Draft only. Do not quote without permission.
mark.sprevak@ed.ac.uk

Final version due to appear in *Behavioral and Brain Sciences*

Commentary on Bruineberg, J., Dolega, K., Dewhurst, J. & Baltieri, M. (2021) 'The Emperor's new Markov blankets', *Behavioral and Brain Sciences*, 1–63: doi:10.1017/S0140525X21002351

What realism about agents requires

Mark Sprevak
University of Edinburgh

4 January 2022

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 – i.e. 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 agent. 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.

Bibliography

Craver, C. F. (2009). "Mechanisms and natural kinds". In: *Philosophical Psychology* 22, pp. 575–594.

Dennett, D. C. (2017). *From Bacteria to Bach and Back: The Evolution of Minds*. New York, NY: W. W. Norton and Company.

Huebner, B. (2014). *Macrocognition*. Oxford: Oxford University Press.

Kingma, E. (2019). "Were you a part of your mother?" In: *Mind* 128, pp. 609–646.