The Neural Organ Explains the Mind

Jakob Hohwy

The free energy principle says that organisms act to maintain themselves in their expected states and that they achieve this by minimizing their free energy. This corresponds to the brain's job of minimizing prediction error, selective sampling of sensory data, optimizing expected precisions, and minimizing complexity of internal models. These in turn map on to perception, action, attention, and model selection, respectively. This means that the free energy principle is extremely ambitious: it aims to explain *everything* about the mind. The principle is bound to be controversial, and hostage to empirical fortune. It may also be thought preposterous: the theory may seem either too ambitious or too trivial to be taken seriously. This chapter introduces the ideas behind the free energy principle and then proceeds to discuss the charge of preposterousness from the perspective of philosophy of science. It is shown that whereas it is ambitious, controversial and needs further evidence in its favour, it is not preposterous. The argument proceeds by appeal to: (i) the notion of inference to the best explanation, (ii) a comparison with the theory of evolution, (iii) the notion of explaining-away, and (iv) a "bio-functionalist" account of Bayesian processing. The heuristic starting point is the simple idea that the brain is just one among our bodily organs, each of which has an overall function. The outcome is not just a defence of the free energy principle against various challenges but also a deeper anchoring of this theory in philosophy of science, yielding an appreciation of the kind of explanation of the mind it offers.

Keywords
Explaining-away | Free energy principle | Functionalism | Inference to the best explanation | Prediction error minimization | Scientific explanation

Author

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

Commentator

Dominic 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 The brain and other organs

Many organs in the body have a fairly specific main function, such as cleaning or pumping blood, producing bile, or digesting. Nothing is ever simple, of course, and all the organs of the body have highly complex, interconnected functional roles. The digestive system involves many different steps; the kidneys help regulate blood pressure; while the heart changes the way it pumps in a very complex and context-dependent manner. Experts in different areas of human biology have a wealth of knowledge about the morphology and physiology of organs, at multiple levels of description. For example, much is known about what cellular and molecular processes occur as the kidneys filter blood, or as food is digested. Knowledge about the functions of organs is not yet complete, but there is reasonable agreement about the overall picture—namely, which organs have what function.

But the brain seems different. There is much less agreement about what is its main function, and much less knowledge about how it fulfills the various functions attributed to it. Of course, everyone agrees that the brain subserves perception, decision-making, and action—and perhaps that it is the seat of consciousness, self and soul. There is a reasonable degree of knowledge about some aspects of the brain, such as the mechanism behind action potentials, and about what happens when neurons fire. But most would agree that it would be controversial or even preposterous to claim that there is one main function of the brain, on a par with the heart's pumping of blood.

Yet there is an emerging view that claims that the brain has one overarching function. There is one thing the brain does, which translates convincingly to the numerous other functions the brain is engaged in. This chapter will introduce this idea and will show that, whereas it may be controversial, the idea is not preposterous. It will help us understand better all the things that the brain does, how it makes us who we are, and what we are.

The main version of the idea is labeled the free energy principle, and was proposed by Karl Friston (2010). It unifies and develops a number of different strands of thinking about the brain, about learning, perception and decision-making, and about basic biology. The principle says that biological organisms on average and over time act to minimize free energy. Free energy is the sum of prediction error, which bounds the surprise of the sensory input to the system. Put one way, it is the idea that brains are hypothesis-testing neural mechanisms, which sample the sensory input from the world to keep themselves within expected states. Generalizing greatly, one might say that, just as the heart pumps blood, the brain minimizes free energy.

Before moving on to introduce and defend this idea, it will be useful to explain why the analogy to the functions of other organs is apt. Once a function is identified it serves as a unifying, organizing principle for understanding what the organ does. For example, even though the heart acts very differently during rest and exercise, it still pumps blood. Similarly, even though the brain acts very differently during the awake state and during sleep it still minimizes free energy. Taking such a general approach therefore helps to provide a unified account of the brain.

Related to this, there is a type of objection that will have little bite on the organ-focused account of the brain. To see this, consider again the heart. The heart pumps blood, and this function is realized in part by the way the contraction of the heart muscle occurs—a process that depends on intricate ion flows across heart cell membranes. One should not object to the notion that the heart pumps blood by referring to the fact that what happens in the heart is an intricate cellular ion flow. This is so even though one might be able to understand much about the heart just by being told the cellular and molecular story. The story about the function and the story about a level of realization of that function are not in conflict with each other. Similarly, one cannot object to the free energy principle by pointing to facts about what the brain does (e.g., what happens as action potentials are generated, or as long-term potentiation is instantiated). The reason for this is that those low-level processes might be ways of realizing free energy minimization. At best such objections are calls for explanatory work of the sort “how can the generation of action potentials be realizations of free energy minimization?”

These two points together suggest that the functional, organ-based account of the brain is reductionist in two ways (familiar from discussions in philosophy of science). On the one hand it seeks to reduce all the different things the brain does to one principle, namely free energy minimization. This is a kind of theory reduction, or explanatory unification. It says that one theory explains many different things. On the other hand, it is consistent with a kind of metaphysical reduction where the overall function is in the end realized by a set of basic physical processes. Here, mental function is fully physical and fully explained by free energy minimization. It is interesting to note that no one would object to such a two-fold reductionism for the heart and other organs, yet it is controversial or even preposterous to do so for the organ that is the brain. For these reasons, it is useful to keep in mind the simple idea that the brain is also an organ. Much of the discussion in this chapter revolves around these two reductive aspects: how can the free energy principle *explain everything*? And can it provide the *functional* scaffolding that would allow realization by brain activity?

2 Minimizing free energy (or average prediction error minimization)

Consider the following very broad, very simple, but ultimately also very far-reaching claim: the brain’s main job is to maintain the organism within a limited set of possible states. This is a fairly trivial claim, since it just reflects that there is a high probability of finding a given organism in some and not other states, combined with the obvious point that the organism's brain, when in good working order, helps explain this fact. It is the brain's job to prevent the organism from straying into states where the organism is not expected to be found in the long run. This can be turned around such that, for any given organism, there is a set of states where it is expected to be found, and many states in which it would be surprising to find it. This is surely an entirely uncontroversial observation: we don't find all creatures with equal probability in all possible states (e.g., in and out of water). Indeed, since an organism's phenotype results from the expression of its genes together with the influence of the environment, we might define the phenotype in terms of the states we expect it to be found in, on average and over time: different phenotypes will be defined by different sets of states. This way of putting it then defines the brain's job: it must keep the organism within those expected states. That is, the brain must keep the organism out of states that are surprising given the organism it is—or, in general, the brain must minimize surprise (Friston & Stephan 2007).

Here surprise should not be understood in commonsense terms, in the way that a surprise party, say, is surprising. "Surprise" is technically surprisal or self-information, which is a concept from information theory. It is defined as the negative log probability of a given state, such that the surprise of a state increases the more improbable it is to find the creature in that certain state (in this sense a fish out of water is exposed to a lot of surprise). Surprise is then always relative to a model, or a set of expectations (being out of water is not surprising given a human being's expectations). States in which an organism is found are described in terms of the causal impact from the environment on the organism (for example, the difference to the fish between being in water and being out of water). This, in turn, can be conceptualized as the organism's sensory input, in a very broad sense, including not just visual and auditory input but also important aspects of sensation like thermoreception, proprioception, and interoception. Surprising states are then to be understood as surprising sensory input, and the brain's job is to minimize the surprise in its sensory input—to keep the organism within states in which it will receive the kind of sensory input it expects.

To be able to use this basic idea about the brain's overall function in an investigation of all the things minds do we need to ask how the brain accomplishes the minimization of surprise. It cannot assess surprise directly from the sensory input because that would require knowing the relevant probability distribution as such. To do this it would need to, impossibly, average over an infinite number of copies of itself in all sorts of possible states in order to figure how much of a surprise a given sensory input might be. This means that to do its job, the brain needs to do something else; in particular it must harbor and finesse a model of itself in the environment, against which it can assess the surprise of its current sensory input. (The model concerns expected sensory states, it is thus a model of the states of the brain, defined by the sensory boundary in both interoceptive and exteroceptive terms, see Hohwy 2014.)

Assume then that the brain has a model—an informed guess—about what its expected states are, and then uses that model to generate hypotheses that predict what the next sensory input should be (this makes it a generative model). Now the brain has access to two quantities, which it can compare: on the one hand the predicted sensory input, and on the other the actual sensory input. If these match, then the model is a good one (*modulo* statistical optimization). Any difference between them can be conceived as prediction error, because it means that the predictions were erroneous in some way. For example, if a certain frequency in the auditory input is predicted, then any difference from what the actual auditory input turns out to be is that prediction's error.

The occurrence of prediction error means the model is not a good fit to the sensory samples after all, and so, to improve the fit, the overall prediction error should be minimized. In the course of minimizing prediction error, the brain averages out uncertainty about its model, and hence implicitly approximates the surprise. It is guaranteed to do this by minimizing the divergence between the selected hypothesis and the posterior probability of the hypothesis given the evidence and model. The guarantee stems from the facts that this is a Kullback-Leibler divergence (KL-divergence) which is always zero (when there is no divergence) or positive (when there is prediction error), and which therefore creates an upper bound on the surprise—minimizing this bound will therefore approximate surprise.

The key notion here is that the brain acts to maintain itself within its expected states, which are estimated in prediction error minimization. This is known as the free energy principle, where free energy can be understood as the sum of prediction error (this and the following is based on key papers, such as Friston & Stephan 2007, Friston 2010, as well as introductions in Clark 2013 and Hohwy 2013). Prediction error minimization itself instantiates probabilistic, Bayesian inference because it entails that the selected hypothesis becomes the true posterior, given evidence and model. On this view, the brain is a model of the world (including itself) and this model can be considered the agent, since it acts to maintain itself in certain states in the world.

3 Varieties of prediction error minimization

The central idea here is that, on average and over the long run, surprising states should be avoided, or, prediction error should be minimized. Prediction error minimization can occur in a number of ways, all familiar from debates on inference to the best explanation and many descriptions of scientific, statistical inference.

First, the model parameters can be revised in the light of prediction error, which will gradually reduce the error and improve the model fit. This is perception, and corresponds to how a scientist seeks to explain away surprising evidence by revising a hypothesis. This perceptual process was alluded to above.

Slightly more formally, this idea can be expressed in terms of the free energy principle in the following terms. The free energy (or sum of prediction error) equals the negative log probability of the sensory evidence, given the model (the surprise) + a KL-divergence between the selected hypothesis (the hypothesis about the causes of the sensory input, which the system can change to change the free energy), and the true posterior probability of the hypothesis given the input and model. Since the KL-divergence is never negative, this means that the free energy will bound (be larger than) the surprise. Therefore, the system just needs to minimize the divergence to approximate the surprisal.

Second, the model parameters can be kept stable and used to generate predictions—in particular, proprioceptive predictions, which are delivered to the classic reflex arcs and fulfilled there until the expected sensory input is obtained. This is action, and corresponds to how a scientist may retain a hypothesis and control the environment for confounds until the expected evidence obtains. Since action is prediction error minimization with a different direction of fit, it is labeled active inference.

Slightly more formally (and still following Friston), this notion of action arises from another reorganization of the free energy principle. Here, free energy equals complexity minus accuracy. Complexity may be taken as the opposite of simplicity, and is measured as a KL-divergence between the prior probability of the hypothesis (i.e., before the evidence came in) and the hypothesis selected in the light of the evidence. Intuitively, this divergence is large if many changes were made to fit the hypothesis—that is, if the hypothesis has significant complexity compared to the old hypothesis. Accuracy is the surprise about the sensory input given the selected hypothesis—that is, how well each hypothesis fits the input. Free energy is minimized by changing the sensory data, such that accuracy increases. If the selected hypothesis is not changed, then this amounts to sampling the evidence selectively such that it becomes less surprising. This can only happen through action, where the organism re-organizes its sensory organs or whole body, or world, in such a way that it receives the expected sensory data (e.g., holding something closer in order to smell it).

There are further questions one must ask about action: how are goals chosen and how do we work out how to obtain them? The free energy principle can be brought to bear on these questions too. In a very basic way, our goals are determined by our expected interoceptive and proprioceptive states, which form the basis of homeostasis. If we assume that we can approximate these expected states, as described above, what remains is a learning task concerning how we can maintain ourselves in them. This relies on internal models of the world, including, crucially, modeling how we ourselves, through our action, impact on the sensory input that affects our internal states. Further, we need to minimize the divergence between, on the one hand, the states we can reach from a given point and, on the other, the states we expect to be in. Research is in progress to set out the details of this ambitious part of the free energy program.

Third, the model parameters can be simplified (cf. complexity reduction), such that the model is not underfitted or overfitted, both of which will generate prediction error in the long run. This corresponds to Bayesian model selection, where complexity is penalized, and also to how a scientist will prefer simpler models in the long run even though a more complex model may fit the current evidence very well. The rationale for this is quite intuitive: a model that is quite complex is designed to fit a particular situation with particular situation-specific, more or less noisy, interfering factors. This implies that it will generalize poorly to new situations, on the assumption that the world is a fairly noisy place with state-dependent uncertainty. Therefore, to minimize prediction error in the long run it is better to have less complex models. Conversely, when encountering a new situation, one should not make too radical changes to one’s prior model. One way to ensure this is to pick the model that makes the least radical changes but still explains the new data within expected levels of noise. This is just what Bayesian model selection amounts to, and this is enshrined in the formulations of the free energy principle. A good example of this is what happens during sleep, when there is no trustworthy sensory input and the brain instead seems to resort to complexity reduction on synthetic data (Hobson & Friston 2012).

Fourth, the hypotheses can be modulated according to the precision of prediction error, such that prediction error minimization occurs on the basis of trustworthy prediction error; this amounts to gain control, and functionally becomes attention. This corresponds to the necessity for assessment of variance in statistical inference, as well as to how a scientist is guided by, and seeks out, measurements that are expected to be precise more than measurements that are expected to be imprecise.

Precision optimization is attention because it issues in a process of weighting some prediction errors more than others, where the weights need to sum to one in order to be meaningful. Hence, peaks across the prediction error landscape reflect both the magnitude of the prediction error per se and the weight given to that error based on how precise it is expected to be. This moves the prediction error effort around, much like one would expect the searchlight of attention to move around.

Within this framework, there is room for both endogenous and exogenous attention. Endogenous attention is top-down modulation of prediction error gain based on learned patterns of precision. Exogenous attention is an intrinsic gain operation on error units, sparked by the current signal strength in the sensory input; this is based on a very basic learned regularity in nature, namely that strong signals tend to have high signal to noise ratio—that is, high precision.

In all this, there is a very direct link between perception, action, and attention, which will serve to illustrate some of the key characteristics of the framework. In particular, expected precision drives action such that sensory sampling is guided by hypotheses that the system expects will generate precise prediction error. A very simple example of this is hand movement. For hand movement to occur, the system needs to prioritize one of two competing possible hypotheses. The first hypothesis is that the hand is *not* moving, which predicts a particular kind of (unchanging) proprioceptive and kinesthetic input; the second hypothesis is (the false one) that the hand *is* moving, which predicts a different (changing) flow of proprioceptive and kinesthetic input. Movement will only occur if the second hypothesis is prioritized, which corresponds to the agent harboring the belief that the hand is actually moving. If this belief wins, then proprioceptive predictions are passed to the body, where classic reflex arcs fulfill them. Movement is then conceived as a kind of self-fulfilling prophecy.

A crucial question here is how the actually false hypothesis might be prioritized, given that the actually true hypothesis (that the agent is not moving) has evidence in its favor (since the agent is in fact not moving). Here expected precisions play a role, which means that action essentially turns into an attentional phenomenon: in rather revisionist terms, agency reduces to self-organisation guided by long term prediction error minimization. Hypotheses can be prioritized on the basis of their expected precision: hence if future proprioceptive input is expected to be more precise than current proprioceptive input, the gain on the current input will be turned down, depriving the hypothesis that the agent is not moving of evidence. Now the balance shifts in favor of the actually false hypothesis, which can then begin to pass its predictions to the sensorimotor system. This rather inferential process is then what causes movement to occur. It is an essentially attentional process because acting occurs when attention is withdrawn from the actual input (Brown et al. 2013).

The outstanding issue for this story about what it takes to act in the world is why there is an expectation that future proprioceptive input will be more precise than the current input. One possibility here is that this is based on a prior expectation that exploration (and hence movement) yields greater prediction error minimization gains in the long run than does staying put. Conversely, this is the expectation that the current state will lose its high-precision status over time. Writ large, this is the prior expectation concerning precisions (i.e., a hyperprior), which says that the world is a changing place such that one should not retain the same hypotheses for too long: when the posterior probability of a hypothesis becomes the new prior, it will soon begin to decrease in probability. This is an important point because it shows that the ability to shift attention around in order to cause action is not itself an action performed by a homunculus. Rather, it is just a further element of extracting statistical information (about precisions) from the world.

4 Hierarchical inference and the recapitulating, self-evidencing, slowing brain

A system that obeys the free energy principle minimizes its free energy, or prediction error, on average and over time. It does this through perception, belief updating, action, attention, and model simplification. This gives us the outline of a very powerful explanatory mechanism for the mind. There is reason to think that much of this explanatory promise can be borne out (Clark 2013; Hohwy 2013).

This mechanism shapes and structures our phenomenology—it shapes our lived, experienced world. A good starting point for making good on this idea is the notion of hierarchical inference, which is a cornerstone of prediction error minimization.

Conceive of prediction error minimization as being played out between overlapping pairs of interacting levels of processing in the brain. A pair has a lower level receives input, and a higher level that generates predictions about the input at the lower level. Predictions are sent down (or “backwards”) where they attenuate as well as possible the input. Parts of the input it cannot attenuate are allowed to progress upwards, as prediction error. The prediction error serves as input to a new pair of levels, consisting of the old upper level, which is now functioning as lower input level, and a new upper level. This new pair of levels is then concerned with predicting the input that wasn’t predicted lower down. This layering can then go on, creating in the end a deep hierarchy in our brains (and perhaps a more shallow hierarchy in some other creatures). The messages that are passed around in the hierarchy are the sufficient statistics: predictions and prediction errors concerning (1) the means of probability distributions (or probability density functions) associated with various sensory attributes or causes of sensory input out there in the world, and (2) the precisions (the inverse of variance) of those distributions, which mediate the expected precisions mentioned above.

The hierarchy gives a deep and varied empirical Bayes or prediction error landscape, where prior probabilities are "empirical" in that they are learned and pulled down from higher levels, so they do not have to be extracted de novo from the current input. This reliance on higher levels means that processing at one level depends on processing at higher levels. Such priors higher up are called hyperparameters, for expectations of means, and hyperpriors for expectations of precisions.

The key characteristics of the hierarchy are *time and space*. Low levels of the hierarchy deal with expectations at fast timescales and relatively small receptive fields, while higher levels deal with expectations at progressively slower timescales and wider receptive fields. That is, different levels of the hierarchy deal with regularities in nature that unfold over different spatiotemporal scales. This gives a trade-off between detail and time horizon such that low down in the hierarchy, sensory attributes can be predicted in great detail but not very far into the future, and higher in the hierarchy things can be predicted further into the future but in less detail. This is essential to inference because different causal regularities in nature, working at different time scales, influence each other and thereby create non-linearities in the sensory input. Without such interactions, sensory input would be linear and fairly easy to predict both in detail and far into the future. So the temporal organization of the hierarchy reflects the causal order of the environment as well as the way the causes in the world interact with each other to produce the flow of sensory input that brains try to predict.

The structure of the hierarchy in the brain, and thereby the shape of the inferences performed in the course of minimizing prediction error, must therefore mimic the causal order of the world. This is one reason why hierarchical inference determines the shape and structure of phenomenology, at least to the extent that phenomenology is representational. The way inference is put together in the brain recapitulates the causes we represent in perception. Moreover, this is done in an integrated fashion, where different sensory attributes are bound together under longer-term regularities (for example, the voice and the mouth are bound together under a longer-term expectation about the spatial trajectories of people). This immediately speaks to long-standing debates in cognitive science, concerning for example the binding problem and cognitive penetrability (for which see Chs. 5-6 in Hohwy 2013). Though there is, of course, much more to say about how prediction error minimization relates to phenomenology, so far this suggests that there is some reason to think the austere prediction error minimization machine can bear out its explanatory promise in this regard.

Goals and actions are also embodied in the cortical hierarchy. Goals are expectations of which states to occupy. Actions ensue, as described above, when those expected states, which may be represented at relatively long timescales, can confidently be translated into policies for concrete actions fulfilled by the body. There are some thorny questions about what these goals might be and how they are shaped. One very fundamental story says that our expected states are determined by what it takes to maintain homeostasis. We are creatures who are able to harness vast and deep aspects of the environment in order to avoid surprising departures from homeostasis; though this opportunity comes with the requirement to harbor an internal model of the environment. Reward, here, is then the absence of prediction error, which is controlled by using action to move around in the environment, so as to maintain homeostasis on average and in the long run.

Taking a very general perspective, the brain is then engaged in maintaining homeostasis, and it does so by minimizing its free energy, or prediction error. Minimization of prediction error entails building up and shaping a model of the environment. The idea here is very simple. The better the model is at minimizing prediction error the more information it must be carrying about the true causes of its sensory input. This means that the brain does its job by recapitulating the causal structure of the world —by explaining away prediction error, the brain is essentially becomes a deeply structured mirror of the world. This representational perspective is entailed by the brain's efforts to maintain itself in a low entropy or free energy state. This means that we should not understand the brain as first and foremost in the business of representing the world, such that it can act upon it—which may be an orthodox way of thinking about what the brain does. Put differently, the brain is not selected for its prowess in representation per se but rather for its ability to minimize free energy. Even though this means representation is not foundational in our explanation of the brain, it doesn't mean that representation is sidelined. This is because we don't understand what free energy minimization is unless we understand that it entails representation of the world. (This formulation raises the issue of the possibility of misrepresentation in prediction error minimization, for discussion see Hohwy 2013, Chs. 7-8.)

The brain can be seen, then, as an organ that minimizes its free energy or prediction error relative to a model of the world and its own expected states. It actively changes itself and actively seeks out expected sensory input in an attempt to minimize prediction error. This means the brain seeks to expose itself to input that it can explain away. If it encounters a change in sensory input that it cannot explain away, then this is evidence that it is straying from its expected states. Of course, the more it strays from its expected states, the more we should expect it to cease to exist. Put differently, the brain should enslave action to seek out evidence it can explain away because the more it does so, the more it will have found evidence for its own existence. The very occurrence of sensory input that its model can explain away becomes an essential part of the evidential basis for the model. This means the brain is self-evidencing (Hohwy 2014), in that the more input it can explain away, the more it gains evidence for the correctness of the model and thereby for its own existence.

The notions of recapitulation of the world and of self-evidencing can be captured in an exceedingly simple idea. The brain maintains its own integrity in the onslaught of sensory input by *slowing down* and controlling the causal transition of the input through itself. If it had no means to slow down the input its states would be at the mercy of the world and would disperse quickly. To illustrate, a good dam-builder must slow down the inflow of water by slowing down and controlling it with a good system of dams, channels, and locks. This dam system must in some sense anticipate the flows of water in a way that makes sense in the long run and that manages flows well on average. The system will do this by minimizing "flow errors", and it and its dynamics will thereby carry information about—recapitulate—the states of water flow in the world on the other side of the dam. In general, it seems any system that is able to slow the flow of causes acting upon it must be minimizing its own free energy and thereby be both recapitulating the causes and self-evidencing (Friston 2013).

With these extremely challenging and abstract ideas, the brain is cast as an organ that does one thing only: minimize free energy and thereby provide evidence for its own existence. Just as the heart can change its beat in response to internal and external changes, the brain can change its own states to manage self-evidencing according to circumstances: perceive, act, attend, simplify. The weighting between these ways of minimizing prediction error is determined by the context. For example, it may be that learning is required before action is predicted to be efficient, so perception produces a narrow prediction error bound on surprise before action sets in, conditional on expected precisions; or perhaps reliable action is not possible (which may happen at night when sensory input is so uncertain that it cannot be trusted) and therefore the brain simplifies its own model parameters, which may be what happens during sleep (Hobson & Friston 2012).

This is all extremely reductionist, in the unificatory sense, since it leaves no other job for the brain to do than minimize free energy—so that everything mental must come down to this principle. It is also reductionist in the metaphysical sense, since it means that other types of descriptions of mental processes must all come down to the way neurons manage to slow sensory input.

The next sections turn to the question of whether this extreme explanatory and reductionist theory is not only controversial and ambitious but also preposterous.

5 A preposterous principle? Comparing the free energy principle with evolution

One way to curtail the free energy principle is to allow that the idea of a hypothesis-testing mechanism in the brain may be useful for some but *not all* purposes. Thus the idea could explain, say, visual illusions, but not action. Indeed, versions of the idea in this curtailed form have surfaced many times in the history of philosophy of mind, vision science, and psychology (see Hohwy 2013, Introduction). One view would be that evolution very likely has recruited something like hypothesis-testing, such that the brain can represent the world, but that this likely co-exists with many other types of mechanism that the brain makes use of, for good evolutionary reasons. From this perspective, the universal ambition of the free energy principle is preposterous because it goes against the evolutionary perspective of a tinkering, cobbled-together mechanism.

It is possible of course that a limited-use, Bayesian neural mechanism has evolved in this way. There is no strong evidence that there is in fact something like a circumscribed, modular mechanism. For example, Bayes optimal integration seems to work across modalities and types of sensory attributes (Trommershäuser et al. 2011). On the other hand, there is not yet strong empirical evidence for the ubiquitousness of free energy minimization, though there is emerging evidence of its usefulness for explaining a very surprising range of mental phenomena, from visual perception, illusion, movement, decision, and action.

Speaking more conceptually, the free energy principle is not a theory that lends itself particularly well to piecemeal, curtailed application. Recall that the principle concerns the very shape and structure of the brain, mirroring as it does the causal structure of the world. The very hierarchical morphology of the organ is shaped by free energy minimization. This means that other neural mechanisms, that are not involved in prediction error minimization, would have to have evolved in a way parasitic on the free energy principle rather than alongside it. In this sense, the free energy principle would, at the very least, lay the foundation for everything else. Against this, it could be said that perhaps parts of the brain are not, strictly speaking, part of hierarchical inference. Perhaps subcortical nuclei have evolved independently of free energy. This is therefore an argument for which empirical evidence would be important: are there areas of the brain that are not best described in terms of prediction error message passing?

Continuing the very general approach, the free energy principle has such generality that it tends to monopolize explanation. To demonstrate this, consider the theory of evolution, which is also an extremely ambitious theory in the sense that it aims to explain all parts of biology with just a few very basic tools. It is conceptually possible to curtail this theory: perhaps it explains only 70% of life, leaving some other mechanism to explain the rest, or perhaps it explains only non-human life, leaving some deity to fully explain us. This kind of curtailed view would of course ignore the mountain of evidence there is for evolution in absolutely all parts of life (a point we will revisit in a moment), but it would also miss something about the kind of theory that the theory of evolution is. It seems that, as an explanation, evolution is so powerful that it would be incredible that something else would be equally able to explain life.

Whereas it cannot be stipulated that the theory of evolution is true universally, it can be argued that if it is true, it is true everywhere. To see this, consider that if incontrovertible evidence was found that evolution does not explain, say, the eight eyes of most spiders, then for most people that would cast aspersions on the theory of evolution in all other areas—even where it is backed up with overwhelmingly strong evidence. This is not simply to say that some recalcitrant evidence lowers the posterior probability of the theory somewhat, but rather that it would begin to completely undermine the theory. It seems the theory of evolution posits such a fundamental mechanism that anything short of universal quantification would invalidate it.

Perhaps we can describe what goes on in terms of "explaining away" (Pearl 1988). Imagine, for example, that one night the electricity in your house cuts out. You consider two hypotheses: that a possum has torn down the power line to your house, or that the whole neighbourhood has blacked out due to the recent heat wave. Out in the street you see other people checking their fuse boxes and this evidence favours the second hypothesis. Importantly, this evidence considerably lowers the probability of the possum hypothesis even though the two hypotheses could be true together. There is debate about what explaining away really is, but agreement that it exists. Part of what grounds this notion is that our background knowledge of the frequency of events tells us that it would be rather an unusual coincidence if, just as the overall power goes out due to the heat wave, a possum caused the line to go down (unless possums are known to take to power lines during heat waves). In the case of the deity hypothesis and the evolutionary hypothesis, it seems that explaining away is particularly strong. It would be an utterly astrounding coincidence if something as fundamental as speciation and adaptation had two coinciding explanations.

After this excursion into philosophy of science, we can return to the free energy principle. Though it still has nothing like the amount of evidence in its favour that evolution has, it seems that if it is true then it too must apply everywhere, and if not then it must be false. There is no middle way. This again seems to relate to explaining away. It would be too much of a coincidence if two explanations both accounted for something as fundamental as the organism's ability to sustain itself in its expected states. If the principle was directed at only fairly superficial aspects of mentality, such as the nature of visual illusions, then it would not strongly explain away other theories. But this misrepresents how deep the explanatory target actually is.

The issue was whether the explanatory ambition of the free energy principle can be curtailed, in order to make it seem less preposterous. If it is assumed that explaining away is particularly strong for fundamental rather than superficial explanations, then it appears that a principle as fundamental as the free energy principle cannot be curtailed. If it is believed, then it is believed with maximal scope. It is therefore misguided to think that one can take a divide and conquer approach to the free energy principle.

Of course, this can be taken to cement its preposterousness. If it is a hypothesis designed to be universal, then how can it be anything but preposterous? The immediate answer to this lies in comparing it again to the theory of evolution. This venerable theory must be preposterous in just the same way, but of course it isn't—it is true. This means that the issue whether the free energy principle is preposterous cannot be decided just by pointing to its explanatory ambition, since this would also invalidate the theory of evolution. Not surprisingly, it must be resolved by considering the evidence in favour of the free energy principle. As mentioned, this does not yet compare to that of the theory of evolution, though it is noteworthy that evidence is coming in, and that it is coming in from research on a comfortably large suite of mental phenomena.

Consider next the question of what happens with existing, competing theories once something like the free energy principle or the theory of evolution begins to gain explanatory force. Existing theories may have considerable evidence in their favour (this may be a theory about a cognitive or perceptual domain, such as attention or illusion); and they may explain away the existing evidence relatively well and therefore have that evidence in their favour (this contrasts with the comparison with the deity hypothesis, which strictly speaking has no evidence in its favour). Nevertheless, once additional, relevant evidence becomes available, existing theories may begin to lose ground to a new theory, like the free energy principle, even if it as yet has less evidence in its favour. For example, once it is noted that the brain is characterized by plentiful backwards connections, it becomes clear that these must be relevant to phenomena like attention and illusion (for example, disrupting them disrupts attention and illusion). However, if existing theories cannot explain this new evidence, then a new theory can begin to usurp their explanatory job. This means the evidence in their favour begins to wane, even if the new theory is still only enjoying spotty support. Compare again to the electricity blackout example. There might be a very impressive theory of the whereabouts and heat wave-related behavior of possums that very snugly explains the blackout in the house and perhaps other things besides. But the moment we become aware that the whole neighborhood is without electricity, even a poor theory of the blackout that can also address this new evidence ("perhaps it is some central distributor thingamajig that has broken down") becomes much more attractive than the existing possum theory. New evidence and new theories can very quickly wreak havoc on old, cherished theories. The free energy principle should therefore be expected to usurp the explanatory jobs of existing theories, and thereby challenge them, even if it is still a fairly fledgling theory. Of course, explanatory usurpation depends on acknowledging the occurrence of new evidence, such as the presence of backwards connections in the brain. Perhaps it is no surprise that the free energy principle is beginning to gain ground just as imaging brain science is maturing beyond the phase in which it was concerned mainly with collecting new evidence, and on to a new phase in which researchers consider the theoretical significance of the evidence in terms of both functional specificity and effective connectivity.

6 Predictions, distinctness, fecundity

It will be useful to discuss a concrete example of explanatory contest for the free energy principle. A good example comes from Ned Block & Susanna Siegel (2013) who argue against Andy Clark's (2013) version of the predictive processing framework in a way that pertains to the preceding remarks about explanatory prowess and ambition. In a comparison with an existing theory of attentional effects (proposed by Marisa Carrasco), they argue first that the predictive framework makes false predictions, and second that it offers no distinctive explanations.

As to the first point, Block and Siegel consider the effect where covert attention to a weak contrast grating enhances its perceived contrast. They argue that this increased contrast should be unexpected and therefore should elicit a prediction error that in turn should be extinguished, thereby annihilating the perceptual effect that the account was meant to explain in the first place. However, their argument does not rely on the correct version of the free energy account of attention. Block and Siegel overlook the fact that attention is itself predictive, in virtue of the prediction of precision. This means that attention enhances the prediction error from the weak grating, which in turn is explained away under the hypothesis that a strong contrast grating was present in that location of visual space. This conception of attention thus does yield a satisfactory account of the phenomenon that they claim cannot be explained (attentional enhancing), and it does not generate the false predictions they suggest (Hohwy 2013).

Block and Siegel's second point is more difficult to get straight. They argue that the predictive account offers no explanation of attentional findings, in particular relating to receptive field distortions; they then suggest that the account could adopt the existing theory, which asserts that "representation nodes" have shrinking receptive fields. They continue to argue that since the purported prediction error gain relates to error units in the brain rather than representation nodes, the prediction error account cannot itself generate this explanation. The argument is then that if the prediction processing account simply *borrows* that explanation (namely the existing explanation in terms of representation nodes), it hasn't offered anything distinctive. Again, this rests on an incorrect reading of the free energy account: error units are not insulated from representation units. Error units receive the bottom-up signal and this leads to revision of the predictions generated from the representation units. The outstanding question is how the distortions of receptive fields can be explained within the prediction error account.

This question has been addressed within the predictive coding literature. Thus Spratling (2008), who is a proponent of predictive coding accounts of attention, says (referring to the literature on changing receptive fields to which Block and Siegel themselves appeal) “the [predictive processing] model proposes, as have others before, that the apparent receptive field distortion arises from a change in the pattern of feedforward stimulation received by the cell”. That is, increased gain explains the distortion of the receptive field.

In fact, one might speculate that the predictive processing story makes perfect sense of the existence of modulable receptive fields. The receptive field of a given representational unit would, that is, be a function of the prediction error received from below, where—as described earlier—lower levels operate at smaller spatiotemporal scales. To give a toy illustration, assume that a broad receptive field would receive an equal amount of error signal from ten lower units each with smaller receptive fields, whereas a narrow receptive field receives error only from two such units. For the broad receptive field, if the gain on error from lower unit numbers one and two increases due to attention, then the gain on the other eight units decreases (since weights sum to one). Now, the hitherto broad receptive field mainly receives error from two lower units, so its receptive field has automatically shrunk. Attentional effects thus track the effects of expected precisions.

Here a more specific point can be made about Block and Siegel’s argument. The predictive processing account of attention can potentially offer a distinctive explanation of rather finegrained attentional findings. There is also reason to think that this explanation has more promise than existing theories. This is because the existing theories help themselves to the notion of ‘representational nodes’ whereas the free energy principle explains what these are, what they do, and how they connect with other nodes. Moreover, the prediction error account can deal very elegantly for key receptive field properties (Rao & Ballard 1999; Harrison et al. 2007).

This seems to be a good example of the situation outlined earlier with respect to the contest between the free energy principle and existing theories. The free energy principle can explain more types of evidence, under a more unificatory framework, and this immediately begins to undermine existing theories. Specifically, the theory that has no role for prediction error in receptive field modulation and activation only in representation nodes is explained away, even if it has significant evidence in its favour.

Underlying this story, there are some larger issues in the philosophy of science. One issue concerns the role of unification in explanation (Kitcher 1989). This is the idea that there are explanatory dividends in explanations that unify a variety of different phenomena under one theory. Obviously the free energy principle is a strong, ambitious unifier (perception, action, and attention all fall under the principle). Whereas there is discussion about whether this in itself adds to its explanatory ability as such, the ability to unify with other areas of evidence is part of what makes an explanation *better* than others. Noting this aspect of the free energy principle therefore supports it, in an inference to the best explanation (Lipton 2004, 2007). Confronted with a piecemeal explanation of a phenomenon and a unificatory explanation of the same phenomenon, the inference to the latter is stronger. There may be some difficult assessments concerning which explanation best deals with the available evidence. In the case discussed above, the free energy principle can explain less of the attention-specific evidence than the piecemeal explanation, but on the other hand it can explain more kinds of evidence, it provides explanatory tools that are better motivated (roles of representation and error 725 units), and it offers a more unifying account overall.

A second issue from the philosophy of science, in particular concerning inference to the best explanation, is the fecundity of an explanation, which is regarded as a best-maker. The better an explanation is at generating new predictions and ways of asking research questions, the stronger is the inference in its favour. Whereas this is not on its own a decider, it is an important contributor to the comparison of explanatory frameworks. Block and Siegel also seem to suggest that the predictive framework has nothing new to offer, or at least very little compared to existing (piecemeal) theories. Their example of a piecemeal theory is Carrasco's impressive work on attention, which has proven extraordinarily fecund, leading to a series of discoveries about attention. Assessing which theory is the more fecund is difficult, however, and involves considerations of unification. The free energy principle, as described above, does not posit any fundamental difference between perception and action. Both fall out of different reorganisations of the principle and come about mainly as different directions of fit for prediction error minimization (Hohwy 2013, 2014). This means that optimization of expected precisions, and thereby attention, must be central to action as well as to perception. This provides a whole new (and thus fecund) source of research questions for the area of action, brought about by viewing it as an attentional phenomenon. Important modeling work has been done in this regard (Feldman & Friston 2010), age-old questions (such as our inability to tickle ourselves) have been re-assessed (Brown et al. 2013), and new evidence concerning self-tickle has been amassed (Van Doorn et al. 2014). Theoretically, this has led to the intriguing idea that action occurs when attention is withdrawn from current proprioceptive input (described above). This idea points to a fully integrated view of attention, where attention is ubiquitous in brain function (with deep connections to consciousness, Hohwy 2012).

There is thus fecundity on both sides of this debate. It is difficult to conclusively adjudicate which side is more fecund, in part because the new research questions are in different areas and with different theoretical impact. It is surprising to be told that too much attention can undermine acuity—which is an example from Block and Siegel—but it is also surprising to be told that action is an attentional phenomenon.

The third issue from the philosophy of science concerns theory subsumption. It would be very odd if the explanations associated with the free energy principle (e.g., that attention is optimization of expected precision) completely contradicted all existing, more piecemeal explanations of attention. It should be expected that explanations of attention have some overlap with each other, as they are explaining away overlapping bodies of evidence. Indeed, the free energy explanation seems to subsume elements of biased competition theories of attention, as well as elements of Carrasco's theory, as seen above. This raises the question of to what extent a new theory, like the free energy principle's account of attention, really contributes a new and better understanding, especially if it carries within it elements of older theories. One way to go about this question again appeals to inference to the best explanation. The new and the old theories overlap in some respects, but they differ in respect of further elements of unification, theoretical motivation, broadness, fecundity, and so on. It can be difficult to come up with a scheme for precise assessment of these features, but it seems not unreasonable to say that the free energy principle performs best on at least those further elements of what makes explanations best.

At this stage it is tempting to apply the free energy principle to itself. This is an apt move since the idea of the hypothesis-testing brain arose in comparison with scientific practice (Helmholtz 1867; Gregory 1980). On this view, the point of a good scientific theory is to minimize prediction error as well as possible, on average and in the long run. This imputes an overall weighting of all the very same elements to science as we have ascribed to the brain above: revise theories in the light of evidence, control for confounds by making experimental manipulations, be guided by where highly precise evidence is expected to be found, adopt simple theories that diverge minimally from old theories, and let theories have a hierarchical structure such that they can persist in the face of non-linearities (due to causal interactions) in the evidence. All of these considerations speak in favour of the free energy principle over piecemeal, existing theories. By absorbing and revising older theories under the hierarchically imposed scientific "hyperparameter" of the free energy principle, it seems a very reasonable weighting of all these aspects can be achieved. For example, aspects of Carrasco's theory are subsumed, but under revised accounts of its notions of the functional role of representation nodes; due to the hierarchical aspect it is able to account for evidence arising under attentional approaches to action; in addition, this subsumption may be fecund, since we could expect it to lead to new findings in action (for example, a prediction that there will be attentional enhancement in the sensorimotor domain, leading to "illusory action").

7 The triviality worry

There is a different worry about preposterousness, also related to the issue of evidence. This worry is that the free energy principle is so general that anything the brain does can be construed as minimizing its prediction error. This is most clearly seen once the idea is cast in Bayesian terms. The brain harbours priors about the causes in the environment, and it calculates likelihoods that it combines with the priors to arrive at posterior probabilities for the hypotheses in question. One way to make this story apply to a particular case is to ascertain what is believed and then in a retrodictive fashion, posit priors and likelihoods accordingly unto the brain in question. If this can always be done, then the theory is trivialised by "just-so" stories and explains nothing. It is then preposterous because it pretends to be fundamental but is just trivial.

This triviality worry alerts the defender of the free energy principle to some pitfalls, but it is not a critical worry. To see this, an appeal can again be made to the theory of evolution. It is clear that when described in very general terms, anything can be described as enhancing fitness. For example, in an infamous hoax, Ramachandran gave a ridiculous, just-so adaptationist account of why gentlemen prefer blondes (Ramachandran & Blakeslee 1998), which some reportedly took seriously. Yet, no one serious thinks this invalidates the theory of evolution. The reason is, to repeat, that there is abundant solid, non-trivial evidence in favour of evolution. In other words, the presence of just-so triviality at some level of description can coexist with non-trivial explanations at the level of detailed, quantifiable evidence. Therefore the free energy principle cannot be invalidated just because it invites just-so stories. Of course, it is then hostage to translation into more precise, constricted applications to various domains, where predictions can be quantified and just-so stories avoided. Though there is nowhere near the same evidence that we have for the theory of evolution, evidence of this sort is becoming available (some is reviewed in Hohwy 2013).

Whereas the triviality worry does not invalidate the free energy principle, it does alert to some pitfalls. In particular, when forming hypotheses and when explaining phenomena in Bayesian terms, priors should not be stipulated independently of other evidence. If there is independent reason for asserting a prior with an explanatory role, then it is less likely that this prior is part of a just-so story. Similarly, discovery and manipulation of priors has a particularly important role in the defence of the free energy principle as applied to perception. For example, there is independent evidence that we expect light to come more or less from above (Adams et al. 2004), that objects move fairly slowly (Sotiropoulos et al. 2011), and that we expect others to look at us (Mareschal et al. 2013). Once established on independent grounds, researchers are better able to appeal to such priors in other explanations. This then helps avoid the just-so pitfall.

The triviality worry was that *everything* we do can be made to fit with the free energy principle. A different worry is that almost *nothing* we do fits with the free energy principle. If the free energy principle basically says the brain is an organ that tries to slow down the causal impact upon it from the world, then why don't organisms with brains just seek out sensory deprivation such as dark, silent rooms (Friston et al. 2012)? This dark room problem is aired very often and is natural on first thought when considering prediction error minimization. However, it also rests on a fundamental misreading of the free energy principle. The principle is essentially about maintaining the organism in its expected states, homeostatically defined, on average and in the long run. Locking oneself up in a dark silent room will only produce transitory free-energy minimization, as the demands of the world and the body will not be avoided for long. Soon, action is required to seek food, and soon the local council will come round to switch off the gas. It is much better for the brain to harness the deep model of the world in order to control its movement through the environment and thereby maintain itself more efficiently in its expected states.

Notice that this point harks back to a very basic hyperprior mentioned above—namely that the world is a changing place so that occupying the same state for too long will incur increasing free energy costs. This means that even if you currently have the prior that sensory deprivation is the right strategy for minimizing free energy, and even if this strategy works initially (as it does after a long and stressful day), that prior will decrease in strength as time goes by—leading to action and thus escape from sensory deprivation.

This response to the dark room problem in fact has a parallel in evolutionary theory. It has been argued that the free energy principle is false, essentially because not every action contributes directly to instantaneous prediction-error minimization and, analogously, it could be objected that evolutionary theory is false because not every trait directly contributes to instantaneous fitness. But of course this is a poor objection because fitness is measured over longer timescales and some traits, such as spandrels, contribute indirectly to fitness.

This and the preceding two sections have considered whether the explanatory ambition of the free energy principle is preposterous. By comparing the principle with the theory of evolution, and casting the worry in terms of philosophy of science, it can be seen that the explanatory ambition is not preposterous in and of itself. The verdict on the principle must come down to the quality of the explanations it offers and the amount of evidence in its favour. The principle is bound to be controversial, however, because it strongly explains away competing theories.

Of course, there are further issues to explore regarding the analogy between the free energy principle and the theory of evolution, and no doubt the analogy will have its limits. One interesting issue concerns the possibility of theory revision and thereby the possibility that the original statement of a theory is strictly speaking, false, even if it is one of those theories with extreme explanatory scope. The notion of natural selection as the only mechanism behind evolution is, for example, put under pressure by the discovery of genetic drift. This has led to revision of the theory of evolution, to encompass drift. Could something similar happen to the free energy principle, or is it in effect so ambitious that it is unrevisable? Conversely, is there any conceivable evidence that could falsify the current version of the theory in a wholesale fashion, rather than the piecemeal, detailed fashion discussed above? There are various answers available here, all of which reflect the peculiar theory emerging from the free energy principle.

First, the current form of the principle itself results from a long series of revisions of the basic idea that the brain engages in some kind of inference. Helmholtz' and Ibn Al Haytham's original ideas (reviewed briefly in Hohwy 2013) have been greatly revised, particularly in response to the mathematical realisation that the inversion of generative models presents an intractable problem, thus calling for variational Bayesian approaches to approximate inference. These developments occured partly in concert with the empirical discovery that the brain, as mentioned above, is characterized by massive backwards connectivity. It is then not unreasonable to say that older feed-forward versions of computational, information theoretical (e.g., infomax) theories of cognition constitute earlier versions of the free energy principle and that the latter is a revision in the light of formal and empirical discovery. The analogy with theory of evolution can thus be maintained in at least this backwards-looking respect.

Second, a more forward-looking example concerns the nature of the backwards connectivity in the brain. The free energy principle deems these descending signals *predictions,* but crucially it needs them to be of two kinds, namely predictions of the means of the underlying level's representations, and, as mentioned briefly above, predictions of the precisions of the underlying representations (thus encompassing sufficient statistics). There is some direct and some circumstantial evidence in favour of this dual role for descending signals, but the empirical jury is still out. Should it be found that descending signals do not mediate expected precisions, this would falsify the free energy principle. Notice that this falsification would be specific to the free energy principle, since the element of expected precisions is not found in some of the much broader theories in the academic marketplace that seem to countenance a predictive element in cognition. Notice also that a failure to identify top-down expectations of precision would amount to a wholesale falsification of the principle, since these "second-order" expectations are crucial not only for perception but also for action and action initiation (as explained above).

Third, and speaking much more generally, the principle would be falsified if a creature was found that did not act at all to maintain itself in a limited set of states (in our changing world). Such a creature should not on average and over time change its model parameters or active states and yet it would be able to prevent itself from being dispersed with equal probability among all possible states. This is a clear notion of a strong falsifier, and it speaks to the beauty of the free energy principle since it showcases its deep link between life and mind. However, it is not a very feasible falsifier because there is significant doubt that we would classify such a 'creature' as being alive or being a creature at all. Consider, for example, that a simple rock would serve as a falsifier in this sense since it is maintained on average and over the long run (that is, its states do not immediately disperse). One possibility here (Friston 2013) is to require that the scope is restricted to creatures that are space-filling, that is, who visits the individual states making up their overall set of expected states. A falsifier would then be a creature that manages to be space-filling but who does not manage this by changing its internal and active states via variational Bayes.

One nice question, in all of this, is whether the theory of evolution and the free energy principle can co-exist—and if so, how. This is a substantial issue, and a pertinent one, since both theories are fundamental and pertain to some of the same aspects—such as morphology, phenotypes, and life. Here is not the place to try to answer this interesting question, though inevitably some initial moves are made that might start to integrate them.

8 How literally is the brain Bayesian?

Bayes' rule is difficult to learn and takes considerable conscious effort to master. Moreover, we seem to flout it with disturbing regularity (Kahneman et al. 1982). So it is somewhat hard to believe that the brain unconsciously follows Bayes' rule. This raises questions about how literally we should think of the brain as a Bayesian hypothesis-tester. In blog correspondence, Lisa Bortolotti put the question succinctly:

> Acknowledging that prior beliefs have a role in perceptual inference, do we need to endorse the view that the way in which they constrain inference is dictated by Bayes' rule? Isn't it serendipitous that something we came up with to account for the rationality of updating beliefs is actually the way in which our brain unconsciously works?

Part of the beauty of the free energy principle is that even though it begins with the simple idea of an organism that acts to stay within expected states, its mathematical formulation forces Bayesian inference into the picture. Expected states are those with low surprisal or self-information. That is they have high probability given the model (low negative log probability).

These states cannot be estimated directly because that would require already knowing the distribution of states one can be in. Instead it is estimated indirectly, which is where the free energy comes in. Free energy, as mentioned above, is equal to the surprisal plus the divergence between the probability of the hypothesis currently entertained by the brain's states and the true posterior of the hypothesis given the model and the state. This much follows from Bayes' rule itself. This means that if the brain is able to minimize the divergence, then the chosen hypothesis becomes the posterior. This is the crucial step, because a process that takes in evidence, given a prior, and ends up with the posterior probability, as dictated by Bayes, must at least implicitly be performing inference (Friston 2010).

Hence, if the free energy principle is correct, then the brain must be Bayesian. How should this be understood? Consider what happens as the divergence is minimized. Formally this is a Kullback-Leibler divergence (or cross entropy), which measures the dissimilarity between two probability distributions. The KL-divergence can be minimized with various minimization schemes, such as variational Bayes. This plays an important role in machine learning and is used in simulations of cognitive phenomena using the free energy principle. Given the detail and breadth of such simulations, it is not unreasonable to say that brain activity and behavior are describable using such formal methods.

The brain itself does not, of course, know the complex differential equations that implement variational Bayes. Instead its own activity is brought to match (and thereby slow down) the occurrence of its sensory input. This is sufficient to bring the two probability distributions closer because it can only do this if it is in fact minimizing prediction error. This gives a mechanistic realization of the hierarchical, variational Bayes. The brain is Bayesian, then, in the sense that its machinery implements Bayes not serendipitously but necessarily, if it is able to maintain itself in its expected states. (There is discussion within the philosophy of neuroscience about what it means for explanations to be computational. See papers by Piccinini 2006, Kaplan 2011, Piccinini & Scarantino 2011, Chirimuuta 2014.)

The notion of realization (or implementation, or constitution) is itself subject to considerable philosophical debate. A paradigmatic reading describes it in terms of what plays functional roles. Thus a smoke alarm can be described in terms of its functional role (*i.e.*, what it, given its internal states, does, given a certain input). The alarm has certain kinds of mechanisms, which realize this role. This mechanism may comprise radioactive ions that react to smoke and causes the alarm to sound. The analogy between the smoke alarm and the brain seems accurate enough to warrant the paradigmatic functionalist reading of the way neuronal circuitry implements free energy minimization and therefore Bayes. Perhaps it is in some sense a moot point whether the ions in the smoke alarm "detect smoke" or whether they should merely be described in terms of the physical reactions that happen when they come into contact with the smoke particles. Rather than enter this debate it seems better to return to the point made at the start, when the brain was compared to other organs such as the heart. Here the point was that it is wrong to retract the description of the heart as a blood pump when we are told that no part of the cardiac cells are themselves pumps. The brain is literally Bayesian in much the same sense as the heart is literally a pump.

Behind this conceptual point is a deeper point about what kind of theory the free energy principle gives rise to (the following discussion will be based on Hohwy 2014). As described above, the Bayesian brain is entailed by the free energy principle. Denying the Bayesian brain then requires denying the free energy principle and the very idea of the predictive mind. This is, of course, a possible position that one could hold. One way of holding it is to "go down a level" such that instead of unifying everything under the free energy principle, theories just describe the dynamical causal interactions between brain and world. This would correspond to focusing more on systematic elements in the realization than in the function (looking at causal interaction between the heart and other parts of the body, and the individual dynamics of the cells making up the heart, rather than understanding these in the light of the heart being a pump). Call this the "causal commerce" position on the brain. Given the extensive and crucial nature of causal commerce between the brain and the world, this is in many ways a reasonable strategy. It seems fair to characterize parts of the enactive cognition position on cognitive science as informed primarily by the causal commerce position (for a comprehensive account of this position, see Thompson 2007; for an account that brings the debate closer to the free energy principle, see Orlandi 2013).

From this perspective, the choice between purely enactive approaches and inferential, Bayesian approaches becomes methodological and explanatory. One key question is what is accomplished by re-describing the causal commerce position from the more unified perspective of the free energy principle. It seems that more principled, integrated accounts of perception, action, and attention then become available. The more unified positioin also seems to pull away from many of the lessons of the enactive approach to cognition, because the free energy principle operates with a strict inferential veil between mind and world—namely the sensory evidence behind which hidden causes lurk, which must be inferred by the brain. Traditionally, this picture is anathema to the enactive, embodied approaches, as it lends itself to various forms of Cartesian skepticism, which signals an internalist, secluded conception of mind. A major challenge in cognitive science is therefore to square these two approaches: the dynamical nature of causal commerce between world, body, and brain and the inferential free energy principle that allows their unification in one account. On the approach advocated here, modulo enough empirical evidence, denying that the free energy principle describes the brain is on a par with denying that the heart is a pump. This means that it is not really an option to deny that the brain is inferential. This leaves open only the question of *how* it is inferential.

One line of resistance to subsuming everything under the free energy principle has to do with intellectualist connotations of Bayes. Somehow the idea of the Bayesian brain seems to deliver a too regularized, sequential, mathematical desert landscape—it is like a picture of a serene, computational mechanism silently taking in data, passing messages up and down the hierarchy, and spitting out posterior probabilities. This seems to be rather far from the somewhat tangled mess observed when neuroscientists look at how the brain is in fact wired up. In one sense this desert landscape is of course the true picture that comes with the free energy principle, but there need be nothing serene or regularized about the way it is realized. The reason for this goes to the very heart of what the free energy principle is. The principle entails that the brain recapitulates the causal structure of the world. So what we should expect to find in the brain will have to be approximating the far-from-serene and regularized interactions that occur between worldly causes. Just as there are non-linearly interacting causes in the world there will be convolving of causes in the brain; and just as there are localized, relatively insulated causal "eddies" in the world there will be modularized parameter spaces in the brain.

Moreover, there is reason to think the brain utilizes the fact that the same causes are associated with multiple effects on our senses and therefore builds up partial models of the sensorium. This corresponds to cognitive modules and sensory modalities allowing processing in conditionally independent processing streams, which greatly enhances the certainty of probabilistic inference. In this sense the brain is not only like a scientist testing hypotheses, but is also like a courtroom calling different, independent witnesses. The courtroom analogy is worth pursuing in its own right (Hohwy 2013), but for present purposes it supports the suggestion that when we look at the actual processing of the brain we should expect a fairly messy tangle of processing streams. (Clark 2013 does much to characterize and avoid this desert landscape but seems to do so by softening the grip of the free energy principle.)

9 Functionalism and biology

So far the free energy principle has been given a functionalist reading. It describes a functional role, which the machinery in the brain realizes. One of the defining features of functionalism is that it allows multiple realization. This is the simple idea that the same function can be realized in different ways, at least in principle. For example, a smoke alarm is defined by its functional role but can be realized in different ways. There is on-going debate about whether something with the same causal profile as the human brain could realize a mind. Philosophers have been fond of imaging, for example, a situation in which the population of Earth is each given a mobile phone and a set of instructions about whom to call and when, which mimics the "instructions" followed by an individual neuron (Block 1976). The question then is whether this mobile phone network would be a mind. Though this is not the place to enter fully this debate, it seems hard for the defender of the free energy principle to deny that, if these mobile phone-carrying individuals are really linked up in the hierarchical message-passing manner described by the equations of the free energy principle, if they receive input from hidden causes, and if they have appropriate active members, then they do constitute a mind.

However, a different issue here is to what extent the free energy principle allows for the kind of multiple realization that normally goes with functionalism. The mathematical formulations and key concepts of the free energy principle arose in statistical physics and machine learning, and hierarchical inference has been implemented in computer learning (Hinton 2007). So there is reason to think that prediction error minimization can be realized by computer hardware as well as brainware. There is also reason to think that within the human brain the same overall prediction error minimization function can be realized by different hierarchical models. Slightly different optimizations of expected precisions would determine the top-down vs. bottom-up dynamics differently, but may show a similar ability to minimize prediction error over some timeframes. Different weightings of low and high levels in the hierarchy can lead to the same ability to minimize prediction error in the short and medium term. This is similar to how a dam can be controlled with many small plugs close to the dam wall, or by fewer connected dam locks operating at longer timescales further back from the dam wall. In some cases, such different realizations may have implications for the organism over the long run, however (for example, building locks in a dam may take time, and thus allow flows in the interim; whereas many small plugs prevent flows in the short run term but may be impractical in the long run). Such differences may show up in our individual differences in perceptual and active inference (for an example, see Palmer et al. 2013), and may also be apparent in mental illness (Hohwy 2013, Ch. 7).

Functionalist accounts of the mind are widely discussed in the philosophical literature, and there are various versions of it. A key question for any functionalism is how the functional roles are defined in the first instance (for an overview see Braddon-Mitchell & Jackson 2006). Some theories—psychofunctionalism or empirical functionalism—posit that functional roles should be informed by best empirical science ("pain is caused by nociceptor activation… etc."). The consequence is that their domain is restricted to those creatures for whom that empirical science holds. Other theories—commonsense functionalism—begin with conceptual analysis and use that to define the functional roles ("pain is the state such that it is caused by bodily damage, gives rise to pain-avoidance behavior, and relates thus and so to internal states…"). The consequence of taking the commonsense approach is that such functionalisms apply widely, including to creatures science has never reached, in so far as they have something realizing that functional role.

There are some nice questions here about what we should really say about creatures with very different realizations of the same functions (e.g., "Martian pain"), and creatures with very similar realizations but different functions (e.g., "mad pain"; see Lewis 1983). Setting those issues aside for the moment, one question is which kind of functionalism goes with the free energy principle. There is no straightforward answer here, but one possibility is that it is a kind of “biofunctionalism”, where the basic functional role is that of creatures who manage to maintain themselves within a subset of possible states (in a space-filling or active manner) for a length of time. Any such creature must be minimizing its free energy and hence engaging in inference and action. It is biological functionalism because it begins by asking for the biological form—the phenotype—of the candidate creature.

This is an extremely abstract type of functionalism, which allows considerable variation amongst phenotypes and hence minds. For example, it has no problem incorporating both Martians and madmen in so far as they maintain themselves in their expected states. It will however specify the mental states of the organism when it becomes known in which states it maintains itself. This follows from the causal characterization of sensory input, internal states, and active output that fully specify a prediction error minimizing mechanism. Once these states are observed, the states of the system can be known too, and the external causes rendered uninformative (i.e., the sensory and active states form a Markov blanket; Friston 2013).

What drives biofunctionalism is not species-specific empirical evidence, as in psychofunctionalism. And it does not seem to be commonsense conceptual analysis either. Rather, it begins with a biological, statistical observation that is as basic as one can possibly imagine—namely that creatures manage to maintain themselves within a limited set of states. As seen at the very start of this paper, this defines a probability density for a given creature, which it must approximate to do what it does. For an unsupervised system, it seems this can only happen if the organism minimizes its free energy and thereby infers the hidden causes of its sensory input, and then acts so as to minimize its own errors. This is an empirical starting point at least in so far as one needs to know many empirical facts to specify which states a creature occupies. But it is, arguably, also a conceptual point in so far as one hasn’t understood what a biological creature is if one does not associate it at least implicitly with filling some specified subset of possible states.

The upshot is that the free energy principle sits well with a distinct kind of functionalism, which is here called biofunctionalism. It remains an open question how this would relate to some versions of functionalism and related views, such as teleosemantics (Neander 2012), which relies on ideas of proper function, and information theoretical views (Dretske 1983). The biofunctionalism of the free energy principle seems to have something in common with those other kinds of positions though it has no easy room for the notion of proper function and it doesn’t rely on, but rather entails, information theoretical (infomax) accounts.

Setting aside these theoretical issues, note that biofunctionalism has a rather extreme range because it entails that there is Bayesian inference even in very simple biological organisms in so far as they minimize free energy. This includes for example *E. coli* that with its characteristic swimming-tumbling behavior, maintains itself in its expected states. And it includes us, who with our deeper hierarchical models maintain ourselves in our expected states (with more space-filling and for longer than *E. coli*). Of course, one might ask where, within such a wide range of creatures, we encounter systems that we are comfortable describing as minds—that is, as having thought, as engaging in decision-making and imagery, and not least as being conscious. This remains a challenge for the free energy principle, just as it is a challenge for any naturalist theory of the mind to specify where, why, and how these distinctions between creatures arise.

10 The neural organ can explain the mind

The brain is an organ with a function, namely to enable the organism to maintain itself in its expected states. According to the free energy principle, this is to say that it minimizes prediction error on average and over the long run. This is a controversial idea, with extreme explanatory ambition. It might be considered not only controversial but also preposterous. But the philosophy of science-based discussions above have sought to show that it is not in fact preposterous. The different ways in which it might be preposterous either do not apply, misunderstand the principle, or would also apply to the paradigmatically non-preposterous theory of evolution. The free energy principle yields a theory that should, indeed, strongly explain away competing theories. The free energy principle is an account that displays a number of explanatory virtues such as unification and fecundity. It is therefore not reasonable to detract from the principle by claiming it is preposterous or too ambitious. Scientifically speaking, what remains is to assess the evidence for and against the free energy principle and consider how, more specifically, it explains our mental lives (a task I undertake in Hohwy 2013). Speaking in terms of philosophy of mind, there remain questions about what type of functionalist theory the free energy principle is, how it performs vis-à-vis traditional questions about functionalism and the realizers of functional roles, and, finally, some more metaphysical questions about what it says about the nature of the mind in nature. None of these philosophical issues are apparently more damning for the free energy principle than they are for other, previously proposed accounts of the nature of the mind, and there is reason to think that with the free energy principle a new suite of answers may become available.

References

Adams, W. J., Graf, E. W. & Ernst, M. O. (2004). Experience can change the ‘light-from-above’ prior. *Nature Neuroscience*, *7* (10), 1057-1058. 10.1038/nn1312

Block, N. (1976). Troubles with functionalism. *Minnesota Studies in the Philosophy of Science*, *9*, 261-325.

Block, N. & Siegel, S. (2013). Attention and perceptual adaptation. *Behavioral and Brain Sciences*, *36* (3), 205-206. 10.1017/S0140525X12002245

Braddon-Mitchell, D. & Jackson, F. (2006). *The philosophy of mind and cognition: An introduction.* London, UK: Wiley-Blackwell.

Brown, H., Adams, R. A., Parees, I., Edwards, M. & Friston, K. (2013). Active inference, sensory attenuation and illusions. *Cognitive Processing*, *14* (4), 411-427. 10.1007/s10339-013-0571-3

Chirimuuta, M. (2014). Minimal models and canonical neural computations: The distinctness of computational explanation in neuroscience. *Synthese*, *191* (2), 127-153. 10.1007/s11229-013-0369-y

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral & Brain Sciences*, *36* (3), 181-204. 10.1017/S0140525X12000477

Dretske, F. (1983). *Knowledge and the flow of information.* Cambridge, MA: MIT Press.

Feldman, H. & Friston, K. (2010). Attention, uncertainty and free-energy. *Frontiers in Human Neuroscience*, *4* (215), 1-23. 10.3389/fnhum.2010.00215

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11* (2), 127-138. 10.1038/nrn2787

——— (2013). Life as we know it. *Journal of the Royal Society Interface*, *10* (86). 10.1098/rsif.2013.0475

Friston, K., Thornton, C. & Clark, A. (2012). Free-energy minimization and the dark room problem. *Frontiers in Psychology*, *3* (130), 1-7. 10.3389/fpsyg.2012.00130

Friston, K. & Stephan, K. (2007). Free energy and the brain. *Synthese*, *159* (3), 417-458. 10.1007/s11229-007-9237-y

Gregory, R. L. (1980). Perceptions as hypotheses. *Philosophical Transactions of the Royal Society B, Biological Sciences*, *290* (1038), 181-197. 10.1098/rstb.1980.0090

Harrison, L. M., Stephan, K. E., Rees, G. & Friston, K. J. (2007). Extra-classical receptive field effects measured in striate cortex with fMRI. *NeuroImage*, *34* (3), 1199-1208. 10.1016/j.neuroimage.2006.10.017

Hinton, G. E. (2007). Learning multiple layers of representation. *Trends in Cognitive Sciences*, *11* (10), 428-434. 10.1016/j.tics.2007.09.004

Hobson, J. A. & Friston, K. J. (2012). Waking and dreaming consciousness: Neurobiological and functional considerations. *Progress in Neurobiology*, *98* (1), 82-98. 10.1016/j.pneurobio.2012.05.003

Hohwy, J. (2012). Attention and conscious perception in the hypothesis testing brain. *Frontiers in Psychology*, *3* (96), 1-14. 10.3389/fpsyg.2012.00096

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

——— (2014). The self-evidencing brain. *Noûs, Early View.* 10.1111/nous.12062

Kahneman, D., Slovic, P. & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases.* Cambridge, UK: Cambridge University Press.

Kaplan, D. M. (2011). Explanation and description in computational neuroscience. *Synthese*, *183* (3), 339-373. 10.1007/s11229-011-9970-0

Kitcher, P. (1989). Explanatory unification and the causal structure of the world. In P. Kitcher & W. Salmon (Eds.) *Scientific explanation* (pp. 410-505). Minneapolis, MN: University of Minnesota Press.

Lewis, D. (1983). Mad pain and martian pain. In D. Lewis (Ed.) *Philosophical papers, Vol. 1* (pp. 122-130). Oxford, UK: Oxford University Press.

Lipton, P. (2004). *Inference to the best explanation.* London, UK: Routledge.

——— (2007). Précis of Inference to the best explanation. *Philosophy and Phenomenological Research*, *74* (2), 421-423. 10.1111/j.1933-1592.2007.00027.x

Mareschal, I., Calder, A. J. & Clifford, C. W. G. (2013). Humans have an expectation that gaze is directed toward them. *Current Biology*, *23* (8), 717-721. 10.1016/j.cub.2013.03.030

Neander, K. (2012). Teleological theories of mental content. *The Stanford encyclopedia of philosophy*, *Spring 2012 Edition* E. N. Zalta (Ed.) http://plato.stanford.edu/archives/spr2012/entries/content-teleological/

Orlandi, N. (2013). Embedded seeing: Vision in the natural world. *Noûs*, *47* (4), 727-747. 10.1111/j.1468-0068.2011.00845.x

Palmer, C. J., Paton, B., Hohwy, J. & Enticott, P. G. (2013). Movement under uncertainty: The effects of the rubber-hand illusion vary along the nonclinical autism spectrum. *Neuropsychologia*, *51* (10), 1942-1951. 10.1016/j.neuropsychologia.2013.06.020

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference.* San Fransisco, CA: Morgan Kaufmann Publishers.

Piccinini, G. (2006). Computational explanation in neuroscience. *Synthese*, *153* (3), 343-353. 10.1007/s11229-006-9096-y

Piccinini, G. & Scarantino, A. (2011). Information processing, computation, and cognition. *Journal of Biological Physics*, *37* (1), 1-38. 10.1007/s10867-010-9195-3

Ramachandran, V. S. & Blakeslee, S. (1998). *Phantoms in the brain.* London, UK: Fourth Estate.

Rao, R. P. N. & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive field effects. *Nature Neuroscience*, *2* (1), 79-87. 10.1038/4580

Sotiropoulos, G., Seitz, A. R. & Seriés, P. (2011). Changing expectations about speed alters perceived motion direction. *Current Biology*, *21* (21), R883-R884. 10.1016/j.cub.2011.09.013

Spratling, M. W. (2008). Predictive coding as a model of biased competition in visual attention. *Vision Research*, *48* (12), 1391-1408. 10.1016/j.visres.2008.03.009

Thompson, E. (2007). *Mind in life: Biology, phenomenology, and the sciences of mind.* Harvard, MA: Harvard University Press.

Trommershäuser, J., Körding, K. & Landy, M. (Eds.) (2011). *Sensory cue integration.* Oxford, UK: Oxford University Press.

Van Doorn, G., Hohwy, J. & Symmons, M. (2014). Can you tickle yourself if you swap bodies with someone else? *Consciousness and Cognition*, *23*, 1-11. 10.1016/j.concog.2013.10.009

von Helmholtz, H. (1867). *Handbuch der Physiologischen Optik.* Leipzig, GER: Leopold Voss.

From Explanatory Ambition to Explanatory Power

A Commentary on Jakob Hohwy

Dominic L. Harkness

The free energy principle is based on Bayesian theory and generally makes use of functional concepts. However, functional concepts explain phenomena in terms of how they should work, not how they in fact do work. As a result one may ask whether the free energy principle, taken as such, can provide genuine explanations of cognitive phenomena. This commentary will argue that (i) the free energy principle offers a stronger unification than Bayesian theory alone (strong unification thesis) and that (ii) the free energy principle can act as a heuristic guide to finding multilevel mechanistic explanations.

Keywords
Active inference | Bayesian enlightenment | Bayesian fundamentalism | Bayesian theory | Free energy | Free energy principle | Functional | Mechanisms | Precision | Prediction errors | Preposterous | Strong unification thesis | Weak unification thesis

Commentator

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

Target Author

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

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 has far-reaching implications for cognitive science. In fact, the free energy principle seeks to explain everything related to the mind. Due to this explanatory ambition, it has been deemed preposterous by researchers. Jakob Hohwy challenges the opponents of the free energy principle and its applications by demonstrating that this framework is everything but preposterous. Rather, he compares the free energy principle with the theory of evolution in biology. The theory of evolution is not discarded due to its unifying power; and the free energy principle shouldn't be either. In this paper I will present a negative as well as two positive theses: first, the free energy principle will be contrasted to Bayesian theory with regard to the degree of unification they offer. I will argue that the unification resulting from the free energy principle can be regarded as stronger since it attempts to empirically ground its conclusions in the brain via neuroscience and psychology. The negative thesis consists in the suggestion that one major flaw of the free energy principle, taken as such, lies within its explanatory *power*. As a result of being a functional theory, the concepts it employs are also functional. Yet functional concepts, at least when it comes to explaining the brain and cognitive phenomena, do not explain how a certain phenomenon actually works, but rather how it should work. To improve this situation, the second positive thesis of this paper makes use of a suggestion by Piccinini & Craver (2011), namely that functional analyses are mechanism sketches, i.e., incomplete descriptions of mechanisms. In other words, functional concepts (such as precision) must be enriched with mechanistic concepts that include known structural properties (such as "dopamine") in order to count as a full explanation of a given phenomenon. The upshot of this criticism lies within the free energy principle's potential to act as a heuristic guide for finding multilevel mechanistic explanations. Furthermore, this paper will not advocate that functional concepts should be fully replaced or eliminated, but that functional and mechanistic descriptions complement each other.

2 The free energy principle

In his article "The Neural Organ Explains the Mind", Jakob Hohwy (this collection) proposes that the brain, as every other organ in the human body, serves one basic function. Just as one might say that the basic function of the heart is to pump blood through the body or that of the lungs is to provide oxygen, the basic function of the brain is to minimise free energy (Friston 2010). However, this is a very general claim that does not yet establish how the minimisation of free energy is realised in humans. How is this done?

Very generally, the brain stores statistical regularities from the outer environment or, in other words, it forms an internal model about the causal structure of the world. This model is then used to predict the next sensory input. Consequently, we have two values that can be compared with each other: the predicted sensory feedback and the actual sensory feedback. When perceiving, the brain predicts what its own next state will be. Depending on the accuracy of the prediction, a divergence will be present between the predicted and the actual sensory feedback. This divergence is measured in terms of prediction errors. The larger the amount of prediction error, the less accurately the model fits the actual sensory feedback and thus the causal structure of the world. Crucially, the model that fits best, i.e., that which brings forth the smallest amount of prediction error, also determines consciousness. In this framework, free energy amounts to the sum of prediction errors. Thus, minimizing prediction errors always entails the minimisation of free energy.

The minimization of prediction error can generally be achieved in two ways: either the brain can change its models according to the sensory input or, vice versa, it can change the sensory input according to its models. In this scheme the former mode can be seen as veridical perception, whereas the latter can be seen as action, or more formally active inference—the fulfillment of predictions via classic reflex arcs (Friston et al. 2009; Friston et al. 2011). Furthermore, two other factors play a large role in the minimization of prediction error: first, the precision, or "second-order statistics" (Hesselmann et al. 2012), which ultimately encodes how "trustworthy" the actual sensory input is. Precision is realised by synaptic gain, and it has been established that the modulation of precision corresponds to attention (Hohwy 2012). Second, model optimization ensures that models are reduced in complexity in order to account for the largest number of possible states in the long run, i.e., under expected levels of fluctuating noise. For example, sleep has been associated with this type of model optimization (Hobson & Friston 2012). More detailed descriptions of these four factors, i.e., perception, active inference, precision, and model optimization can be found in Hohwy's article.

Additionally, models are arranged in a cortical hierarchy (Mumford 1992). This hierarchy is characterised, as Hohwy points out (this collection, p. 7), by time and space: models higher up in the hierarchy have a larger temporal scale and involve larger receptive fields than models lower down in the hierarchy, which concern predictions at fast time scales and involve small receptive fields (p. 7). This hierarchy implies a constant message-passing amongst different levels. Once a sensory signal arrives at the lowest level it is compared to the predictions coming from the next higher level (in this case level two).[1] If prediction errors ensue they are sent to the higher level (still level two). Here they are predicted by the next higher level (now level three). This process goes on until prediction errors are minimised to expected levels of noise.

Now the general scheme of prediction error minimization can be presented: the brain builds models that represent the causal structure of the world. These models are, in turn, used to generate predictions about what the next sensory input might be. The two resulting values, i.e., the predicted and the actual sensory feedback, are continuously compared. The divergence between these two values is the prediction error, or free energy. Since it is the brain's main function to minimise the amount of free energy and therefore prediction error, it will either change its models or engage in active inference. Decisions about which path will be taken depend on the precision of the incoming sensory signal (or prediction error). Signals with high precision are taken to be "trustworthy", and therefore model changes can follow. Low precision signals, however, require further investigation since noise could be the principal factor in an ambiguous input. In addition, models during wakefulness are changed "on-the-fly", thus leading to highly idiosyncratic and complex models. This complexity is reduced, for example during sleep (Hobson & Friston 2012), to increase the generalizability of models, since noise is always present.

3 Bayesian theory and unification

As mentioned above, all this serves the basic function of the brain: the minimization of free energy. This strategy is employed in every aspect of cognition; thus the free energy principle (Friston 2010) is a grand unifying theory. But from where does the free energy principle derive its unifying power?[2]

The free energy principle makes use of Bayesian theory, which can be regarded as its foundation. For some years now, Bayesian theory has been applied to many cognitive phenomena, since it may "offer a new interpretation of the goals of cognitive systems, in terms of inductive probabilistic inference […][,] can make the assumptions of Bayesian models more transparent than in mechanistically oriented models […][and] may have the potential to explain some of the most complex aspects of human cognition […]" (Jones & Love 2011, p. 170). Yet Jones & Love (2011) also address the fact that Bayesian theories, although aiming at researching and investigating the human brain and its workings, remain unconstrained by psychology and neuroscience "and are generally not grounded in empirical measurement" (ibid., p. 169). They term this approach "Bayesian Fundamentalism", since it entails that all that is necessary to explain human behaviour is rational analysis. Supporters of this position rely on the mathematical framework of Bayesian theory as the origin of its explanatory power and unification. The positive thesis of Jones & Love (2011) consists in arguing for "Bayesian Enlightenment" that tries to include mechanistic explanation in Bayesian theory. To give more detail, they propose that, rather than following Bayesian Fundamentalism and thus being "logically unable to account for mechanistic constraints on behavior […] one could treat various elements of Bayesian models as psychological assumptions subject to empirical test" (Jones & Love 2011, p. 184). Similarly, Colombo & Hartmann (2014) argue that although "the Bayesian framework […] does not necessarily reveal aspects of a mechanism[,] Bayesian unification […] can place fruitful constraints on causal-mechanical explanation" (Colombo & Hartmann 2014, p. 1).

According to Colombo & Hartmann (2014), many Bayesian theorists falsely equate unification with explanatory power. But Bayesian theories derive their unificatory power

[1] The numerical values for the levels have no scientific relevance. They are used only for illustrative purposes.

[2] At this point I would like to thank one of the reviewers for her or his substantial advice and constructive comments.

from their mathematical framework. However, just because different cognitive phenomena can be mathematically unified does not entail a causal relationship between them, and nor does the mathematical unification tell us anything about the causal history of these phenomena. However, as will be presented in the next section, explanatory power, at least from a mechanistic point of view, results from investigating structural components and their causal interactions that give rise to a certain phenomenon. For example Kaplan & Craver (2011) write that "[…] the line that demarcates explanations from merely empirically adequate models seems to correspond to whether the model describes the relevant causal structures that produce, underlie, or maintain the explanandum phenomenon" (p. 602). Yet in the case of Bayesian theory—and Bayesian Fundamentalism in particular—, this cannot be achieved, since they "say nothing about the spatio-temporally organized components and causal activities that may produce particular cognitive phenomena […]" (Colombo & Hartmann 2014, p. 5). But not everything is lost concerning the explanatory role of Bayesian theories. Even if Bayesian theory cannot provide mechanistic explanations, it may nonetheless be beneficial to cognitive science by offering constraints on causal-mechanical explanation (Colombo & Hartmann 2014).

This brings us to the free energy principle. As noted, the free energy principle is, at its core, a theory that makes use of Bayesian theory; consequently it inherits all of Bayesian theory's pros and cons. Thus, since unification in the free energy principle is also grounded in its mathematical foundations "[…] the real challenge is to understand how [the free energy principle] manifests in the brain" (Friston 2010, p. 10). With regard to Jones & Love's (2011) distinction, the free energy principle can be considered to belong to Bayesian Enlightenment, since it attempts to ground its findings in neurobiology and psychology rather than remaining unconstrained by these sciences. Furthermore, due to the fact that the free energy principle integrates neuroscientific findings into its conclusions, it can offer more precise constraints on causal-mechanical explanations than Bayesian theory alone. For example, the free energy principle tries to incorporate neuroscientific facts about brain structure and its hierarchical organization, or tries to link concepts such as "precision" to neurophysiological phenomena such as "dopaminergic gating" (Friston et al. 2012).[3] The latter example will be presented in greater detail in section 5.

> In sum, the free energy principle offers a form of unification that exceeds that offered by Bayesian theory alone. It makes statements about how the free energy principle could be realised in the brain and does not solely rely on its mathematical framework. Thus, one could term the former a "strong unification thesis" (SUT) and the latter a "weak unification thesis" (WUT).

If the free energy principle is true it creates a backdrop against which other theories must be evaluated. This also implies a kind of explanatory monopolization, since "the free energy principle is not a theory that lends itself particularly well to piecemeal" (Hohwy this collection, p. 9). In other words, as Hohwy highlights on many occasions, the free energy principle is an all-or-nothing theory. He compares it to the theory of evolution in biology and states that, just like the free energy principle, "evolution posits such a fundamental mechanism that anything short of universal quantification would invalidate it" (p. 10). Due to this large explanatory ambition, some researchers have described the free energy principle as preposterous. Yet "the issue whether the free energy principle is preposterous cannot be decided just by pointing to its explanatory ambition […] [but] by considering the evidence in favour of the free energy principle" (p. 11). This is a very important transition, i.e., the switch from explanatory ambition to explanatory power, since, from a mechanistic viewpoint, the former gives no statement about the veridicality of its assumptions, whereas the latter does.

[3] However, I'd like to point out that the free energy principle does not make any commitments to one single neuroscientific theory. Rather, it tries to find entities that may realize the free energy principle in the brain; what these entities are remains to be inquired.

In the remainder of this paper, I will argue that one major shortcoming of the free energy principle lies in its explanatory *power*. The main issue to be discussed consists in the fact that most concepts employed in the free energy principle, or in its applications such as predictive coding (Friston 2005; Rao & Ballard 1999) or predictive processing (Clark 2013; Hohwy 2013), are principally functional concepts. Yet, at least in the case of the free energy principle, functional concepts do not hold much explanatory power, since they "describe how things ought to work rather than how they in fact work" (Craver 2013, p. 18). For example, the concept of "precision" represents the amount of uncertainty in the incoming sensory signal that may arise due to noise. Thus the precision of the incoming sensory inputs determines how an agent interacts with its environment next: it can either change its models or its sensory input. Yet, this description holds no commitments as to how precision is realised in the brain; it only describes what effect precision *should* have on a given cognitive system. Therefore the free energy principle seems to be of a normative, rather than descriptive, nature.[4] On the other hand, there are mechanistic explanations that, according to Craver (2007), can also count as such, since they don't describe how things should work but how they in fact *do* work.

Yet these two types of epistemic strategies don't necessarily exclude each other. Here I want to introduce Piccinini & Craver's (2011) claim that functional analyses can serve as "mechanism sketches". The upshot lies within the free energy-principle's unifying power: it can act as a kind of conceptual guide for revealing mechanistic explanations. Once physiological concepts are mapped onto the functional concepts derived from the free energy-principle, multilevel mechanistic explanations follow. But before this is elaborated the next section will give a short introduction to mechanistic explanation (Craver 2007).

4 Mechanistic explanation

Mechanistic explanation claims that in order "[t]o explain a phenomenon, [...] one has to know what its components are, what they do and how they are organized [...]" (Craver & Kaplan 2011, p. 269). It does not suffice to merely be able, e.g. to accurately predict a phenomenon. Craver & Kaplan (2011, p. 271) show this by referring to the example of a heat gauge on a car. Despite the fact that the gauge represents engine heat and that one can also predict when the engine will overheat by looking at the gauge, it doesn't explain why the engine is overheating. It only states that it is—not how it came about. Thus, mechanists introduced the "model-to-mechanism-mapping" (3M) requirement for explanatory models:

> (3M) A model of a target phenomenon explains that phenomenon when (a) the variables in the model correspond to identifiable components and organizational features of the target mechanism that produces, maintains, or underlies the phenomenon, and (b) the causal relations posited among these variables in the model correspond to the activities or operations among the components of the target mechanism. (Kaplan 2011, p. 272)

This requirement can serve as a demarcation criterion as to when a model can actually be seen as explanatory. But how does mechanistic explanation progress? Two principal approaches are described by Craver & Kaplan (2011): reductionism and integrationism. The former tries to reduce mental phenomena into ever-smaller entities. Its most radical form, "ruthless reductionism", is advocated by John Bickle (2003), who states that neuroscience should reduce "[...] psychological concepts and kinds to molecular-biological mechanisms and pathways" (Bickle 2006, p. 412). In other words, mental phenomena should be explained with low-level concepts. The integrationist approach, on the other hand, claims that explanations can be found across a hierarchy of mechanisms (Craver 2007), since every mechanism is itself embedded into a higher-level mechanism. Consequently, reductionism isn't the only option, since "[...] mechanistic explanation requires consideration not just of the parts and operations in the mechanism but also of the organization

[4] This does not mean that the free energy principle is false. On the contrary, this paper will present an attempt to increase its explanatory potential.

within the mechanism and the environment in which the mechanism is situated" (Bechtel 2009, p. 544). In particular, multilevel mechanistic explanations consider three viewpoints on any given mechanism: the etiological, constitutive, and contextual aspects (Craver 2013). At the etiological level, the causal history of a given mechanism is investigated at the same level of the hierarchy. Yet mechanisms can also be broken down into smaller, more specialised mechanisms. When investigating the internal mechanisms that give rise to a mechanism at a higher level, one can speak of the constitutive aspect of mechanistic explanation. This strategy resembles reductionism most. But, as mentioned before, every mechanism is also embedded in a higher-level mechanism. Thus, one must also investigate how a given mechanism contributes to the next higher-level mechanism. This has been termed the contextual aspect, because it situates a mechanism into a higher-order context. After this short introduction into mechanistic explanation, the next section will show how this relates to the problem above, i.e., that applications of the free energy principle operate with functional concepts and thus can't serve as full explanations.

5 The free energy principle as heuristic guide

Here I will follow Piccinini & Craver's (2011) proposal that functional descriptions are nothing other than mechanism sketches that derive their "[...] explanatory legitimacy from the idea that [they][...] capture something of the causal structure of the system" (Piccinini & Craver 2011, p. 306). Mechanism sketches are simply outlines of mechanisms that haven't been fully investigated with regard to their structural properties. Thus, functional descriptions serve as placeholders until a mechanistic explanation can fully account for a given phenomenon by enriching functional concepts with concepts related to its structural properties.[5] The explanatory gaps[6] resulting from the functional nature of the free energy principle could then be closed, leading to a shift from explanatory ambition to explanatory power. This also directly relates to the alleged preposterousness of the free energy principle, since the process of "filling-in" will diminish any residual doubts about the theory's truthfulness. This can be applied to the free energy principle, which works with functional concepts such as "precision", "prediction error", "model optimization" or "attention": "[o]nce the missing aspects are filled in, a functional analysis turns into a full-blown mechanistic explanation" (Piccinini & Craver 2011, p. 284). Take the concept of precision in the free energy principle as an example. As described above, precision gives an estimate concerning the "trustworthiness" of a given sensory signal and its ensuing prediction errors. Taken as such, precision is clearly a functional concept since it is "[...] specified in terms of effects on some medium or component under certain conditions" (Piccinini & Craver 2011, p. 291) without committing to any structural entities that could realise these functional properties. However, according to Friston et al. (2012), "[...] dopaminergic gating may represent a Bayes-optimal encoding of precision that enhances the processing of particular sensory representations by selectively biasing bottom-up sensory information (prediction errors)" (p. 2). In turn, "dopaminergic gating" involves the neurotransmitter dopamine, a molecule that can be structurally described. Crucially, now that the functional concept of precision, derived from the free energy principle, has been linked with dopaminergic gating, one can make further inferences as to how this entity is situated in a multilevel mechanism. For example, the modulation of precision has been associated with attention (Feldman & Friston 2010; Hohwy 2012), and since precision is realised via dopamine mediation, one can investigate the effects of dopamine on attentional mechanisms.[7] On the other hand, if empirical evidence regarding precision or in particular predictions of precisions (hyperpriors) find "[...] that

[5] However, as a preliminary note, both functional and structural properties are needed for a full mechanistic explanation (cf. Piccinini & Craver 2011, p. 290).

[6] In this paper, the term "explanatory gap" is not used in the sense of "an *explanatory gap* [...] between the functions and experience" (Chalmers 1995, p. 205; see Levine 1983 for the classical reference), as we see in the philosophy of mind. Rather, it describes the lack of neurobiological details in functional concepts.

[7] Of course, to do so one would also have to know all the components involved in the mechanism responsible for attention.

descending signals do not mediate expected precisions, this would falsify the free energy principle" (p. 16). This further accentuates the need for mechanistic explanations.

As a more elaborate example, the phenomenon of biased competition will shortly be introduced. In biased competition, two stimuli are presented at a topographically identical location. However, only one of these stimuli is actually perceived. Thus the principal question: by which means does the brain "select" any given stimuli? In the free energy principle, the most obvious answer would be the stimulus that best minimises free energy or prediction error. However, in these cases, the stimuli are equally accurate, i.e., they both represent the causal structure of the world equally well. As a consequence, the stimuli will "[...] compete for the responses of cells in visual cortex" (Desimone 1998, p. 1245). Crucially, Desimone (1998) brings up a preliminary study by Reynolds et al. (1994) that states "[...] that attention serves to modulate the suppressive interaction between two or more stimuli within the receptive field [...]" (Desimone 1998, p. 1250). Thus, attention could be the determining factor as to which stimulus is perceived at a given moment. From the perspective of the free energy principle and in accordance with these findings, Feldman & Friston (2010) propose that "[...] attention is the process of optimizing synaptic gain to represent the precision of sensory information (prediction error) during hierarchical inference" (p. 2). These two views agree, since synaptic gain also entails a suppressive effect upon the other competing stimuli. Also, as just mentioned, Friston et al. (2012) identify precision weighting with dopaminergic gating, i.e., they argue that dopamine mediation realises the precision of incoming stimuli or prediction errors.

Now a fuller picture can be presented. This much more complete picture allows us to see how the free energy principle or prediction error minimization framework can prove to be beneficial with regard to mechanistic explanation. The phenomenon to be explained is biased competition. The mechanism that realises, or resolves, biased competition, i.e., the competition between two identically accurate and topographically identical stimuli, is precision weighting. This represents the etiological level of description since it describes how biased competition is resolved at a level of description that doesn't refer to lower-level processes nor to how they are embedded into a higher order mechanism. It remains at the same level in the hierarchy of mechanisms. At the constitutive level we have the fact presented by Friston et al. (2012), that precision weighting is neurophysiologically realised by dopaminergic gating. This *constitutes* precision weighting and is located at a lower level. Last, precision weighting is embedded into the higher-order mechanism of attention. Precision weighting contributes to this higher order mechanism, or, from the other perspective, attention is constituted by precision weighting. This represents the contextual description.

The upshot is that, just as "[e]volutionary thinking can be heuristically useful as a guide to creative thinking about what an organism or organ is doing [...]" (Craver 2013, p. 20), the free energy principle can be a useful guide in finding multilevel mechanistic explanations concerning how the mind works. Due to its unifying power, the free energy principle offers a grand framework that seeks to explain every aspect of human cognition. Thus, filling increasingly more mechanistic concepts into functional placeholders will enable an understanding of the mind in terms of how it does work instead of how it ought to work. The explanatory worth of the free energy principle would then be preserved, since "[i]f these heuristics contribute to revealing some relevant aspects of the mechanisms that produce phenomena of interest, then Bayesian unification has genuine explanatory traction" (Colombo & Hartmann 2014, p. 3).

However, this should not be seen as an attempt to eliminate functional concepts by reducing them to mechanistic ones. Instead, as mentioned above, the integrationist account emphasises that functional and mechanistic concepts are both necessary for mechanistic explanations, since "structural descriptions constrain the space of plausible functional descriptions, and functional descriptions are elliptical mechanistic descriptions" (Piccinini & Craver 2011,

p. 307). Furthermore, once every functional term has a mechanistic counterpart, the 3M requirement posed by mechanists can be fulfilled in the case of the free energy principle.

Last, as a general remark, searching for structural properties seems important if researchers want to ground the free energy principle in the human brain. Functional theories are subject to multiple realizability. This means that not only humans or mammals could be bound to the free energy principle, but also Martians or bacteria or anything that could possess the "hardware" to do so. Hohwy suggests that the free energy principle can be seen as a biofunctionalist theory (this collection p. 20). In principle this means that the free energy principle can be multiply realised as long as that creature acts in such a way as to maintain itself in a certain set of expected states. These expected states then determine the creature's phenotype. In seeking to explain human cognition, functional theories have to be enriched with mechanistic concepts relating to structural properties, since otherwise we could also be investigating Martians.

6 Conclusion

The negative thesis of this paper states that the free energy principle's explanatory power, unlike its unificatory power, can be regarded as weak, since it does not fulfil the 3M requirement posited by mechanists. This follows from the fact that the free energy principle is a functional theory, thus also employing functional concepts. Yet these do not explain how a given phenomenon in fact does work but only how it should work. However, Piccinini & Craver (2011) propose that functional analyses, ultimately, are nothing else but mechanism sketches, i.e., incomplete mechanistic explanations.

In this paper I have tried to make a positive contribution to the discussion by arguing for two claims: first, since the free energy principle incorporates empirical results from psychology and neuroscience it provides a stronger case of unification (SUT) than the unification provided by Bayesian theory alone. By not solely relying on its mathematical foundation, the free energy principle can try to ground its findings empirically in the brain. As a result, both the free energy principle and theories from psychology and neuroscience can constrain each other, thus being beneficiary to one another. Second, I argue that the free energy principle can act as a guide to finding multilevel mechanistic explanations. By linking mechanistic concepts with functional concepts from the free energy principle, the 3M requirement posited by mechanists can be fulfilled, consequently leading to actual explanations. This relates to the accused preposterousness of the free energy principle: with increasing explanatory power it becomes more and more difficult to deny that the free energy principle itself is, in fact, true.

References

Bechtel, W. (2009). Looking down, around, and up: Mechanistic explanation in psychology. *Philosophical Psychology*, *22* (5), 543-564. 10.1080/09515080903238948

Bickle, J. (2003). *Philosophy and neuroscience: A ruthlessly reductive account.* Dordrecht, NL: Kluwer Academic.

——— (2006). Reducing mind to molecular pathways: Explicating the reductionism implicit in current cellular and molecular neuroscience. *Synthese*, *151*, 411-434. 10.1007/s11229-006-9015-2

Chalmers, D. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, *2* (3), 200-219. 10.1093/acprof:oso/9780195311105.003.0001

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36* (3), 181-204. 10.1017/S0140525X12000477

Colombo, M. & Hartmann, S. (2014). Bayesian cognitive science, unification, and explanation. [Pre-Print]. (Unpublished)

Craver, C. F. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience.* New York, NY: Oxford University Press.

——— (2013). Functions and mechanisms: A perspectivalist view. *Synthese Library*, *363*, 133-158. 10.1007/978-94-007-5304-4_8

Craver, C. F. & Kaplan, D. M. (2011). Towards a mechanistic philosophy of neuroscience. In S. French & J. Saatsi (Eds.) *Continuum companion to the philosophy of science* (pp. 268-290). London, UK: Continuum Press.

Desimone, R. (1998). Visual attention mediated by biased competition in extrastriate visual cortex. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *353* (1373), 1245-1255. 10.1098/rstb.1998.0280

Feldman, H. & Friston, K. J. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience*, *4* (215), 1-23. 10.3389/fnhum.2010.00215

Friston, K. J. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Science*, *360* (1456), 815-836. 10.1098/rstb.2005.1622

——— (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*, 127-138. 10.1038/nrn2787

Friston, K. J., Daunizeau, J. & Kiebel, S. J. (2009). Active inference or reinforcement learnin? *PLoS ONE*, *4* (7), e6421. 10.1371/journal.pone.0006421

Friston, K. J., Mattout, J. & Kilner, J. (2011). Action understanding and active inference. *Biological Cybernetics*, *104* (1-2), 137-160. 10.1007/s00422-011-0424-z

Friston, K. J., Shiner, T., FitzGerald, T., Galea, J. M., Adams, R., Brown, H., Bestmann, S., Dolan, R. J., Moran, R. & Stephan, K. E. (2012). Dopamine, Affordance and Active Inference. *PLoS Computational Biology*, *8* (1), e1002327. 10.1371/journal.pcbi.1002327

Hesselmann, G., Sadaghiani, S., Friston, K. J. & Kleinschmidt, A. (2012). Predictive coding or evidence accumulation? False inference and neuronal fluctuations. *PLoS ONE*, *5* (3), e9926. 10.1371/journal.pone.0009926

Hobson, J. A. & Friston, K. J. (2012). Waking and dreaming consciousness: Neurobiological and functional considerations. *Progress in Neurobiology*, *98* (1), 82-98. 10.1016/j.pneurobio.2012.05.003

Hohwy, J. (2012). Attention and conscious perception in the hypothesis testing brain. *Frontiers in Psychology*, *3* (96), 1-14. 10.3389/fpsyg.2012.00096

——— (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.

Jones, M. & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of bayesian models of cognition. *Behavioral and Brain Sciences*, *34* (4), 168-188. 10.1017/S0140525X10003134

Kaplan, D. M. (2011). Explanation and description in computational neuroscience. *Synthese*, *183* (3), 339-373. 10.1007/s11229-011-9970-0

Kaplan, D. M. & Craver, C. F. (2011). The explanatory force of dynamical and mathematical models in neuroscience: A mechanistic perspective. *Philosophy of Science*, *78* (4), 601-627. 10.1086/661755

Levine, J. (1983). Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly*, *64*, 354-361.

Mumford, D. (1992). On the computational architecture of the neocortex. II. The role of cortico-cortical loops. *Biological Cybernetics*, *66* (3), 241-251. 10.1007/BF00202389

Piccinini, G. & Craver, C. F. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese*, *183* (3), 283-311. 10.1007/s11229-011-9898-4

Rao, R. P. N. & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2* (1), 79-87. 10.1038/4580

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.