

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, neuro-cognitive 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 Angel 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<sup>a</sup> and Daniel Graham<sup>b</sup> 

<sup>a</sup>Department of Physics, Hobart and William Smith Colleges, Geneva, NY 14456, USA and <sup>b</sup>Department of Psychological Sciences, Hobart and William Smith Colleges, Geneva, NY 14456, USA

[spector@hws.edu](mailto:spector@hws.edu)

[graham@hws.edu](mailto: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

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](mailto:rainer_spiegel@hotmail.com)

[rainer.spiegel@usb.ch](mailto: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