The Diversity of Bayesian Explanation

A Reply to Dominic L. Harkness

Jakob Hohwy

My claim is that, if we understand the function of the brain in terms of the free energy principle, then the brain can explain the mind. Harkness discusses some objections to this claim, and proposes a cautious way of solidifying the explanatory potential of the free energy principle. In this response, I sketch a wide, diverse, and yet pleasingly Bayesian conception of scientific explanation. According to this conception, the free energy principle is already richly explanatory.

Keywords
Bayesian explanation | Free-energy principle | Functionalism | Mechanistic explanation | Philosophy of neuroscience | Scientific explanation | Scientific unification

Author

Jakob Hohwy
jakob.hohwy@monash.edu
Monash University
Melbourne, Australia

Commentator

Dominic L. Harkness
dharkness@uni-osnabrueck.de
Universität Osnabrück
Osnabrück, Germany

Editors

Thomas Metzinger
metzinger@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt
jennifer.windt@monash.edu
Monash University
Melbourne, Australia

1 Introduction

The free energy principle free energy principle (FEP) is ambitiously touted as a unified theory of the mind, which should be able to explain everything about our mental states and processes. Dominic L. Harkness discusses the route from the principle to actual explanations. He reasonably argues that it is not immediately obvious how explanations of actual phenomena can be extracted from the free energy principle, and then offers positive suggestions for understanding FEP's potential for fostering explanations. The argument I focus on in Hohwy (this collection) is that FEP is not so preposterous that it cannot explain at all; Harkness's commentary thus raises the important point that there may be other obstacles to explanatoriness than being preposterous.

A further aspect of Harkness' approach is to make contact between the discussion of FEP's explanatory prowess and discussions in philosophy of neuroscience about computational and mechanistic explanation. This matters, since, if FEP is really set to dominate the sciences of the mind and the brain, then we need to understand it from the point of view of philosophy of science.

In this response, I will attempt to blur some distinctions between notions currently discussed in the philosophy of science. This serves to show that there is a diversity of ways in which a theory, such as FEP, can be explanatory. I am not, however, advocating explanatory pluralism; rather, I am roughly sketching a unitary Bayesian account of explanation according to which good explanation requires balancing the diverse ways in which evidence is explained away. This seems to me an attractive approach to scientific explanation—not least because it involves applying FEP to itself. The upshot is that even though FEP is not yet a full explanation of the mind, there are several ways in which it already now has impressive explanatory prowess.

2 Explanations, functions and mechanisms

Harkness employs existing views in the philosophy of science to create a divide between functions and mechanisms: functions specify what some phenomenon of interest ought to be doing, they don't specify how it actually does it. For that, a mechanism is needed which, in addition to specifying a functional role, also names the parts of the mechanism that perform this role (i.e., the realisers of the function), for example in the brain. This is thought to limit the explanatory power of FEP, which at its mathematical heart is just functionalist.

Whilst I accept the divide between functions and realisers, I don't think there is much explanatory mileage in naming realisers. If I already know what functional role is being realized, I don't come to understand a phenomenon better by being given the names of the realizing properties. This can be seen by imagining any mechanistic explanation (encompassing both functional role and realisers) where the names of the realizing properties are exchanged for other names. Such a move might deprive us of knowledge of which parts of the world realize this function, but this is not in itself explanatory knowledge. For example, I get to understand the heart by being told the functional role realized by atria and ventricles; I don't lose understanding if we rename the atria "As" and ventricles "Bs".

This is not to deny that we can gain understanding from learning about mechanisms. In particular, if I don't know about a phenomenon of interest, then I might explore the realizer of a particular case, and thereby get clues about the functional role. For example, in the 17th century William Harvey was able to finally comprehensively explain the functional role of the heart by performing vivisection on animals. Indeed, the point of such an exercise is to arrive at a clear and detailed description of a functional role (recall the difference between behaviourism and functionalism is that for the latter, the functional role is not just an input–output profile but also a description of the internal states and transitions between states).

Importantly, exploration (e.g., via vivisection, or via functional magnetic resonance imaging) of a mechanism is not the only way to eventually arrive at explanations. There can be multiple contexts of discovery. In particular, there can be very broad empirical observations as well as conceptual arguments. In the case of FEP, a key observation is that living organisms exist in this changing world. That is, organisms like us are able to maintain themselves in a limited number of states. This immediately puts constraints on any mechanistic explanation, which must cohere with this basic observation. Further, since an organism cannot know a priori what its expected states are, there must be an element of uncertainty reduction going on within the organism in order to estimate its expected states, or model. In a world with state-dependent uncertainty, this must happen through hierarchical inference. With these simple notions, FEP itself is well on its way to being established.

So I don't think it is explanatory power that is limited by being confined, as FEP fundamentally is, to functional roles. This mainly seems to impose a limit on our knowledge of *which* objects realize a given functional role, or it might curb our *progress* in finessing the functional role in question. Whereas it is right to say that FEP is limited because it is merely functional, this limit does not apply to its explanatory prowess.

3 Explanations and mechanism sketches

In assessing the explanatoriness of a functional theory like FEP it is useful, as Harkness proposes, to consider it as a mechanism sketch. Sketchiness, however, comes in degrees, and it is hard to think of any extant scientific account that is not sketchy in some respects—no matter how abundantly mechanistic it is. There doesn't seem to be any principled point at which a sketchy functional account passes over into being a non-sketchy mechanistic account. Rather, an account may become less and less sketchy as the full functional role and its realisers are increasingly revealed. This would be one respect in which the explanation in question would expand: more types and ranges of evidence would be explained, accompanied by a richer understanding of the functional workings of the mechanism.

The idea here is that mechanistic explanation comes in degrees, which makes it hard to say clearly when something is a mechanism sketch. Speaking of organs, consider again the case of the heart. Harvey is often said to have provided the first full account of pulmonary circulation, and it might be true that his account is less sketchy than that of his precursors, such as the much earlier Ibn al-Nafis. Yet even Harvey had areas of ignorance about the heart, and had to deduce some parts of his theory from his hypothesis about the overall function of the heart. Indeed, he readily acknowledges the difficulty of his project:

> When I first gave my mind to vivisections, as a means of discovering the motions and uses of the heart, and sought to discover these from actual inspection, and not from the writings of others, I found the task so truly arduous, so full of difficulties, that I was almost tempted to think, with Fracastorius, that the motion of the heart was only to be comprehended by God. (Harvey 1889, p. 20)

A key question then is how sketchy FEP is—is it more like Harvey's rather comprehensive sketch of the heart, or is it like that of al-Nafis? (If it is not completely misguided, like Galen's claim that there are invisible channels between the ventricles.) Harkness suggests that part of the attraction of FEP is that it comes with more empirical specification than mere Bayesian theory. It is true that much of the literature on FEP tries to map mathematical detail onto aspects of neurobiology. However, the mathematical detail of FEP itself is devoid of particular empirical fact—it is purely functionalist. (We might even say FEP is more fundamental than the Bayesian brain hypothesis, since the latter seems to be derivable from the former.)

However, this austerity with respect to specification of particular types of fact does not make FEP inherently sketchy. The starting point for FEP is the trivial but contingent fact that the world is a changing place and yet organisms exist—that is, that they can maintain themselves in a limited set of fluctuating states. This very quickly leads to the idea that organisms must be recapitulating (modelling) the structure of the world, and that they must be approximating Bayesian inference in their attempt to figure out what their expected states are.

This starting point for FEP gives us a lot of structure to look for in the brains of particular creatures. It calls for hierarchical structures the levels of which can encode sufficient statistics (means and variances) of probability distributions, pass these as messages throughout the system, and engage in explaining away and updating distributions over various time-scales. This has a much more mechanistic flavour than a more pure appeal to Bayes' rule, which leaves many more questions about the inferential mechanistics of the brain unanswered. (Part of the difference here is that FEP suggests that the brain implements approximate Bayesian inference, described in terms of variational Bayes.)

It is reasonable, then, to say that, even when stripped of extraneous neurobiological scaffolding, FEP is not inherently sketchy. It might not have the wealth of particular fact that would make it analogous to Harvey's theory of the heart. But it gives a surprisingly very rich description of the functional role implemented by the brain of living organisms.

4 Explanation and types of functionalism

One might still insist on the point that Harkness raises, namely that, even if FEP is not particularly sketchy when stripped of empirical content, it is really only an account of what the system *should* do, rather than what it *actually* does. There is of course some truth to this, since the mathematical formulation of FEP is an idealization of a system engaged in variational Bayes.

However, perhaps FEP is in a peculiar functionalist category. Its starting point, as I mentioned earlier, is the trivial truth that organisms exist, from which it follows that they must be acting to maintain themselves in a limited set of states, from which it in turn follows that they must be reducing uncertainty about their model. Thus the function described by FEP is not about what the system should or ought to be doing but about what it *must* be doing, given the contingent fact that it exists.

This starting point differs from commonsense functionalism because it is not based on conceptual analysis but is instead based on a basic observation, plus statistical notions. It also differs from empirical functionalisms (cf. psychofunctionalism) because it does not specify functional roles in terms of proximal input–output profiles for particular creatures. Neither are the functional roles it sets out defined in terms of teleologically-defined proper functions (cf. teleosemantics), except in so far as it could be said that the proper function of an organism is to exist.

This category of functionalism, which I dubbed “biofunctionalism”, seems intriguingly different from other kinds of functionalism. It provides a foundational functional role, which *must* be realized in living organisms, and from which more specific processes can be derived (for perception, action, attention etc.). This differs from austere functionalisms, which only say how things ought to be working, and it differs from fully mechanistic functionalisms, which specify how particular types of things actually work.

5 Explanation by unification, and by mechanism revelation

Explanation in science is not just a matter of revealing the full detail of the parts and processes of mechanisms. Explanation is many things, as evidenced by the literature on the topic in philosophy of science. Most commonly, explanation is sought to reveal causes, and the contemporary discussion of mechanisms contributes substantially to this discussion. A different idea is that *unification* is explanatory—and yet explanation by unification is a multifaceted and disputed notion.

I think FEP explains by unification because it is a principle that increases our understanding of many very different phenomena, such as illusions, social cognition, the self, decision, movement, and so on (see *The Predictive Mind*, Hohwy 2013, for examples and discussion). FEP teaches us something new and unexpected about these phenomena, namely that they are all *related* as different *instances* of prediction-error minimization. For example, we are surprised to learn that visual attention and bodily movement are not only both engaged in prediction error minimization, they are essentially identical phenomena. FEP thus explains by providing a new, unified and coherent view of the mind.

In this manner, FEP is explanatory partly in ways that are separate from mechanistic explanation, and also from the discussion of how the functionalist and mechanistic approaches relate to each other.

6 Explanation is itself Bayesian

The comments I have provided so far appear to pull somewhat in different directions. I have argued that there is no sharp delineation between functional and mechanistic accounts, and yet I acknowledged that the functional aspects of FEP do set it apart from fully mechanistic accounts. I have argued that merely naming realisers is not explanatory, yet I have acknowledged that mechanistic accounts are explanatory. I have argued (with Harkness) that FEP explains by guiding particular mechanistic accounts, but also by unification. In each of these cases, there seems to be much diversity, or even tension, in how FEP is said to be explanatory.

This diversity and tension, however, is by design. Explanation is not a one-dimensional affair; rather, a hypothesis, h, can be explanatory in a number of different ways. This can be seen by applying the overall Bayesian framework to scientific explanation itself. The strength of the case for h is consummate with how much of the evidence, e, h can explain away. As we know from the discussion of FEP, explaining away can happen in diverse ways: by changing the accuracy, the precision, or the complexity of h, or by intervening to obtain expected, high precision e. As discussed for FEP in Hohwy (this collection), we can also consider h's ability to explain away e over shorter or longer time scales: if h has much fine-grained detail it will be able to explain away much of the short term variability in e but may not be useful in the longer term, whereas a more abstract h is unable to deal with fine-grained detail but can better accommodate longer prediction horizons.

Sometimes these diverse aspects of Bayesian explaining-away pull in different directions. For example, an attempt at unification via de-complexifying h may come at the loss of explaining some particular mechanistic instantiations. Conversely, an overly complex h may be overfitted and thereby explain away occurrent particular detail extremely well but be at a loss in terms of explaining many other parts of e.

In constructing a scientific explanation, how should one balance these different aspects of Bayesian explanation? Again we can appeal to FEP itself for inspiration: a good explanation minimizes prediction error on average and in the long run. That is, a good explanation should not generate excessively large prediction errors, and should be robust enough to persist successfully for a long time. This is intuitive, since we don't trust explanations that tend to generate large prediction errors, nor explanations that cease to apply once circumstances change slightly.

Formulating the goal of scientific explanation in this way immediately raises the question of what it means for prediction error to be "large" or for a hypothesis to survive a "long time". The answer lies in expected precisions and context dependence. In building a theory, the scientist also needs to build up expectations for the precision (i.e., size) of prediction errors, and for the spatiotemporal structure of the phenomenon of interest. Not surprisingly, these aspects are also found in the conception of hierarchical Bayesian inference.

Achieving this balanced goal requires a golden-mean-type strategy: explanations should not be excessively general nor excessively particular, given context and expectations. That is, h should be able to explain away e in the long term without generating excessive prediction errors in the short term, as guided by expectations of precision and domain.

I think FEP is useful for attaining this golden mean, and that this is what makes FEP so attractive and promising. As a scientific hypothesis, it does not prioritise one type of explanatory aspect over another, but instead balances explanatory aspects against each other such that prediction error concerning the workings of the mind is very satisfyingly minimized on average and in the long run (and this indeed is the message of *The Predictive Mind*). Rather poetically, in my view, this means that we should evaluate FEP's explanatory prowess by applying it to itself.

7 Conclusion

I have agreed, to a large extent, with the points Harkness makes in his commentary. I have however also sought to suggest a more pluralistic perspective on scientific explanation. This ensures that the free energy principle, as it applies to the neural organ, has great potential to explain many aspects of the mind. I went one step further, however, and suggested that behind this explanatory pluralism lies a unified, Bayesian account of explanation, which perfectly mimics the unifying aspects of the free energy principle itself.

References

Harvey, W. (1889). *On the motion of the heart and the blood in animals.* London, UK: George Bell & Sons.

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

——— (2015). The neural organ explains the mind. In T. Metzinger & J. M. Windt (Eds.) *Open MIND.* Frankfurt a. M., GER: MIND Group.