J. Allan Hobson[1]
and Karl J. Friston[2]

A Response to Our Theatre Critics

Abstract: *We would like to thank Dolega and Dewhurst (2015) for a thought-provoking and informed deconstruction of our article, which we take as (qualified) applause from valued members of our audience. In brief, we fully concur with the theatre-free formulation offered by Dolega and Dewhurst and take the opportunity to explain why (and how) we used the Cartesian theatre metaphor. We do this by drawing an analogy between consciousness and evolution. This analogy is used to emphasize the circular causality inherent in the free energy principle (aka active inference). We conclude with a comment on the special forms of active inference that may be associated with self-awareness and how they may be especially informed by dream states.*

Keywords: consciousness; prediction; free energy; neuronal coding; inference; neuromodulation.

Introduction

We enjoyed reading Dolega and Dewhurst's (2015) critique of our earlier paper and thinking about the issues it raised. We begin our response by stating our position clearly — in terms of the few key

Correspondence:
Karl Friston, Wellcome Trust Centre for Neuroimaging, Institute of Neurology, Queen Square, London WC1N 3BG, UK. *Email: k.friston@.ucl.ac.uk*

1 Division of Sleep Medicine, Harvard Medical School, Boston, Massachusetts 02215, USA.

2 The Wellcome Trust Centre for Neuroimaging, University College London, Queen Square, London WC1N 3BG.

points — and then substantiate these points with more detailed arguments. Our response can be summarized as follows:

- We are thoroughgoing physicalists. We are dualists only in asserting that, while the brain is material, the mind is immaterial. Both are physically grounded and both are causal, each upon the other. This circular causality assumption is dear to our hearts.
- The mind is entirely dependent upon the brain, but the mind can exert a force upon the brain (through free energy gradients). These assumptions are in no way Cartesian. We reject Cartesian dualism summarily.
- Our theatre is empirically, not theoretically, inspired. We are conscious of being conscious in waking. When we dream, we erroneously suppose ourselves to be awake. When we become lucid we become aware that we are dreaming.

In what follows, we will revisit these points, motivating them from the theoretical perspective of the free energy principle — and calling upon empirical results in neurobiology and sleep research as evidential support. To motivate the importance of circular causality (our first point), we draw on an analogy between active inference and evolution. This may seem rather odd; however, it allows us to clarify our formal arguments.

The Free Energy Principle and Circular Causality

We take as our starting point the free energy principle: the free energy principle borrows from statistical thermodynamics and population dynamics to provide a description of any (biotic) self-organizing system (Friston, 2013). This principle asserts that any system that conserves its boundaries (known technically as a Markov blanket) can be described as modelling its external milieu on the basis of its sensory impressions (or sensorium) (*cf.* Conant and Ashby, 1970). In neuroscience, this leads to the notion of the embodied Bayesian brain (Knill and Pouget, 2004) and active inference (Friston, Mattout and Kilner, 2011; Friston *et al*., 2015a).[3]

[3] The term active inference is preferred to 'action oriented predictive processing' because it has a more precise meaning — and does not conflate action with a particular (i.e. predictive) process theory: formally, active inference is a corollary of the free energy principle — and subsumes action and perception. The minimization of free energy by action and perception constitutes a state (as if) theory. Predictive coding is a process

The free energy principle applies to any scale, from a virus to an ecosystem. This means the underlying principle remains unchanged, irrespective of whether we are talking about thermodynamics (Evans, 2003), a conscious brain (Friston, Kilner and Harrison, 2006), or the evolution of a species (Sella and Hirsh, 2005). In fact, the link between evolution and the Bayesian brain is more than analogous: it is fairly easy to show that population dynamics in evolution, described by the replicator equation, are formally equivalent to Bayesian filters that have been proposed for perceptual synthesis — such as predictive coding or processing (Harper, 2010). See also Fernando, Szathmary and Husbands (2012). In brief, in natural selection, each new generation corresponds to a Bayesian update, converting a prior distribution over phenotypic characteristics into a posterior distribution. Even more simply, this means that evolution is the process of predicting which phenotypes are best adapted to their econiche. So why is evolution a useful analogy for consciousness?

The first thing it brings to the table is an inherent dualism between the genotype and a phenotype that is encoded by the genotype. This is reflected in the distinction between sufficient statistics and a probability distribution (i.e. probabilistic belief) that is encoded by sufficient statistics. It is this dual aspect that motivated our focus on Cartesian (and property) dualism. From the point of view of the brain, the sufficient statistics correspond to biophysical states like synaptic activity and efficacy, while beliefs are probability distributions that describe the products of conscious (or unconscious) processing. From the point of view of evolution, the genotype corresponds to the genomic make-up that prescribes a phenotype, which actively engages with its econiche. Indeed, one can explain morphogenesis by a genetic encoding of prior beliefs about phenotypic form, that are realized epigenetically through free energy minimization (Friston *et al.*, 2015b).

With the analogy between consciousness and evolution in place, let us consider the key argument of our critics:

> Either conscious phenomena are causally efficacious in virtue of being realized at a particular physical locus, in which case Hobson and Friston end up being committed to Cartesian materialism (since they speak of this locus as a theatre), or the locus of consciousness is identified with a

theory that might mediate free energy minimization and perceptual inference — and has nothing to say about action.

> virtual construct, whose role and relationship to the wider system remains unexplained, rendering it epiphenomenal. This is the crux of our argument: Hobson and Friston's proposal is caught between the two horns of Cartesian materialism and epiphenomenalism. (Dolega and Dewhurst, 2015, pp. 121–2)

Like the authors (and presumably most readers), we are not committed to Cartesian materialism in the sense of a (neurophysiological) locus of consciousness. As noted above, the formal principles underlying the self-organizing and self-evidencing brain (Hohwy, 2014) transcend any particular scale and are equally applicable to a dendritic tree through to the embodied nervous system. This means, there is no locus — there are as many (possibly uncountable) loci as there are molecular, cellular, or neuroanatomical systems (with Markov blankets that are conserved over time) that constitute a brain.

The evolution analogy can usefully clarify this: assuming the existence of a locus of consciousness is as untenable as assuming a locus for evolution. For example, evolution cannot be located in a particular genotype — it is a process that entails population dynamics over multiple phenotypes in constant exchange with their econiche (and other phenotypes). The analogy with evolution further suggests that, like evolution, consciousness is a process not a phenomenon. For example, one might ask what the purposes of qualia and self-awareness are. However, this would be as meaningful as asking what the purposes of phenotypes and species are. Phenotypes and species are the products of an evolutionary process; in the same way that qualia and self-awareness are the products of a conscious process. So what is the purpose of a conscious process? Again, this question is as meaningful (or meaningless) as asking what is the purpose of evolution? Evolution is the process of selecting phenotypes that persist for extended periods of time. Similarly, consciousness is the process of selecting (probabilistic) beliefs that persist for non-trivial periods of time. This line of argument suggests that an understanding of consciousness will have the same form as an understanding of evolution. Inherent in this understanding is a circular causality between the genotype and the phenotype. The genotype encodes the phenotype, while the phenotype determines fluctuations in the prevalence of a genotype. In exactly the same way, sufficient statistics (neuronal activity) encode probability distributions (beliefs), while beliefs determine fluctuations in sufficient statistics. This brings us to the heart of our argument and refutation of epiphenomenalism (see first point).

Our position is that there is a necessary duality to conscious processing that distinguishes between sufficient statistics and the (probabilistic) beliefs they entail. This is not unrelated to the distinction between the genotype and phenotype in natural selection. So could the products of conscious processing (i.e. conscious phenomena) be epiphenomenal? This is possible, provided beliefs do not couple back to the sufficient statistics in a causal fashion. However, this reciprocal causality is exactly what the free energy principle describes: it states that the biophysical (material) states of any self-organizing system are driven by free energy gradients (see second point), where free energy is a functional of a probability distribution or (immaterial) belief. This means, mathematically, there is a circular causality that precludes epiphenomenalism. In short, circular causality binds the two aspects of Cartesian dualism into an inseparable whole.

Note that we are assuming beliefs are entailed or encoded by sufficient statistics. Therefore, there is an isomorphism between sufficient statistics and their probability distributions that is more than nomological — it is a mathematical equivalence. This follows because the definition of sufficient statistics is that they are sufficient to describe a probability distribution. Having said this, we fully concur with the authors that the isomorphic relationship between sufficient statistics and beliefs is nuanced; for example, the belief that 'this is lasting forever' may itself be fleeting — and, clearly, the encoding of a probability distribution over velocity, by motion sensitive neurons in V5, does not mean the neurons are moving.

The Cartesian Theatre and Virtual Reality

So why did we emphasize the Cartesian theatre metaphor for virtual reality? In retrospect, this was a little philosophically naïve. As noted above, the inspiration was more empirical than philosophical. We were using the theatre metaphor to emphasize the role of a generative model or virtual reality in probabilistic inference — associating the 'as if' nature of generative models with a 'play' on a stage or screen (see third point).

Although potentially dangerous, there may be some mileage in the theatre metaphor: recall from above that any neuronal system — from a cellular compartment to entire brain systems — can be characterized as exchanging sufficient statistics across its boundary (Markov blanket). In this sense, there may be many (possibly uncountable)

homunculi — all watching each other through the sufficient statistics they exchange.

This perspective also illuminates the fallacy of a locus in Cartesian materialism. In other words, it highlights the physically distributed and nested organization of putative loci that are necessarily coupled to each other. If we assume that the internal states of each system or locus correspond to sufficient statistics, does this mean we can localize a belief to each system? The answer to this is yes and no.

The answer is no because the sufficient statistics over all loci encode a single probability distribution — because the brain entails a single generative model (by virtue of every neuron being connected to every other neuron, at least vicariously). It is this probability distribution that determines the free energy (or Bayesian model evidence), which describes the exchange of sufficient statistics (or their proxies like prediction error) among the loci. This probability distribution has a coherent and unitary aspect; for example, it may have a single peak that coincides with an expectation, where the expectation (i.e. mean) is a single point in a high dimensional state-space. This means the probabilistic belief can have many dimensions or attributes that are subtended by sufficient statistics in distributed (functionally segregated) brain systems (e.g. 'this red rose smells nice'). The belief *per se* cannot be localized to any one locus, in the same way that the prevalence of a particular phenotype in a population cannot be localized to a single gene. In evolution, (the prevalence of) every gene is connected to every other gene statistically, through their mutual contributions to the phenotype and its adaptive fitness or free energy (Sella and Hirsh, 2005).

The answer is also (a nuanced) yes; for particular generative models with deep hierarchical structure — of the sort found in the brain. This hierarchical structure rests upon a sparsity of connections and statistical dependencies (that reflects the sparse causal structure of the processes generating the sensorium). Mathematically, this sparsity enables approximate Bayesian inference in terms of posterior beliefs that can be factorized. This is known in statistical physics as a mean field approximation and provides a powerful way to solve otherwise intractable Bayesian inverse problems.

Heuristically, this factorization means that there may be subsets of neurons or neuronal populations that encode marginal beliefs about single attributes of the causes of sensations. For example, one brain locus (e.g. V4) could encode colour as a parsimonious explanation for the relative intensities of wavelength selective sensory input, given the

ambient illumination (Zeki and Shipp, 1988). This means, one can have a classical neuropsychological scenario in which the qualia (qualitative experience) of colour is absent — and this dimension is lost from the conscious experience (e.g. the rose ceases to be red). Achromatopsia and related conditions (Zeki, 1990) suggest that multiple loci of conscious experience exist and, crucially, speak to a special form of generative model with sparse dependencies. We will return to this theme later when considering self-consciousness. From the point of view of evolution, the encoding of marginal probability distributions over phenotypic traits could correspond to monogenetic traits. For example, the genetic determinants of achromatopsia (Remmer *et al.*, 2015).

Our use of the virtual reality metaphor seems to be more tenable, from the perspective of Dolega and Dewhurst (2015). From the perspective of the free energy principle, the virtual reality stands in for inversion or fitting of a generative model (aka forward model). It was used largely in the sense of Hobson (2009b). However, it also speaks nicely to the notion of perception as inverse optics or graphics (Kawato, Hayakawa and Inui, 1993; Kersten, 1997). In other words, virtual reality systems rely on rendering and computer graphics, which — for vision — play the role of a forward or generative model. As articulated nicely by Helmholtz, in his treatment of physiological optics:

> Objects are always imagined as being present in the field of vision as would have to be there in order to produce the same impression on the nervous mechanism. (Helmholtz, 1866/1962, p. 25)

In short, perception is in the game of generating a virtual sensorium that approximates sensory impressions (and therefore minimizes prediction error). Dolega and Dewhurst (2015) seem more comfortable with the notion of a virtual reality — and relate it to other perspectives; e.g. Grush (2000), Lenggenhager *et al.*, 2007) that sound entirely consistent.

Consciousness *per se*

We have said relatively little about consciousness as such. We were working at a rather simple (and formal) level in which consciousness is simply the process of optimizing beliefs through inference. Implicit in this argument is equivalence between probabilistic beliefs and the products or phenomena of consciousness. Clearly, some of these beliefs may be subpersonal and others not; e.g. self-consciousness.

One might ask what special aspects of generative models support self-awareness. There is a growing consensus that this probably entails a deep (hierarchical) generative model with, crucially, beliefs about the future. This has been discussed nicely in terms of counterfactual richness (Palmer, Seth and Hohwy, 2015; Seth, 2014), where counterfactual states of the world necessarily imply a generative model that predicts not just the current state of affairs, but what could happen under different actions in the future (Friston *et al.*, 2015a; Seth, 2014).

Counterfactual processing is potentially very important because it could distinguish between the sort of inference producing subpersonal or unconscious influences and the conscious inference implicit in self-awareness. In brief, if a generative model has prior beliefs about the future, it must believe it will minimize free energy. See Friston *et al*. (2015a) for a more detailed treatment. Put simply, this means a generative model that has beliefs about the future must have beliefs about itself. In this sense, there is a reprise of the theatre metaphor, in which mindful agents must have the capacity to make inferences about their own (counterfactual or fictive) behaviour and experiences. This capacity is beautifully illustrated by lucid dreaming, when we infer, correctly, that our experiences are fictive (see third point) (Hobson, 2009a; Voss *et al.*, 2009).

The physiological correlate of dream lucidity — and waking awareness of awareness — is activation of the frontal lobe. We therefore localize awareness of awareness and dream lucidity to the executive functions of the frontal cortex. We hypothesize that activation of this region is critical to self-consciousness — and repudiate any suggestion that 'there is a little man seated in our frontal cortex' or that 'it all comes together' there. We insist only that without frontal lobe activation the brain is not fully conscious.

In summary, we could say, perhaps provocatively, that (self-) consciousness is like a theatre in that one watches something like a play, whenever the frontal lobe is activated. In waking, the 'play' includes the outside world. In lucid dreaming the 'play' is entirely internal. In both states, the 'play' is a model, hence virtual. But it is always physical and is always brain-based.

Conclusion

We hope that this clarifies our position in light of Dolega and Dewhurst's (2015) thoughtful critique. We acknowledge a little philosophical naïveté — and apologize for this. Our hope is that the

mathematical formalism of active inference and the empirical neurobiology of sleep can offer useful constraints on the philosophical issues raised by Dolega and Dewhurst.

Acknowledgments

This work was funded by the Wellcome Trust, the US National Institute of Mental Health, the National Science Foundation, and the MacArthur Foundation. We would like to thank an anonymous reviewer of this work for helpful guidance in presenting these ideas.

References

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

Dolega, K. & Dewhurst, J. (2015) Curtain call at the Cartesian theatre, *Journal of Consciousness Studies*, **22** (9–10), pp. 109–128.

Evans, D.J. (2003) A non-equilibrium free energy theorem for deterministic systems, *Molecular Physics*, **101**, pp. 15551–15554.

Fernando, C., Szathmary, E. & Husbands, P. (2012) Selectionist and evolutionary approaches to brain function: A critical appraisal, *Frontiers in Computational Neuroscience*, **6**, art. 24, doi: 10.3389/fncom.2012.00024.

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

Friston, K., Kilner, J. & Harrison, L. (2006) A free energy principle for the brain, *Journal of Physiology — Paris*, **100** (1–3), pp. 70–87.

Friston, K., Mattout, J. & Kilner, J. (2011) Action understanding and active inference, *Biological Cybernetics*, **104**, pp. 137–160.

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T. & Pezzulo, G. (2015a) Active inference and epistemic value, *Cognitive Neuroscience*, pp. 1–28, doi: 10.1080/17588928.2015.1020053.

Friston, K., Levin, M., Sengupta, B. & Pezzulo, G. (2015b) Knowing one's place: A free-energy approach to pattern regulation, *Journal of the Royal Society Interface*, **12** (105), doi: 10.1098/rsif.2014.1383.

Grush, R. (2000) Self, world and space: The meaning and mechanisms of ego- and allocentric spatial representation, *Brain and Mind*, **1** (1), pp. 59–92.

Harper, M. (2010) *The Replicator Equation as an Inference Dynamic*, arXiv: 0911.1763 [math.DS].

Helmholtz, H. (1866/1962) Concerning the perceptions in general, in Southall, J. (trans.) *Treatise on Physiological Optics*, vol. III, New York: Dover.

Hobson, J.A. (2009a) The neurobiology of consciousness: Lucid dreaming wakes up, *International Journal of Dream Research*, **2** (2), pp. 41–44.

Hobson, J.A. (2009b) REM sleep and dreaming: Towards a theory of protoconsciousness, *Nature Reviews Neuroscience*, **10** (11), pp. 803–813.

Hohwy, J. (2014) The self-evidencing brain, *Noûs*, doi: 10.1111/nous.12062.

Kawato, M., Hayakawa, H. & Inui, T. (1993) A forward-inverse optics model of reciprocal connections between visual areas, *Computation in Neural Systems*, **4**, pp. 415–422.

Kersten, D. (1997) Inverse 3-D graphics: A metaphor for visual perception, *Behavior Research Methods, Instruments, & Computers*, **29** (1), pp. 37–46.
Knill, D.C. & Pouget, A. (2004) The Bayesian brain: The role of uncertainty in neural coding and computation, *Trends in Neurosciences*, **27** (12), pp. 712–719.
Lenggenhager, B., Tadi, T., Metzinger, T. & Blanke, O. (2007) Video ergo sum: Manipulating bodily self-consciousness, *Science*, **317** (5841), pp. 1096–1099.
Palmer, C.J., Seth, A.K. & Hohwy, J. (2015) The felt presence of other minds: Predictive processing, counterfactual predictions, and mentalising in autism, *Consciousness & Cognition*, **36**, pp. 376–389.
Remmer, M.H., Rastogi, N., Ranka, M.P. & Ceisler, E.J. (2015) Achromatopsia: A review, *Current Opinion in Ophthalmology*, **26** (5), pp. 333–340.
Sella, G. & Hirsh, A.E. (2005) The application of statistical physics to evolutionary biology, *Proceedings of the National Academy of Sciences*, **102**, pp. 9541–9546.
Seth, A.K. (2014) A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synesthesia, *Cognitive Neuroscience*, **5** (2), pp. 97–118.
Voss, U., Holzmann, R., Tuin, I. & Hobson, J.A. (2009) Lucid dreaming: A state of consciousness with features of both waking and non-lucid dreaming, *Sleep*, **32** (9), pp. 1191–1200.
Zeki, S. (1990) A century of cerebral achromatopsia, *Brain*, **113** (pt 6), pp. 1721–1777.
Zeki, S. & Shipp, S. (1988) The functional logic of cortical connections, *Nature*, **335**, pp. 311–317.[4]

[4] Editorial note: *JCS* will not be considering further papers in this particular debate. However, this paper will be uploaded to the *JCS* blog following publication (http://www.imprint.co.uk/category/jcs-blog/), where the authors concerned (and *JCS* readers generally) can continue to discuss the topic further.