Behavioral and Brain Sciences

Redressing the Emperor in Causal Clothing

--Manuscript Draft--

Manuscript Number:	
Full Title:	Redressing the Emperor in Causal Clothing
Short Title:	Redressing the Emperor in Causal Clothing
Article Type:	Open Peer Commentary
Corresponding Author:	David Lagnado UCL London, UNITED KINGDOM
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	UCL
Corresponding Author's Secondary Institution:	
First Author:	Victor Btesh, BA, MSc
First Author Secondary Information:	
Order of Authors:	Victor Btesh, BA, MSc
	Neil Bramley, BA, MSc, PhD
	David Lagnado, BA, MSc, PhD
Order of Authors Secondary Information:	
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.

Target article authors: Bruineberg, Dolega, Dewhurst and Baltieri

Word counts:
Abstract: 60 words
Main text: 1000 words
References: 787 words

Redressing the Emperor in Causal Clothing

Victor Btesh (corresponding author)
Experimental Psychology Department, University College London, United Kingdom
E-mail: victor.btesh.19@ucl.ac.uk

Neil Bramley
Psychology Department, University of Edinburgh, United Kingdom
E-mail: neil.bramley@ed.ac.uk
Home page: https://www.bramleylab.ppls.ed.ac.uk/

David Lagnado
Experimental Psychology Department, University College London, United Kingdom
E-mail: d.lagnado@ucl.ac.uk
Home page: https://www.ucl.ac.uk/lagnado-lab/david_lagnado.html

Abstract (60 words)

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.

Commentary (1000 words)

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 possess a Pearl blanket (which therefore say 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 et al., 2019), body substructures such as the brain (Seth & Friston, 2016) and eyes (Parr & Friston, 2018) as well as larger organisms (Buckley et al., 2017; Veissière et al., 2019). This plurality of *blankets* is acknowledged by Parr (2019) and celebrated by Kirchhoff et al. (2018) as evidence for the ubiquity of the Free Energy

Principle. 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, e.g., 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 et al., 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 et al., 2020), and do everything else associated with cognition such as thinking, planning, imagining and explaining (Sloman & Lagnado, 2015). While Constant 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", e.g., 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 et al., 2017; Lagnado & Sloman, 2004, 2006; Steyvers et al., 2003). Thus, a second form of inference is through active interventions, local alterations to the world that allow the learner to identify causal effects – e.g., 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 et al., 2015, 2017, 2018, 2019; Lagnado & Sloman, 2004; Hagmayer et al., 2007; Rothe et al., 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, i.e. pressing a button (Blaisdell et al., 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 et al., 2007; Tenenbaum et al., 2006) and should exhibit behaviour aimed at disambiguating these asymmetries. As 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 in self-regulatory (Barrett, 2017; Pezzulo et al., 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.

Conflict of interest statements

Victor Btesh: none
Neil Bramley: none
David Lagnado: none

Funding statements

Victor Btesh: This research received no specific grant from any funding agency, commercial or not-for-profit sectors.
Neil Bramley: This research received no specific grant from any funding agency, commercial or not-for-profit sectors.
David Lagnado: This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

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*, *311*(5763), 1020–1022. https://doi.org/10.1126/science.1121872

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

Bramley, N., Gerstenberg, T., Mayrhofer, R., & Lagnado, D. (2019). Intervening in Time. *Time and Causality across the Sciences*, 86–115. https://doi.org/10.1017/9781108592703.006

Bramley, N., Lagnado, D., & 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.

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

Fernbach, P. M., & Sloman, A. (2009). Causal Learning With Local Computations. *Journal of Experimental Psychology: Learning Memory and Cognition*, *35*(3), 678–693. https://doi.org/10.1037/a0014928

Friston, K., Daunizeau, J., & Kiebel, S. J. (2009). Reinforcement learning or active inference? *PLoS ONE*, *4*(7). 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, A., Lagnado, D., & Waldmann, M. R. (2007). Causal reasoning through intervention. In *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). https://doi.org/10.1098/rsif.2017.0792

Lagnado, D., & Sloman, 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., & Sloman, 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., Waldmann, M. R., Hagmayer, Y., & Sloman, A. (2007). Beyond covariation: Cues to causal structure. In *Causal learning: Psychology, philosophy, and computation*. (Vol. 44, Issue 0, pp. 1–48).

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., & 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., & 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

Rottman, B. M., & Keil, F. C. (2012). Causal structure learning over time: observations and interventions. *Cognitive Psychology*, *64*(1–2), 93–125. https://doi.org/10.1016/j.cogpsych.2011.10.003

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). 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, A., & Lagnado, D. (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. In *Cognitive Science* (Vol. 27, Issue 3). 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*. https://doi.org/10.1017/S0140525X19001213