentience and the origins of consciousness: From Cartesian duality to Markovian monism. *Entropy*, *22*(5), 516. https://doi.org/10.3390/e22050516

Kiefer, A. B. (2020). Psychophysical identity and free energy. *Journal of the Royal Society Interface*, *17*(169), 20200370. https://doi.org/10.1098/rsif.2020.0370

Sprevak, M. (2020). Two kinds of information processing in cognition. *Review of Philosophy and Psychology*, *11*(3), 591–611. https://doi.org/10.1007/s13164-019-00438-9

Wiese, W., & Friston, K. J. (2021a). Examining the continuity between life and mind: Is there a continuity between autopoietic intentionality and representationality? *Philosophies*, *6*(11), 18. https://doi.org/10.3390/philosophies6010018

Wiese, W., & Friston, K. J. (2021b). The neural correlates of consciousness under the free energy principle: From computational correlates to computational explanation. *Philosophy and the Mind Sciences*, *2*, 9. https://doi.org/10.33735/phimisci.2021.81

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

Markov blankets: Realism and our ontological commitments

Danielle J. Williams

Department of Philosophy, University of California at Davis, Davis, CA 95616, USA
heywilliams@ucdavis.edu
daniellejwilliams.com

doi:10.1017/S0140525X22000255, e217

Abstract

The authors argue that their target is orthogonal to the realism and instrumentalist debate. I argue that it is born directly from it. While the distinction is helpful in illuminating how some ontological commitments demand a theory of implementation, it's less clear whether different views cleanly map onto the epistemic and metaphysical uses defined in the paper.

Bruineberg and colleagues argue there is a conflation between two uses of Markov blankets. Some use Markov blankets in an epistemic way while others use them to make ontological claims about the physical world. To solve this conflation, they propose that we should classify the former as Pearl blankets and the latter as Friston blankets. While this strategy provides a helpful labeling scheme for different uses, a need for a distinction of this kind is indicative of a more substantial problem. Thus, solving this conflation targets a symptom of a broader problem rather than targeting what is at issue in the first place. The authors note that their discussion is orthogonal to the realism and instrumentalism debate in cognitive science, but I argue that their distinction is better understood as a case study born directly from this debate. Computational models play different roles in our scientific theories. We can understand them as purely formal, or we can take them as literally representing physical systems. But, regardless of our position, we need to say something about how our formal, non-physical models relate to the concrete, physical world.

Pearl blankets are Markov blankets used in the formal sense while Friston blankets are taken to be or to genuinely represent concrete boundaries. This distinction rests on how scientists use Markov blankets in their theorizing. But distinguishing between uses leads to a question of how we should frame the difference between Pearl and Friston blankets as scientific posits, not just how they are used within a theory. We could understand the distinction most straightforwardly as delineating between the formal and the physical. One way to cash this out is by thinking about Markov blankets at either the algorithmic level or the implementation level within the Marrian framework. Pearl blankets are purely formal models at the algorithmic level deployed irrespective of the nuts and bolts of the physical system while Friston blankets are implementations of Markov blankets themselves. Because realism proposes that our best scientific theories provide us with knowledge of the objective world – which ontologically commits us to the entities they posit – Markov blankets understood at the implementation level are a bona fide example of a realist position while Markov blankets understood at the algorithmic level and deployed in the Pearl sense demonstrate an instrumentalist position. Because of this, the distinction is not orthogonal to the realism and instrumentalist debate: it's a case study within it.

The authors argue that Friston users have an additional explanatory task because we can't simply read our ontology off of the mathematics. What is needed is an explanation of how a formal construct can be understood in a such metaphysically laden way. This is exactly correct: To complete the theory an account of implementation is required. What is needed for proper reification is an account that maps the formal mathematical model to the boundaries of the physical world. While it is still an open question how we should formulate the implementation relation, there are some views that could be adopted. One approach is to argue that there must be some resemblance between the model and the target system such that some specified features are necessarily consistent between the two (Curtis-Trudel, 2021). Resemblance may help to alleviate some conceptual issues regarding irregular boundaries. Another viable option comes from Bogacz (2015). Bogacz proposes a theory of implementation that maps different elements of the model onto different neural populations within the cortex where the mapping between the variables in the model and the elements of the neural circuitry may not be "clean" but rather "messy" (Bogacz, p. 209). Different views will map the formal computation onto the physical world in different ways, but what is important is that the relation between the formal model and the physical world is accounted for.

One worry, though, is that the distinction between Pearl blankets and Friston blankets is overly restrictive. There are additional ways to understand how Markov blankets are used over and above the Pearl and Friston senses. For example, one might be a realist without being committed to physical implementation: It is possible to have ontological commitments to mathematical entities at Marr's algorithmic level without ontologically committing oneself to implementation level features. Scientific realism proposes that we are ontologically committed to the existence of the posits that do explanatory work in our best scientific theories. Depending on your view of explanation, non-causal, formal properties can play a robust explanatory role that meets the criterion for scientific realism (Williams & Drayson, forthcoming). This goes beyond the epistemic use and stops just short of the metaphysical use blurring the distinction between Pearl and Friston blankets by neglecting to carve out space for a mathematical ontology. If one can hold ontological commitments about formal entities, do they also have an additional explanatory debt? Do they now count as Friston blankets? Because you can have ontological commitments at both the formal and physical levels, the distinction between Pearl and Friston uses blurs and additional explanatory requirements become unclear.

Different uses of Markov blankets provide a case study within the instrumentalism and realism debate in cognitive science. Some accept the formal model as an epistemic tool while others use the formal model to make ontological claims. As with all formal models, for proper reification, some account of implementation is needed. But, once the distinction is considered within the context of the realism and instrumentalism debate in which it belongs, it become unclear that the distinction is able to do the work that it sets out to do in the first place because it fails to leave room for additional ways in which one can take on a realist stance about formal models.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Conflict of interest. None.

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

References

Bogacz, R. (2015). A tutorial on the free-energy framework for modelling perception and learning. *Journal of Mathematics and Psychology*, *76*, 198–122.

Curtis-Trudel, A. (2021). Implementation as resemblance. *Philosophy of Science*, *88*, 1021–1032.

Williams, D., & Drayson, Z. (forthcoming). The nature of the predictive mind: Realism and instrumentalism in Bayesian cognitive science. In T. Cheng, R. Sato & J. Hohwy (eds.) *Expected experiences: The predictive mind in an uncertain world*. Routledge.

Causal surgery under a Markov blanket

Daniel Yon[a] and Philip Robert Corlett[b]

[a]Department of Psychological Sciences, Birkbeck, University of London, London WC1E 7HX, UK and [b]Department of Psychiatry, Connecticut Mental Health Center, Yale University, New Haven, CT 06519, USA
d.yon@bbk.ac.uk
philip.corlett@yale.edu
www.danielyon.com
https://medicine.yale.edu/lab/corlett/

doi:10.1017/S0140525X22000218, e218

Abstract

Bruineberg et al. provide compelling clarity on the roles Markov blankets could (and perhaps should) play in the study of life and mind. However, here we draw attention to a further role blankets might play: as a hypothesis about cognition itself. People and other animals may use blanket-like representations to model the boundary between themselves and their worlds.

In their impressive target article Bruineberg et al. describe two radically different ways we can use Markov blankets. *Pearl blankets* are tools that allow scientists to identify (in)dependence between variables when modelling complex systems. In contrast, *Friston blankets* are tools philosophers may use to parse the world into internal and external states, distinguishing agents from the rest of their worlds.

We wholeheartedly agree that this distinction is important, but feel this dichotomy neglects a third possibility: blankets as a hypothesis about cognition itself. In this way of thinking, cognising creatures may use processes that approximate Bayesian modelling to track which states of the world depend on or are independent of their actions. In so doing, these creatures construct a *cognitive blanket* that captures their beliefs about what they can and cannot control.

This *cognitive blanket* hypothesis makes distinctive predictions about how agents estimate agency and control over their bodies and the world. Many have suggested that humans and other animals determine what they can control by tracking correlations between actions and outcomes (Dickinson & Balleine, 1994; Yon, Bunce, & Press, 2020). However, building a cognitive blanket – mapping causal dependencies between actions and states – allows an agent to entertain counterfactual scenarios and to intervene on the world to test connections implied by their model. This kind of hypothesis testing – evocatively dubbed "causal surgery" (Pearl, Glymour, & Jewell, 2016) – allows agents to refine beliefs about their own causal power by acting on the world in informative ways.

Psychologists can test for *cognitive blankets* by investigating whether agents are sensitive to counterfactual information and engage in "causal surgery" to test what they can and cannot control. In humans, there is some evidence of sensitivity to counterfactual information – we feel a greater sense of control when we believe we could have acted differently and this could have altered outcomes (Kulakova, Khalighinejad, & Haggard, 2017). There is also tentative evidence that human agents perform exploratory actions when judging control over events in the external world (Wen et al., 2020). This kind of exploration could be a hallmark of "causal surgery" that tests hypotheses about our influence. However, it is also possible that apparently exploratory behaviour emerges from noise in decision and action systems (Findling, Skvortsova, Dromnelle, Palminteri, & Wyart, 2019). Targeted tests are thus needed to establish whether humans engage in genuine causal surgery when estimating control – possibly by determining whether explorations about control depend on the agent's uncertainty about action–outcome relationships.

The same tests could also be applied by comparative cognitive scientists. It has long been debated how far nonhuman animals represent their behaviour as "causes" of environmental changes (Penn & Povinelli, 2007). In our way of thinking, empirical evidence of causal surgery in different species would suggest the animal is constructing a *cognitive blanket* – testing hypotheses about how action and outcome connect. As with humans, it would be important to distinguish uncertainty-driven hypothesis testing in animals from blind exploration. Such efforts could exploit apparent signatures of "confidence" detectable in animals (Kepecs, Uchida, Zariwala, & Mainen, 2008), or could investigate how animals respond to different varieties of environmental uncertainty (Yon & Frith, 2021). For instance, if a creature's exploratory behaviour responds to volatility in action–outcome relationships, this may be indicative of causal surgery: The active probing of the agent's blanket-like model to test what they can and cannot influence.

Furthermore, *cognitive blankets* could illuminate the disturbances of action awareness that occur in psychiatric illness. Patients with psychosis often develop delusions about action and control: They claim to control things they objectively cannot (grandiosity) and deny controlling some actions they have genuinely authored (passivity; Frith, Blakemore, and Wolpert, 2000). These strange beliefs might arise from a disordered blanket that draws the boundary between world and agent in an unusual way (much like that depicted in Bruineberg et al.'s Fig. 7c). If intervention and exploration are essential ingredients in building up an accurate *cognitive blanket*, it may be fruitful for clinical scientists to investigate processes of causal surgery in psychosis. If these patients are less likely to intervene on the world to test what they can control, unusual beliefs about the self and the world may persist unchecked. Indeed, one could speculate that a vicious cycle obtains in psychosis, where negative symptoms dampening the drive to act (e.g., apathy, catatonia) rob patients of action–outcome experiences that could challenge positive symptoms (i.e., delusions about action; see Bortolotti & Broome, 2012; Corlett, Honey, & Fletcher, 2016). We note with interest the role that dopamine signalling appears to play in learning, confidence, causal inference, and their derangement in psychosis-like states (Redgrave & Gurney, 2006; Schmack, Bosc, Ott, Sturgill, & Kepecs, 2021; Sharpe et al., 2017).

Our third way of thinking about blankets – as representations in the heads of agents – departs from both *Friston* and *Pearl*

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

blankets. Even if *Friston blankets* cannot pick out the objective boundaries between agents and their worlds, blanket-like computations may still be the processes by which some creatures (suboptimally) identify where these boundaries lie.

Moreover, the ability of scientists to model an agent using *Pearl blankets* cannot tell us whether the agent uses a *cognitive blanket* to model itself. For example, creatures that do not build a *cognitive blanket* may rely on simple, lean psychological processes like associative learning to navigate their environments (i.e., without building a causal graph). However, these simpler psychological processes are only adaptive because they also gear creatures into the causal structure of their environments (Papineau & Heyes, 2006). Simple creatures may thus be well modelled by *Pearl blankets*, even if they do not have a *cognitive blanket* of their own.

In conclusion, Bruineberg et al. provide an important perspective on how scientists and philosophers should use Markov blankets to describe the boundaries between agents and their worlds. However, it is also important to consider (and test) whether blanket-like representations are at the heart of how agents construct these boundaries in their own minds.

Financial support. The authors received no specific funding to support this work.

Conflict of interest. None.

References

Bortolotti, L., & Broome, M. R. (2012). Affective dimensions of the phenomenon of double bookkeeping in delusions. *Emotion Review*, *4*(2), 187–191. https://doi.org/10.1177/1754073911430115

Corlett, P. R., Honey, G. D., & Fletcher, P. C. (2016). Prediction error, ketamine and psychosis: An updated model. *Journal of Psychopharmacology*, *30*(11), 1145–1155. https://doi.org/10.1177/0269881116650087

Dickinson, A., & Balleine, B. (1994). Motivational control of goal-directed action. *Animal Learning & Behavior*, *22*(1), 1–18. https://doi.org/10.3758/BF03199951

Findling, C., Skvortsova, V., Dromnelle, R., Palminteri, S., & Wyart, V. (2019). Computational noise in reward-guided learning drives behavioral variability in volatile environments. *Nature Neuroscience*, *22*(12), 2066–2077. https://doi.org/10.1038/s41593-019-0518-9

Frith, C. D., Blakemore, S. J., & Wolpert, D. M. (2000). Abnormalities in the awareness and control of action. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *355*(1404), 1771–1788. https://doi.org/10.1098/rstb.2000.0734

Kepecs, A., Uchida, N., Zariwala, H. A., & Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature*, *455*(7210), 227–231. https://doi.org/10.1038/nature07200

Kulakova, E., Khalighinejad, N., & Haggard, P. (2017). I could have done otherwise: Availability of counterfactual comparisons informs the sense of agency. *Consciousness and Cognition*, *49*, 237–244. https://doi.org/10.1016/j.concog.2017.01.013

Papineau, D., & Heyes, C. (2006). Rational or associative? Imitation in Japanese quail. In M. Nudds & S. Hurley (Eds.), *Rational animals* (pp. 187–195). Oxford University Press.

Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: A primer*. Wiley.

Penn, D. C., & Povinelli, D. J. (2007). Causal cognition in human and nonhuman animals: A comparative, critical review. *Annual Review of Psychology*, *58*, 97–118. https://doi.org/10.1146/annurev.psych.58.110405.085555

Redgrave, P., & Gurney, K. (2006). The short-latency dopamine signal: A role in discovering novel actions? *Nature Reviews. Neuroscience*, *7*(12), 967–975. https://doi.org/10.1038/nrn2022

Schmack, K., Bosc, M., Ott, T., Sturgill, J. F., & Kepecs, A. (2021). Striatal dopamine mediates hallucination-like perception in mice. *Science* (*New York, N.Y.*), *372*(6537), eabf4740. https://doi.org/10.1126/science.abf4740

Sharpe, M. J., Chang, C. Y., Liu, M. A., Batchelor, H. M., Mueller, L. E., Jones, J. L., … Schoenbaum, G. (2017). Dopamine transients are sufficient and necessary for acquisition of model-based associations. *Nature Neuroscience*, *20*(5), 735–742. https://doi.org/10.1038/nn.4538

Wen, W., Shibata, H., Ohata, R., Yamashita, A., Asama, H., & Imamizu, H. (2020). The active sensing of control difference. *iScience*, *23*(5), 101112. https://doi.org/10.1016/j.isci.2020.101112

Yon, D., Bunce, C., & Press, C. (2020). Illusions of control without delusions of grandeur. *Cognition*, *205*, 104429. https://doi.org/10.1016/j.cognition.2020.104429

Yon, D., & Frith, C. D. (2021). Precision and the Bayesian brain. *Current Biology*, *31*(17), R1026–R1032. https://doi.org/10.1016/j.cub.2021.07.044

Authors' Response

The Emperor Is Naked: Replies to commentaries on the target article

Jelle Bruineberg[a], Krzysztof Dołęga[b], Joe Dewhurst[c] and Manuel Baltieri[d,e]

[a]Department of Philosophy, Macquarie University, Sydney, NSW 2109, Australia; [b]Institut für Philosophie II, Fakultät für Philosophie und Erziehungswissenschaft, Ruhr-Universität Bochum, 44801 Bochum, Germany; [c]Fakultät für Philosophie, Wissenschaftstheorieund Religionswissenschaft, Munich Center for Mathematical Philosophy, Ludwig-Maximilians-Universität München, 80539 Munich, Germany; [d]Araya, Inc., Tokyo, Japan and [e]School of Engineering and Informatics, University of Sussex, Brighton BN1 9RH, UK
jelle.bruineberg@mq.edu.au
krzysztof.dolega@rub.de
joseph.e.dewhurst@gmail.com
manuel_baltieri@araya.org

doi:10.1017/S0140525X22000656, e219

Abstract

The 35 commentaries cover a wide range of topics and take many different stances on the issues explored by the target article. We have organised our response to the commentaries around three central questions: Are Friston blankets just Pearl blankets? What ontological and metaphysical commitments are implied by the use of Friston blankets? What kind of explanatory work are Friston blankets capable of? We conclude our reply with a short critical reflection on the indiscriminate use of both Markov blankets and the free energy principle.

R1. Introduction

> "What's this?" thought the Emperor. "I can't see anything. This is terrible! Am I a fool? Am I unfit to be the Emperor? What a thing to happen to me of all people! – Oh! It's very pretty," he said. "It has my highest approval." And he nodded approbation at the empty loom. Nothing could make him say that he couldn't see anything.

In H. C. Andersen's folktale *The Emperor's New Clothes*, two swindlers convince the king to let them make him a set of special clothes that is invisible to anyone who is either too stupid or incompetent. One by one, the Emperor's advisors go and check with the swindlers and, afraid of being thought a fool, pretend they see the excellent patterns and beautiful colours of the new garments. When the Emperor goes out to show the new robe to his citizens, it is only an innocent child who bursts out: "But he hasn't got anything on" – and even then, the Emperor proudly continues his procession (Anderson, 1837).

Both **Raja, Baggs, Chemero, and Anderson (Raja et al.)** and **Hesp** ask what part of the folktale we are alluding to with the title of our paper. Raja et al. had expected us to "follow the child's

lead" and expose Friston blankets as indeed being nothing more than the Emperor's New Clothes, while Hesp asks us to clarify whether we intended to attribute intentional deceit to free energy principle (FEP) researchers. Let us start by clarifying our position here, which will provide the leitmotif for the rest of our response. First of all, we would like to be clear that our criticism is levelled against the content of academic proposals and the way certain concepts are used, not against the people themselves. To stay with the metaphor, our target article focused on the putative patterns and colours of the clothes, not on the Emperor and his courtiers. Second, there is an important distinction between intentional deceit and what Frankfurt (2005) refers to with the technical term *bullshitting*. Deceit requires knowing the truth and intentionally hiding it from others, while bullshitting involves attempting to persuade without taking truth into consideration. Importantly, bullshitting can be done unintentionally. One way in which bullshitting might happen unintentionally is when speculations are underconstrained, as is sometimes the case when mathematical constructs are applied outside of their original formal context. In such scenarios, the argumentative structure resembles that of ordinary science, but fails to engage meaningfully with the underlying formal constructs. The result is armchair philosophy dressed up as an empirical research programme, which **Allen** goes so far as to label as "cargo cult" science.

Third, the FEP research programme has been an ambitious, daring, and speculative endeavour from the start. This is what attracted each of the authors of the target article to work on the FEP in the first place: the potential for a formalisation of big philosophical concepts like agency, intentionality, life, and consciousness. But in order to make this work, it is necessary to intricately weave together mathematical modelling, philosophy of science, philosophy of mind, and empirical research. This interdisciplinarity, along with the rapid pace at which papers on the FEP are published, introduces vulnerabilities: it is a difficult literature to keep up with, and even experts can sometimes be confronted with things that they do not fully understand. It was this feeling of epistemic dissonance, central to the Emperor's New Clothes, that we wanted to allude to with our title. Andersen's folktale captures a common sentiment among those getting interested in the FEP: am I too incompetent to understand this, are the conclusions that are being drawn not supported by the theory?

Our aim for the target article was to critically assess the current state of both philosophical and scientific literature using Markov blankets within the FEP: what are the assumptions required to transform Markov blankets from a technical notion in statistics and probability theory to an ontological notion used to settle philosophical and scientific disputes? In other words, can we at least agree on what kind of clothes the Emperor is supposed to be wearing?

In analysing the literature, we found a number of technical slippages and conceptual unclarities: an ambiguity between Markov (Pearl) blankets as modelling tools and Markov (Friston) blankets as real-world boundaries, an ambiguity between instrumentalist, realist, and literalist statements about the latter Friston blanket construct, and a conflation of different types of explanatory project that might use either kind of blanket. We hoped that the range of options we presented in the target article could serve as an open invitation for those working with the FEP to make clear what their commitments really are, and how those commitments can support the claims they make.

We are extremely grateful that this challenge has been taken up by a great number of commentators, from both inside and outside the literature on the FEP. We would like to thank the authors of all 35 commentaries for their insightful critiques and friendly suggestions. The commentaries cover a wide range of topics and take many different stances: some push back against our framing of the problems surrounding Markov blankets and the FEP, some suggest ways to fix the problems we identified, and yet others highlight a number of additional problems. In Table R1 we present a thematic overview of the commentaries as we understood them.

The range of attitudes adopted in the commentaries shows just how widespread disagreement and confusion is in the FEP literature about the conceptual, metaphysical, and methodological commitments implied by the use of Markov blankets. Crucially, the issues we raised in the target article are interrelated: you can only claim that the structures identified in your model carve out real boundaries in the world if the mapping between model and target is structure preserving (see **Parr**). The usefulness of instrumentally treating organisms *as if* they are models of their environments does not imply that it will also be useful to treat them *as if* their physical boundaries emerge within this model. The proven utility of Markov blankets as modelling tools in the probabilistic inference literature does not directly support the claim that to *be* a thing is to *have* a Markov blanket. There is a thin line between a scientific endeavour whose metaphors are pushed just a bit too far, and a heavy-duty metaphysical theory dressed up as an empirical research programme. To better capture the diversity of views on offer we have divided our reply into several, mutually supporting sections, with a number of commentaries being mentioned in multiple sections.

R2. Are Friston blankets just Pearl blankets?

The first major issue we would like to tackle in our reply to the commentaries is the distinction (or lack thereof) between Pearl and Friston blankets. One of the major contributions of the target article was to make explicit the previously unacknowledged shift from Pearl blankets to Friston blankets. A number of commentaries argue that the two kinds of entities are, in fact, cut from the same cloth, and that our distinction just captures two different ways of referring to the same Markov blanket formalism.

Parr and **Friston** point out that a Markov [Pearl] blanket is defined in terms of conditional probabilities and, therefore, can only be delineated in the context of a model described in terms of probability distributions. Whether the distributions involved will be sourced from a steady-state density of some dynamical system, or just some static model (i.e., one not evolving over time) should not make a difference for the identification of a Markov [Pearl] blanket. **Ramstead** claims that Friston blankets simply *are* Markov blankets, that is, that Friston blankets and Pearl blankets are the same abstract mathematical objects denoting conditional independence, deployed in different modelling contexts. **Hesp** concurs and points out that much of the confusion we identified stems from differing background assumptions about the kinds of causal relationships and types of systems that Friston and others are trying to study. This is echoed by **Kiverstein and Kirchhoff** who write that "Friston blankets are interpretations given of the Markov blankets formalism in the context of the FEP that purport to describe autopoietic processes." But are Friston blankets and Pearl blankets really one and the same thing?

Our answer is a resounding no, and we also deny that there is any straightforward route to derive Friston blankets from Pearl blankets. The difference between Pearl blankets and Friston blankets is not simply due to differing causal commitments, as

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

Table R1. Thematic overview of commentaries

Theme	Focus	Commentaries
Formalism	Markov blanket formalism	Virgo, Rosas, and Biehl; Friston; Aguilera and Buckley
	Free energy and physics	Spector and Graham; Stoffregen and Heath; Ramstead
	Spatial boundaries	Parr
Philosophy of science	Realism, instrumentalism, and literalism	Sánchez-Cañizares; Kiefer and Hohwy; Williams; Kiverstein and Kirchhoff; Ramstead; Rorot, Korbak, Litwin, and Miłkowski; Hipólito and van Es; Friston; Colombo; Seth, Korbak, and Tschantz; Wiese; Menary and Gillett
	Causality and interventions	Btesh, Bramley, and Lagnado; Yon and Corlett
	Reification	Andrews
	Unification	Gomez-Marin
	Models and abstraction	Nave; Spiegel; Ciaunica; Beck
Philosophy of mind and life	Internalism and externalism	Facchin; Menary and Gillett
	Autopoiesis	Nave; Suzuki, Miyahara, and Miyazono; Raja, Baggs, Chemero, and Anderson; Kiefer and Hohwy; Dengsø, Robertson, and Constant
	Ecological psychology	Stoffregen and Heath; Raja, Baggs, Chemero, and Anderson
	Agents and selves	Sprevak; Colling; Seth, Korbak and Tschantz; Yon and Corlett; Wiese
Interdisciplinary	Boundary objects	Fox; Allen
	Evolution	Veit and Browning

suggested by **Btesh, Bramley, and Lagnado** (**Btesh et al.**), since the use of Markov blankets (Pearl or Friston) is not tied to models interpreted under some causal semantics (Pearl, 2009). Similarly, commentaries such as **Hesp**, **Kiverstein and Kirchhoff**, and **Wiese** that either invoke, or seem to imply the necessity of, causality in Bayesian models, fall short of explaining where this causal interpretation comes from in the first place. Furthermore, Friston blankets are not just Pearl blankets transformed by technical assumptions (cf. Biehl, Pollock, & Kanai, 2021), or buttressed with additional philosophical commitments. In his commentary, **Friston** correctly argues that Markov (i.e., Pearl) blankets are merely statements about conditional independence, such that for generic Markov chains "the 'present' is a Markov [Pearl] blanket that separates the 'past' from the 'future'." However, this is apparently not what a Friston blanket is, as Friston himself (Friston, Da Costa, & Parr, 2021a) writes:

> "For Markovian systems, the states at the current time are the blanket states that separate states in the future from states in the past. However, these are not Markov [Pearl] blankets of the steady-state density."

This tells us that the past and future of a random variable, which would be part of a "naive" Pearl blanket in dynamic settings (Pearl, Geiger, & Verma, 1989), are not part of a Friston blanket. To really get to the core definition of what a Friston blanket is, we can further look at a reply to the example found in Biehl et al. (2021) that describes their mathematical assumptions. Friston et al. (2021a) once again helpfully explains that

> "[…]; the FEP only applies to Markov blankets that emerge under sparse flows; in particular, when autonomous states are uncoupled from external states (by definition)."

Here, "sparse flows" refers to a specific coupling structure required between partitions (internal, external, active, and sensory states) in dynamical (i.e., time-evolving) settings (see equation [7] in Friston et al. [2021a] or equation [12] in Friston et al. [2022]). More recently this also goes by the name of the "sparse coupling conjecture" (Friston, Heins, Ueltzhöffer, Da Costa, & Parr, 2021b; Friston et al., 2022-version 1, removed in version 2). In practice, this has so far been presented as an arbitrary assumption, a conjecture, that excludes examples such as the one given by Biehl et al. (2021) from falling within the scope of the FEP. Furthermore, there are multiple ways to generalise conditional independence relations to dynamic Bayesian networks,[1] which ultimately means that Friston blankets are not just *the* natural, or even unique, description of conditional independence applied to dynamical systems. Contrary to **Ramstead**'s claims, there are reasons to distinguish Pearl and Friston blankets: a Pearl blanket is just a statement about conditional independence, while a Friston blanket is a finely crafted posit that includes both conditional independence and a number of non-trivial additional assumptions that are necessary in order for the construct to play a particular role in the wider FEP theory.

As we have already seen, the debate over the distinction between Pearl blankets and Friston blankets reveals cracks in the conceptual foundations of the wider free energy framework, with past co-authors disagreeing among each other about the legitimacy of this distinction. **Aguilera and Buckley** and **Virgo, Rosas, and Biehl** (**Virgo et al.**) both speak in favour of the distinction. As these authors point out, finding Markov blankets that can delineate boundaries or sensorimotor loops within the bounds of the assumptions made by the FEP is far more difficult than commonly thought. Quite tellingly, **Friston** himself seems to be conflicted about the difference between the two kinds of blankets. On the one hand, he writes that "Pearl and Friston blankets are just Markov blankets in the usual Markovian sense (Pearl, 2009)." On the other, he follows this directly with a rhetorical question that undermines his previous claim: "Are Markov blankets used in an ontological sense under the free energy principle (FEP)? Yes." This statement not only undermines the previous one, as differentiating between the purely technical and more

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

metaphysically laden uses of the Markov blanket formalism was the whole point of the target article, but also undercuts **Andrews**' accusation that we somehow misread or misinterpret the free energy theorists' metaphysical intentions.

A similar kind of confusion can be found in **Fox**'s proposal that the distinction between the two kinds of blankets will only lead to further debate and will not facilitate modelling complex systems. Instead, he advocates for the position that "it may be useful to frame systems in terms of constructs such as Markov blankets, but without applying all technical details and associated mathematics." We fail to see how this is a more constructive solution or how it could help to alleviate, rather than exacerbate, the problems Fox attributes to our account. After all, a lack of clear distinction between formal technicalities and ontological commitments is what got the field into the situation it is in now. It is to the latter topic that we turn next, namely the metaphysics and ontology implied by the use of Friston blankets to demarcate worldly boundaries.

R3. The metaphysics and ontology of Friston blankets

Several of the commentaries on our target article focused on the metaphysical and ontological aspects of our critique, namely whether the use of Friston blankets to demarcate real-world boundaries requires a commitment to either *literalism* or *realism* about these constructs, and also what exactly (scientific) realism about theoretical entities of this kind really commits one to. One initial point we want to pick up on is the importance of distinguishing between *local* scientific realism (i.e., about our attitude towards some particular theoretical entity, such as Friston blankets) and *global* scientific realism (i.e., about our attitude towards theoretical entities in general). It is possible to be a global scientific realist while still advocating anti-realism or instrumentalism towards some particular theoretical construct, which is the position we took ourselves to be suggesting towards the end of the target article. We did not, as **Kiverstein and Kirchhoff** suggest, intend to argue against scientific realism in general, or even to enthusiastically endorse instrumentalism about Friston blankets, as both **Sánchez-Cañizares** and **Colombo** took us to be doing (rather, we think that instrumentalism might be the best option if one is really committed to making use of Friston blankets, instead of simply discarding the construct as unworkable). Other commentators, such as **Andrews**, **Hipólito and van Es**, and **Ramstead**, do seem to be advocating for a general anti-realism or instrumentalism, which would render the particular question of the status of Friston blankets somewhat less pressing (if all theoretical entities are just instrumental tools, then the ontological status of any particular theoretical entity does not matter so much, or perhaps at all).

This brings us to the next point we would like to comment on, which is that we were quite struck by the sheer diversity of ontological attitudes exhibited by proponents (or at least defenders) of both the FEP, and the role of Markov blankets within it. These range from a steadfast realism of some variety, as in the cases of **Kiefer and Hohwy, Kiverstein and Kirchhoff**, and **Wiese**, the more cautious realism exhibited by **Sprevak** and **Seth, Korbak, and Tschantz** (Seth et al.), all the way to the full-blown instrumentalism of **Andrews**, **Hipólito and van Es**, and **Ramstead**. At the very least we hope that our target article, and the commentaries it elicited, have demonstrated that the ontological status of Markov blankets within the FEP is far from settled, and that those working within the framework might do well to better communicate their ontological commitments, both to each other and to the outside world.

In his commentary on our target article **Friston** claims that Markov blankets "are deployed in various scientific fields to 'carve nature at its joints'," while nonetheless conceding that "there are many ways of carving nature at its joints," which seems to us to be at best hedging the ontological status of Friston blankets. He also expresses yet another ontological option, which is the view that Friston blankets are literally real *because* the real world itself is at base composed of mathematical structures of some kind. This is certainly not scientific realism in the traditional sense, but rather something closer to the "it-from-bit" hypothesis that the universe is fundamentally computational or informational, defended for example by Wheeler (1982), Zuse (1982), and Wolfram (2002). Menary and Gillett (2020) have already indicated that there are many unanswered questions about using this kind of formal ontology to ground Markov blankets, and in their commentary (**Menary & Gillett**) they raise further insightful issues with Friston's claim that Markov blankets could ever be used to carve nature at its joints. **Colombo** also suggests interpreting the literalist ontology in this way, and doubts whether a purely instrumentalist interpretation of Friston blankets is either as successful or uninteresting as we made it out to be.

What this points to is the distinct lack of clarity surrounding Friston's own ontological commitments, which appear to range from fully instrumentalist (e.g., Ramstead, Friston, & Hipólito, 2020) to fully realist (e.g., Friston, Wiese, & Hobson [2020], and in his response to our target article). Insofar as Friston's work on Markov blankets in the FEP is the canonical starting point for much of the research that we discussed in the target article, any appeal to the "standard" use of Markov blankets in the framework is going to be necessarily vague. This is not to say that everyone working on the FEP must agree about their ontological commitments (far from it, why not let a thousand flowers bloom), but rather that researchers must be careful to make their own commitments clear and not to assume that this is a settled question on which they can simply adopt the party line.

Several commentators suggested that our three-fold distinction between literalism, realism, and instrumentalism might either be too strict or unclear in some respects – we freely admit to both charges, and we are happy to see the conversation that we started being carried forward in a positive direction. Of particular note here is **Seth et al.**, who proposed that it might be better to think in terms of a spectrum of options ranging from literalism to instrumentalism, with various forms of realism lying in between. We are sympathetic to this suggestion (see Fig. R1), but we disagree with their categorisation of Pearl blankets at the instrumentalist end of this spectrum and Friston blankets at the realist end. Pearl blankets, as we understand them, are indeed just a statistical modelling tool, but such a tool could be used to identify "real" features of the world (just probably not ones that are directly equivalent to the blankets themselves); and while Friston blankets are intended as a more metaphysically robust entity, one could treat this entity in an instrumentalist fashion, as several of the other commentaries demonstrate. The distinction between Pearl blankets and Friston blankets does not directly correspond to any particular position along the realism–instrumentalism spectrum; one can be a global instrumentalist and make use of Friston blankets, or a global realist and make use of Pearl blankets. A similar claim is made by **Williams**, who understands our distinction "as delineating between the formal and the physical." Williams expands on this by

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

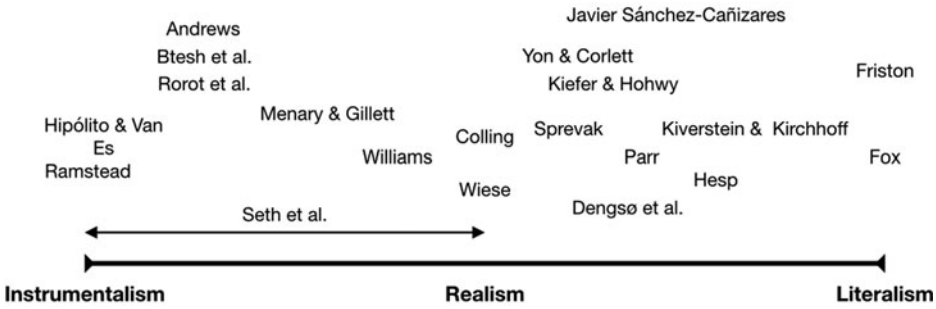


Figure R1. Comparison of the positions espoused by some of the commentaries with regards to their metaphysical commitments to blanket constructs.

proposing that the two kinds of blankets are best understood as being located on different levels of Marr's computational analysis. However, we think that viewing our distinction through the lens of Marr's heuristic division can only blur the boundaries between the instrumental and realist uses of the Friston blanket construct. The fact that a problem solved by a target system can be described using a Markovian formalism does not require a commitment to Markov blankets being explicitly represented (or instantiated) by the system itself. It is only once theorists commit to the target system explicitly possessing or instantiating Markov blankets, either as representations or boundaries, that we can talk about realism.

Some commentators also questioned whether our analysis of the metaphysical uses and misuses of Markov blankets really hits the mark. **Kiverstein and Kirchhoff** argue that our critique relies on what they call the "literalist fallacy," the assumption that realism about the boundaries picked out by Friston blankets implies a belief in the literal existence of Markov blankets in the physical world. This latter belief is indeed what we meant by "literalism," and at least some proponents of the FEP seem to be committed to it. It is hard to see what else is meant when, for example, Allen and Friston (2018) write "the very existence of a system depends upon conserving its boundary, *known technically as a Markov blanket*" (emphasis added) – even if the authors intended to say something else, the most natural reading here is a literalist one, where the physical boundary of a system is identified as a Markov blanket. On the other hand, we by no means wanted to deny the possibility of a more moderate "realism" about the kinds of boundaries that might be picked out by Friston blankets. However, even this more moderate view is not without issues of its own, which we discussed extensively in the target article, and were also mentioned in some of the commentaries. These issues are independent of a literalist interpretation of Friston blankets, and therefore cannot be dismissed by appealing to a "literalist fallacy." At the very least, realism about Friston blankets requires a formalism that guarantees a stable and non-arbitrary mapping between formal model and target system. As it currently stands, the proposed formalism has only been applied to toy models in which such a mapping is already taken for granted, and we have identified in our target article a number of issues that make a stable and non-arbitrary mapping highly unlikely in more realistic (and thus more interesting) cases (see also Aguilera, Millidge, Tschantz, & Buckley, 2021).

Meanwhile, **Andrews** is concerned that while we accuse some proponents of FEP of *reifying* the Markov blanket construct, we fail to give a clear definition of reification and risk conflating it with the metaphysically innocent employment of idealised models. We are happy to at least provisionally accept Andrews' proposed definition of reification as "the mismapping of formal structure onto target phenomena – or theoretical representation thereof – in a manner that leads us to misapprehend the causal structure of nature," although we might add that it is not just the *causal* structure that is at stake here, but rather the structure of the real world more generally (whether or not that structure is either primarily or wholly causal is a further question that we will not attempt to answer here). With this definition on the table, Andrews then argues that our accusations of reification miss the mark, as the use of Markov blankets in the FEP is in fact a "mere modelling gambit" that does not aim to directly uncover the causal structure of the world. Furthermore, Andrews claims that the FEP (or variational free energy minimization) framework does not even *in general* aim to "generate knowledge of natural systems," but is rather just a modelling framework that must be coupled with further auxiliary assumptions before it can generate empirically testable hypotheses. This approach might be congruent with what *some* FEP theorists are doing, and we applaud them for exercising epistemic caution with their uses of the framework, but it seems clear to us that others have taken it far more literally than this, especially when it comes to the use of Friston blankets to demarcate worldly boundaries. Andrews also suggests that our own claim that Markov blankets are "substantiated by the empirical literature" is a case of reification, as formal tools of this kind are not the sort of thing that can be empirically substantiated. Our claim here was not that Markov blankets are empirically substantiated in the sense of being proven to exist, but rather that their repeated application in the empirical literature vindicates their usefulness as a modelling tool – but we are happy to acknowledge that we may have misspoken when making this point, and would invite those whom we have accused of reification to do the same, or else to explain more precisely what they think Friston blankets can tell us about the world.

Finally, some commentaries (**Beck; Ciaunica;** and **Dengsø, Robertson, & Constant [Dengsø et al.]**) discuss the metaphysical possibilities afforded by Pearl blankets and Friston blankets, focusing on the question of how to legitimately connect the map with the territory. Ciaunica emphasises the role of the *cartographer* (i.e., the modeller) in this process, connecting this to older debates between Plato and Aristotle, and somewhat more recent debates between Schlick and Neurath. Beck similarly emphasises the importance of the domain of application of a model, but ultimately considers "metaphysics a pointless endeavor," since "all we have are models." While we agree that it is important to ask who is using a model and for what purpose, we also think that it is important to make clear any theoretical presuppositions about how models connect to the world, that is, what metaphysical or ontological attitude one is taking towards the claims made about Markov blankets and the FEP. Dengsø et al. argue that in order to account for the dynamics of living systems (and to

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

avoid what they call preformationism), Markov blanket theorists need to adopt a process ontology. This is an intriguing option, furthermore demonstrating that many of the interesting philosophical moves available to proponents of FEP are not intrinsic to the framework itself, but rather involve additional metaphysical presuppositions that are necessary in order to make the use of Friston blankets plausible and coherent. In the next section we will accordingly consider what work, if any, Friston blankets *alone* are able to do.

R4. What Friston blankets are not doing

Having defended the validity of our distinction between Pearl and Friston blankets, it is finally time to consider what kind of work Friston blankets, at least in their present form, actually can, and more crucially *cannot*, do. For instance, the causal language adopted by several commentaries reveals the confusion we find when inference with a model is treated as equivalent to inference within a model (as explicitly done by **Kiverstein & Kirchhoff**, **Hesp**, and **Wiese**, and implicitly suggested by **Friston** and **Parr**). This confusion stems from using dependencies intrinsic to models, for example, the fact that changing variable *A* has an effect on variable *B*, to make causal claims about the world extrinsic to that model. We can unpack this better by appealing to an example from Beer's (2004, 2014, 2020) studies of stable patterns in the Game of Life. Do the five rules constituting the game describe interactions that can be treated as causal within the Game of Life universe? Yes, they do. Do the same rules count as causal in our own universe? No, they don't, since the structures described by the Game of Life cannot be straightforwardly mapped onto the structures of our universe. That being said, we could find applications of the Game of Life where it would make sense to treat it as a model of causal relationships between entities in the real world. However, this would only mean that the game has been given a causal interpretation that can also be applied to the dynamics of our world, not that it straightforwardly captures what causation is, or that the Game of Life can by itself be used to discover anything about our world.

Looking closer at the technical commitments defining Friston blankets, the steady-state assumption behind the FEP is perhaps the most controversial one. This assumption naturally leads to questions regarding the relevance of evolutionary dynamics and history more generally (**Veit & Browning**), given the asymptotic independence on initial conditions implied by the even stronger conditions of ergodicity and weak-mixing often assumed by the FEP (see also Di Paolo, Thompson, & Beer [2022] for a related discussion, and Da Costa, Friston, Heins, & Pavliotis [2021] for a technical treatment without assuming ergodicity and weak-mixing, but retaining stationarity). The detailed commentary by **Virgo et al.** demonstrates that a definition of blankets based on steady-state distributions ultimately leads to systems without memory. As clearly portrayed in their example, this is because of the data processing inequality (Cover & Thomas, 2006) applied to Friston blankets, explaining how the internal states of a system of interest (perhaps an agent) cannot store any more information than is present in their current observations at any one point in time. This leads to a severe implication that systems under Friston blankets can only do inference on information located in the perceptual present – past experiences carry no weight for such systems. The only way to allow for some form of memory is to essentially break the stationarity assumption, at least temporarily, as seen in Parr, Da Costa, Heins, Ramstead, and Friston (2021). The implication of breaking this assumption is that the FEP is actually doing its most interesting work *away* from steady state, and therefore *without* Friston blankets. This seems to contradict the large body of work on models based on active inference formulations where past information appears to play a crucial role (see for instance Baltieri & Isomura, 2021; Friston et al., 2015, 2017a, 2017b; Isomura, Shimazaki, & Friston, 2022; Lanillos et al., 2021; Mazzaglia, Verbelen, Çatal, & Dhoedt, 2022; Parr & Friston, 2019 – to give only a few recent examples). It is crucial, however, to highlight that the stipulative nature of Friston blankets is not playing any role in these models – active inference can and does exist without Friston blankets.

The so-called "sparse coupling" assumption, a conjecture that essentially allows the definition of Friston blankets for dynamics starting from Pearl blankets on a steady-state distribution, also leads to questionable consequences. **Aguilera and Buckley** analyse the validity of this assumption, arguing that its high specificity makes it unlikely to realistically characterise relevant (natural and physical) boundaries. In discussing a different derivation of Friston blankets, namely an asymptotic approximation to a weak-coupling equilibrium (Friston et al., 2021b), they also analyse how sparse coupling, however unlikely, is necessary to avoid more dramatic shortcomings: conditional independencies that cannot be guaranteed because of time-dependent relations across different components of a system (cf. **Virgo et al.**). Altogether, we have an assumption that appears to be essential to Friston blankets, sparse coupling, that may unfortunately just not be able to capture any interesting properties of sensorimotor loops, agents, or living systems.

Moving beyond these somewhat technical details, we believe that there are even more fundamental grounds to question the role of Friston blankets within and beyond the FEP. The main worry is captured well by claims that Markov blankets (whether Pearl or Friston) are "formal tools […that] 'carve nature at its joints'" (**Friston**). In the target article, we present and discuss different examples that show why this is simply not the case. In the primordial soup model (Friston, 2013), the experimenter chooses which set of states to designate as internal and which as external, essentially determining *a priori* where the blanket ought to be. Similarly, in the patellar reflex example, which should be properly and more correctly understood following **Spiegel**'s proposed alternative setup, the experimenter once again chooses what constitutes the set of internal states from the vantage point of a scientist who could be asking different possible questions. Contrary to what **Hesp** suggests, our goal with this example was to highlight the ambiguous role of co-parents within a single time slice (as required by Friston blankets), not to look at dependencies across time steps (which are not a part of Friston blankets anyway, as explained above following Friston et al., 2021a). Friston seems to imply that this aspect of experimenter (or modeller) choice is a *feature* of his formulation, but as **Suzuki, Miyahara, and Miyazono** (**Suzuki et al.**) point out, this sidesteps the real question of whether Friston blankets are patterns already present in the "natural" cloth, or patterns that are intentionally introduced by an external agent (either by the tailors themselves, or perhaps just by the imagination of the courtiers).

In a series of commentaries addressing this very issue, it seems to be widely recognised, and in some cases welcomed (**Andrews**; **Hipólito & van Es**; **Kiverstein & Kirchhoff**; **Sprevak**; **Wiese**), that Friston blankets are not foundational to any theory of systems, things, sensorimotor loops, or agents. The heavy lifting is done by a scientist injecting their own assumptions into a

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

model (**Facchin**; **Hipólito & van Es; Kiverstein & Kirchhoff**; **Menary & Gillett**; **Raja et al.**; **Rorot, Korbak, Litwin, & Miłkowski [Rorot et al.]**; **Sprevak**; **Suzuki et al.**; **Wiese**). For example, the choice of which parts (particles) should count as internal in the primordial soup model is essentially arbitrary. As has already been stated, **Friston** believes this is a deliberate and advantageous feature. Some of his defenders view this simply as part of standard modelling practice in science (**Andrews; Hipólito & van Es; Kiverstein & Kirchhoff; Sprevak; Wiese**), and not a reason to dismiss the use of Friston blankets without empirical tests validating their practical applications. However, others (**Facchin; Menary & Gillett; Raja et al.; Rorot et al.; Suzuki et al.**) agree with the critical position we laid out in the target article, and call for clarity in how the explanatorily relevant blankets are identified, and who or what is doing this identification. These two positions are clearly not inconsistent, but they do require that we carefully consider any explanatory role attributed to Friston blankets. Friston blankets cannot be instrumental in defining interesting boundaries, if such boundaries have to be deliberately selected ahead of time for this tool to be successfully applied to. As eloquently argued by **Colling**, a "formalism itself does not licence predictions about which systems are amenable to the formalism and which are [not]." Colling argues that this problem has been faced by other frameworks in the past, such as dynamical systems theory, and points out that defenders of Friston blankets need to "provide an explanation or prediction of which systems are amenable to the formalism – or because the formalism is applicable to every 'thing' (Friston, 2019), which systems are amenable to specific applications of the formalism independent of the particular application of the formalism itself."

This then leads us to some of the core debates regarding teleology, autopoiesis, and the foundational principles of enactive cognitive science in relation to Friston blankets and the FEP, which were discussed in several commentaries. **Veit and Browning** argue that ignoring the teleonomic context of biological systems is ultimately what is causing a lot of the confusion surrounding Friston blankets since previous research has largely overlooked questions about the evolutionary history or function of biological boundaries. We sympathise with this concern, but we worry that the confusion surrounding these issues will only grow, as the ambitions of FEP's proponents seem to have recently moved away from the idea of seeking to explain *only* biological systems (Friston, 2010; Friston, Kilner, & Harrison, 2006) to explaining systems in general (Da Costa et al., 2021; Friston, 2019; Friston et al., 2022). This move could call into question the necessity of teleonomics, but at the same time also raises the issue of whether the FEP is still even trying to say something specific about biology, or cognition, at all. **Raja et al.** and **Suzuki et al.** (see also Di Paolo et al., 2022) carefully observe that Friston blankets are not intrinsically self-determined properties of a system, because as we said above they require a scientist to specify an internal partition, and thus fall short of capturing the *autopoietic* mechanisms found in natural systems (contrary to what, e.g., **Hesp**, **Kiverstein & Kirchhoff**, and **Seth et al.** claim). **Nave** argues forcefully that FEP's abstracting away from the "metabolic turnover" of living systems leads one to "fundamentally misconceive what an organism is." Simply put, FEP ignores that a biological agent's structure not only constrains its own dynamics, but that the dynamics also constitute its structure (cf. Jonas, 1966).

So, what can Friston blankets actually do? It seems clear, at this point, that a Friston blanket-centric metaphysics that makes claims about "blankets of the mind" or "blankets of life" is deeply problematic. It is most certainly not a project that can be accomplished by "just doing the maths," since the relation between Friston blankets and Pearl blankets is all but straightforward, and Friston blankets currently appear to be afflicted by some potentially disqualifying limitations. Furthermore, it should not be presented as a principles-first approach for defining which entities do inference within a model (Friston, 2019), as such entities have to be specified in advance by the modeller. Some authors, such as **Wiese**, acknowledge that the framework is primarily interested in defining a fundamental metaphysics, rather than realising a purely empirical research programme, and we think that this approach at least offers a more perspicuous starting point for future research.

R5. Not everything needs to be a blanket

Having defended the distinction between Pearl blankets and Friston blankets, discussed leading approaches to the metaphysical commitments of the latter, and explored the kind of explanatory work they might be used for, we now conclude our response to the commentaries by looking ahead to the future of Friston blankets and the FEP in science and philosophy.

The most obvious way forward has already been proposed in our target article and involves the tedious, but manageable, task of cleaning up and clarifying the use of the formal constructs in the FEP literature. While we do not expect to see a monolithic consolidation of the framework, confusion about Friston blankets being "just" Pearl blankets still pervades the majority of work on the FEP and must be addressed somehow. There are different interpretations as to who tailors the blankets (**Facchin**; **Friston; Nave**; **Raja et al.**; **Suzuki et al.**), what their role is (**Aguilera & Buckley**; **Menary & Gillett**; **Parr**; **Sprevak**; **Wiese**), and how we should even define a blanket (**Btesh et al.**; **Kiverstein & Kirchhoff**; **Virgo et al.**). The only point that remains uncontroversial is that in order to try and make use of the FEP and its associated formal tools, authors ought to explicitly state what their starting assumptions are.

A viable, but perhaps less appreciated, alternative is to simply make use of the FEP without committing to Friston blankets. While the literature on the FEP has recently been centred around the definition and use of Friston blankets, this hasn't always been the case. As the generalisation of conditional independence relationships to dynamical settings is not in principle unique to Friston blankets, one could simply abandon Friston blankets for a different construct. This could avoid some of the construct's current shortcomings (as discussed by **Aguilera & Buckley** and **Virgo et al.**) by introducing more transparent and explicit assumptions and getting rid of the unhelpful "just do the maths" rhetoric. More interestingly, however, broader work on active inference, predictive coding, and prediction error minimization has been flourishing mostly independently from the notion of Friston blankets (see, e.g., Lanillos et al., 2021; Mazzaglia et al., 2022; Millidge, Seth, & Buckley, 2021; Spratling, 2016 for reviews in neuroscience, robotics, and machine learning). While the mathematical connections between these ideas are unquestionable (especially under Gaussian assumptions), their level of commitment to the full tenets of the FEP varies greatly. This showcases that active inference by itself could be used as a simple alternative to reinforcement learning formulations of behaviour and decision making, one that appeals to a (different) cost function (Millidge, Tschantz, Seth, & Buckley, 2020), without committing to the

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

FEP's extra baggage. Importantly, a version of the FEP that does not appeal to Friston blankets need not collapse into an inherently instrumentalist position, detached from any semblance of empirical validation (about which we are warned by **Colombo**). An active inference alternative to reinforcement learning should be judged on its own merits, according to the empirical results that it generates and the theoretical virtues that it demonstrates.

A third, and yet more radical, answer to the issue of FEP's future brings our attention to the framework as a whole. Although we have already provided a set of actionable solutions to fix the problems surrounding Friston blankets, and have shown that many of the models from the FEP toolkit have seen continuous development outside of the framework, we would like to end by asking about the overall benefit of the FEP for science. With **Allen**'s accusation of a "cargo cult" and **Gomez-Marin**'s worries that the FEP might be "not even wrong" and "appears sufficiently vague to be immune to empirical data," it is worth considering that we might just not need the FEP at all.

Nonetheless, if one wishes to adopt this radical strategy, it is important to do it for the right reasons. For instance, **Spector and Graham** argue that quantities defined by the FEP are needlessly obfuscating the differences between information-theoretic and thermodynamical meanings of terms like "energy," and cannot be easily tied to a meaningful physical interpretation. However, this simply follows a trend in modern stochastic thermodynamics (where the focus is on far from equilibrium systems), and is not as unique or devastating to the FEP as the authors make it out to be. Unlike the classical equilibrium case discussed by these authors, thermodynamic quantities such as heat, free energy, and so on are usually not well defined far from equilibrium, and are best tackled by a formulation based on information theory/geometry (for a recent review see for instance Kim, 2021). Further discussion on the relation of the FEP to physics can be found in Friston (2019) and Friston et al. (2022), which now clarify that the free energy invoked by the FEP is not the same as the free energy defined in thermodynamics.

Similarly, both **Raja et al.** and **Stoffregen and Heath** take the FEP to be incongruent with the commitments of ecological psychology, and see this as a reason to discard it. Stoffregen and Heath seem overly impressed by talk of "inference" within FEP, which does not need to have the intellectualistic commitments that they think it does (cf. Bruineberg & Rietveld, 2014). Raja et al. state that the FEP cannot model relational properties, such as affordances. However, they disregard a productive line of research within ecological psychology that does seek integration with selectionist neuroscience, of which FEP is a variety (see, e.g., Bruineberg & Rietveld, 2019; Reed, 1989, 1996).

In searching for alternatives to the FEP, it might be useful to follow **Btesh et al.** and **Yon and Corlett**, and consider whether a more intervention-focused causal framework (in the sense of Pearl, 2009) could lead to more insightful explanations of cognition. As **Btesh et al.** argue, taking inspiration from an established literature on causality in animal cognition, non-causal alternative frameworks have so far fallen short of explaining any truly high-level features of intelligent and cognitive systems (Marcus, 2018; Pearl & Mackenzie, 2018). On the other hand, Yon and Corlett propose that while Friston blankets may not be the right solution, a different type of blanket, which they named a "cognitive blanket" (cf. the cognitive domain in autopoiesis [Beer, 2014]), based on causal interventions à la Pearl, may provide a more direct bridge to inferential theories of cognition where brains act as hypothesis testing machines by intervening on the environment. Such a framework should still be mindful of the fundamental dichotomy drawn in our target article, between inference *with* or *within* a model.

Overall, we think that a plurality of methods is likely to be the most fruitful approach for a cohesive study of all the diverse aspects of cognition and mind, as evidenced by the wide range of perspectives offered in the 35 commentaries that we received. The tools offered by the FEP, including the new construct of a Friston blanket, might, with some refinement, become a valuable addition to this plurality of methods, but we should keep in mind that not every boundary needs to be a blanket.

R6. Conclusion

In the target article we critically assessed the current state of the philosophical and scientific literature using Markov blankets within the FEP, finding a number of technical slippages and conceptual unclarities. We hoped our target article could serve as an open invitation for those defending the FEP to make clear what their commitments really are, and how those commitments can support the claims they make about Markov blankets. The 35 commentaries we received show that there is widespread disagreement about the role of the Markov blanket construct within the FEP, its ontological status, and the explanatory projects in which it is embedded. Therefore, we do think that the Emperor is currently naked, or at least wearing a different costume each time that he appears. Some commentaries provided fruitful suggestions for how best to redress the Emperor – time will tell whether these attempts will be successful. We hope that the target article, the commentaries, and our reply will all contribute to a more nuanced and productive discussion moving forward.

Financial support. J. B. is funded by a Macquarie Research Fellowship. K. D.'s work is funded by the Volkswagen Stiftung grant no. 87 105.

Conflict of interest. None.

Note

1. Some alternatives to Friston's preferred steady-state assumption + sparse coupling conjecture include:

(a) Friston's very own alternative, that is, an asymptotic approximation to a weak-coupling equilibrium, found in Friston et al. (2021c) and discussed by **Aguilera and Buckley**,
(b) Biehl's example, dropping the above conjecture while retaining the steady-state assumption and adding other constraints, as in Appendix A in Biehl et al. (2021) and Appendices A and B in Friston et al. (2021a),
(c) definitions that don't require stationarity or even Markov properties, based on computational mechanics (Rosas, Mediano, Biehl, Chandaria, & Polani, 2020),
(d) postulating blankets which are not relegated to a single time slice and thus that can in principle consider history (**Virgo, Rosas, & Biehl**), or just
(e) one of the existing alternatives before the advent of Friston blankets, for example, Flesch and Lucas (2007), Materassi and Salapaka (2014), or Koster (1999) in particular for something related to the worries about directed cyclic graphs raised by **Aguilera and Buckley**.

References

Aguilera, M., Millidge, B., Tschantz, A., & Buckley, C. L. (2021). How particular is the physics of the free energy principle? *Physics of Life Reviews*.

Allen, M., & Friston, K. (2018). From cognitivism to autopoiesis: Towards a computational framework for the embodied mind. *Synthese*, *195*(6), 2459–2482.

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press

Anderson, H. C. (1837). The Emperor's New Clothes. English translation by Jean Hersholt available at https://andersen.sdu.dk/vaerk/hersholt/TheEmperorsNewClothes_e.html

Baltieri, M., & Isomura, T. (2021). Kalman filters as the steady-state solution of gradient descent on variational free energy. *arXiv preprint*, arXiv:2111.10530.

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.

Biehl, M., Pollock, F. A., & Kanai, R. (2021). A technical critique of some parts of the free energy principle. *Entropy*, *23*(3), 293.

Bruineberg, J., & Rietveld, E. (2014). Self-organization, free energy minimization, and optimal grip on a field of affordances. *Frontiers in Human Neuroscience*, *8*, 599. https://doi.org/10.3389/fnhum.2014.00599

Bruineberg, J., & Rietveld, E. (2019). What's inside your head once you've figured out what your head's inside of. *Ecological Psychology*, *31*(3), 198–217.

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory* (2nd ed.). Wiley.

Da Costa, L., Friston, K., Heins, C., & Pavliotis, G. A. (2021). Bayesian mechanics for stationary processes. *Proceedings of the Royal Society A*, *477*(2256), 20210518.

Di Paolo, E., Thompson, E., & Beer, R. (2022). Laying down a forking path: Tensions between enaction and the free energy principle. *Philosophy and the Mind Sciences*, *3*, 1–39.

Flesch, I., & Lucas, P. (2007). Independence decomposition in dynamic Bayesian networks. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty* (pp. 560–571). Springer, Berlin, Heidelberg.

Frankfurt, F. (2005). *On bullshit*. Princeton University Press.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138.

Friston, K. (2013). Life as we know it. *Journal of the Royal Society Interface*, *10*(86), 20130475.

Friston, K. (2019). A free energy principle for a particular physics. *arXiv preprint*, arXiv:1906.10184.

Friston, K., Da Costa, L., Sajid, N., Heins, C., Ueltzhöffer, K., Pavliotis, G. A., & Parr, T. (2022). The free energy principle made simpler but not too simple. *arXiv preprint*, arXiv:2201.06387.

Friston, K., Heins, C., Ueltzhöffer, K., Da Costa, L., & Parr, T. (2021b). Stochastic chaos and Markov blankets. *Entropy*, *23*(9), 1220.

Friston, K., Wiese, W., & 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*, *23*, 1076.

Friston, K. J., Fagerholm, E. D., Zarghami, T. S., Parr, T., Hipólito, I., Magrou, L., & Razi, A. (2021c). Parcels and particles: Markov blankets in the brain. *Network Neuroscience*, *5*(1), 211–251.

Friston, K. J., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017a). Active inference: A process theory. *Neural Computation*, *29*(1), 1–49.

Friston, K. J., Kilner, J., & Harrison, L. (2006). A free energy principle for the brain. *Journal of Physiology-Paris*, *100*(1), 70–87.

Friston, K. J., Parr, T., & de Vries, B. (2017b). The graphical brain: Belief propagation and active inference. *Network Neuroscience*, *1*(4), 381–414.

Friston, K. J., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., & Pezzulo, G. (2015). Active inference and epistemic value. *Cognitive Neuroscience*, *6*(4), 187–214.

Isomura, T., Shimazaki, H., & Friston, K. J. (2022). Canonical neural networks perform active inference. *Communications Biology*, *5*(1), 1–15.

Jonas, H. (1966). *The phenomenon of life: Toward a philosophical biology*. Northwestern University Press.

Kim, E. J. (2021). Information geometry, fluctuations, non-equilibrium thermodynamics, and geodesics in complex systems. *Entropy*, *23*(11), 1393.

Koster, J. T. (1999). On the validity of the Markov interpretation of path diagrams of Gaussian structural equations systems with correlated errors. *Scandinavian Journal of Statistics*, *26*(3), 413–431.

Lanillos, P., Meo, C., Pezzato, C., Meera, A. A., Baioumy, M., Ohata, W., … Tani, J. (2021). Active inference in robotics and artificial agents: Survey and challenges. *arXiv preprint*, arXiv:2112.01871.

Marcus, G. (2018). Deep learning: A critical appraisal. *arXiv preprint*, arXiv:1801.00631.

Materassi, D., & Salapaka, M. V. (2014). Notions of separation in graphs of dynamical systems. *IFAC Proceedings Volumes*, *47*(3), 2341–2346. https://doi.org/10.3182/20140824-6-ZA-1003.02661

Mazzaglia, P., Verbelen, T., Çatal, O., & Dhoedt, B. (2022). The free energy principle for perception and action: A deep learning perspective. *Entropy*, *24*(2), 301.

Menary, R., & Gillett, A. J. (2020). Are Markov blankets real and does It matter? In D. Mendonca, M. Curado, & S. S. Gouveia (Eds.), *The philosophy and science of predictive processing* (pp. 39–58). Bloomsbury Academic.

Millidge, B., Seth, A., & Buckley, C. L. (2021). Predictive coding: A theoretical and experimental review. *arXiv preprint*, arXiv:2107.12979.

Millidge, B., Tschantz, A., Seth, A. K., & Buckley, C. L. (2020). On the relationship between active inference and control as inference. In T. Verbelen, P. Lanillos, C. L. Buckley, & C. De Boom (Eds.), *Active Inference. IWAI 2020. Communications in Computer and Information Science* (Vol. *1326*, pp. 3–11). Springer. https://doi.org/10.1007/978-3-030-64919-7_1

Parr, T., Da Costa, L., Heins, C., Ramstead, M. J. D., & Friston, K. J. (2021). Memory and Markov blankets. *Entropy*, *23*(9), 1105.

Parr, T., & Friston, K. J. (2019). Generalised free energy and active inference. *Biological Cybernetics*, *113*(5), 495–513.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Pearl, J., Geiger, D., & Verma, T. (1989). Conditional independence and its representations. *Kybernetika*, *25*(7), 33–44.

Pearl, J., & Mackenzie, D. (2018). *The book of why*. Basic Books.

Ramstead, M., Friston, K., & 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.

Reed, E. S. (1989). Neural regulation of adaptive behavior: An essay review of neural Darwinism. *Ecological Psychology*, *1*(1), 97–117. https://doi.org/10.1207/s15326969eco0101_5

Reed, E. S. (1996). *Encountering the world: Toward an ecological psychology*. Oxford University Press.

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* (pp. 187–198). Springer.

Spratling, M. W. (2016). Predictive coding as a model of cognition. *Cognitive Processing*, *17*(3), 279–305.

Wheeler, J. A. (1982). The computer and the universe. *International Journal of Theoretical Physics*, *21*(6–7), 557–572.

Wolfram, S. (2002). *A new kind of science*. Wolfram Media.

Zuse, K. (1982). The computing universe. *International Journal of Theoretical Physics*, *21*(6–7), 589–600.

https://doi.org/10.1017/S0140525X21002351 Published online by Cambridge University Press