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 ex-

planatory *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 pre-

dictions 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