The Problems with Prediction

The Dark Room Problem and the Scope Dispute

Andrew Sims

There is a disagreement over the scope of explanation for predictive processing. While some proponents think that it is best motivated from—and indeed comprises an explanation of—biological self-organization, others maintain that it should only be a theory of neurocognitive function, or even just of some limited domain of neurocognitive function. Something that these theorists share is an interest in addressing the dark-room problem: at its most naïve, if action is driven by the minimization of surprise then why don't cognitive creatures act to minimize stimuli in general? The dark-room problem is in fact best conceived as a cluster of related concerns, rather than as a single argument against action-oriented predictive processing. These have to do with: i) whether PP (predictive processing) has any substantive empirical content when it is pitched in very general domains; ii) whether a specification can be given of the environmental niche that action moves the organism to occupy, and which is not the dark room; and iii) whether an adequate account can be given within this specification of exploratory and exploitative behaviours. There are interesting conceptual relations between the dark-room problem and the scope dispute. As the putative scope of predictive processing gets wider (culminating in the free energy principle), the resources that are available for answering the concerns about niche-specification become very rich. But increasingly puzzling problems arise as to the implementation of surprise-minimisation within non-paradigmatically cognitive biological systems. On the other hand, under more restrictive construals of the scope of predictive processing, there are new difficulties standing in the way of niche-specification, and new questions about the interface between surprise-minimisation and model-free cognition undermine the promise of predictive processing as a unifier of theories of neurocognitive function in subordinate domains. In this paper I make explicit the dialectic between proponents and critics in order to show how the two problems are related.

Keywords

Active inference | Aneural information processing | Dark-room problem | Explanatory scope | Free energy principle | Minimal cognition

1 Predictive Processing and the Scope Dispute

Predictive processing (PP) is the name for a class of theories in cognitive neuroscience which present the prospect of unifying various accounts of perception, action, and very many other ordinary and pathological cognitive phenomena. Roughly, the central idea is that both perception and action can be explained in terms of a mechanism whose sole function is the minimisation of surprise. There are an astounding range of ordinary and pathological cognitive phenomena that have possible explanations expressed in the theoretical vocabulary of PP, and the number of these continues to grow.

Surprise—often also called surprisal or self-information to distinguish it from phenomenological surprise (Clark 2013a)—is a quantity from information theory that describes how likely some event or set of events is, given some model that assigns probabilities over events. To make the distinction between information-theoretic and phenomenological surprise concrete, Clark gives the example of an elephant being smuggled onstage during a magic show. Given the context, this is not very surprising in an information-theoretic sense. Such things may be expected to happen in the context of a magic show. However, we may expect the audience to experience phenomenological surprise. In its broadest and least controversial formulation, PP states that some neurocognitive functions can be explained

Sims, A. (2017). The Problems with Prediction - The Dark Room Problem and the Scope Dispute.
In T. Metzinger & W. Wiese (Eds.). *Philosophy and Predictive Processing:* 23. Frankfurt am Main: MIND Group. doi: 10.15502/9783958573246

in terms of the minimisation of information-theoretic surprise. In stronger formulations of PP the explanatory scope of surprise minimisation becomes wider, from the extension of the mechanism to action (Friston 2009) all the way up to its application in understanding adaptive behaviour in general (Friston 2013). The details of PP will have already been discussed at length by the other contributors to this volume, and so I will not belabour the details except to remind the reader of its key features.

Consider then the problem of under-determination for visual perception. One pattern of retinal stimulation (a set of sensory states) is compatible with very many interpretations of its causes due to ambiguities that arise from distance, occlusion, and noise; the task in visual perception is to construct the best possible interpretation on the basis of that pattern. This means producing a visual scene on the basis of the retinal stimulation. One way to eliminate ambiguity is to interpret the sensory states on the basis of a model of the causes of those states. That resolves ambiguities by discounting interpretations of the states that are less likely, given that model. Thus, to give an informal example, we are able to perceive a particular pattern of sensory states as an old or young woman rather than as a truly ambiguous figure (Figure 1).

Figure 1: The "young woman/old woman" ambiguous figure (Boring 1930).

PP posits a computational architecture that is capable of successfully implementing a species of this basic top-down strategy (Rao and Ballard 1999). Imagine a model of the causes of sensory states that is *hierarchical*, and for which each level in the hierarchy is predicting the states of the level below it (it is *generative*, since it generates predictions of the sensory states it should encounter if the model is true). So at each level there is a comparison between the actual states which are propagated upwards from lower levels and laterally, and predicted states that are propagated downwards from higher levels and laterally. These can either match or not. If they do not match, then the actual state is propagated upwards as a *prediction error* signal, and the parameters at the higher level are updated on the basis of that prediction error in line with the norms of Bayesian belief-updating. So the higher-up parameters that generate predictions are treated as prior beliefs (in the form of probability distributions) about the way that states at the lower level will behave, and they are updated where they fail to predict those states accurately.[1] This means that the mechanism as a whole can slowly approach being an accurate model of the causes of sensory states, and that it does so by the minimisation of prediction error, which is an upper bound on surprise. PP states that neurocognitive systems implement a hierarchical generative model of this kind, with sensory states at the lowest level in the hierarchy, and that the model of the causes of those states is the content of perception—in this context, the visual scene.

1 "Prior belief" is a term of art here, however. It refers to whatever plays the functional role of top-down prediction in the hierarchical generative model. It's not to be taken as indicating the existence of a propositional attitude.

More expansive versions of this broad kind of proposal make room for multiple perceptual modalities in the hierarchy and include suggestions that it extends upwards into more abstract and amodal predictions. On this hypothesis we posit a more abstract hierarchical model about the causes of sensory states that helps inform perception in multiple modalities. For example, my interpretation of some set of sensory states in audition may be disambiguated by more general beliefs about the context of the states—I may interpret those states as being caused by the clanging of pots and pans on the basis of a prior belief that the origin of those states is from behind a door which I can see is clearly marked "KITCHEN," and on the basis of prior beliefs to do with kitchens more generally. Although it is not uncontroversial as to what each of the levels represents (Vance 2015), one way to make sense of the notion is that each level in the ascending hierarchy represents the world at an ever increasing level of spatiotemporal generality (e.g., Hohwy 2013, pp. 28-30).[2]

For a mechanism like this to work effectively, it needs to be able to distinguish between signal and noise. What that means is that it needs to evaluate how likely some datum is to be genuinely indicative of a causal regularity in the world. A hypothetical example of a situation in which a datum is *not* genuinely informative in the right kind of way can be given in the context of the measurement of population growth. If I am trying to measure long-term growth in tourism in Rio de Janeiro during the Olympics, for example, I will get a reading that is not genuinely informative about tourism growth in that city over a longer period of time. It is a noisy datum. Updating a Bayesian model on the basis of noisy data leads to overfitting; an overfitted model is greatly reduced in its predictive power because it contains redundant parameters and does not easily generalise to new data. So an effective Bayesian model needs to modulate its updating on the basis of the reliability of error, that is, how likely an error is to be genuinely informative about a causal regularity. In PP, this function is performed by the precision weighting of prediction error. A precise prediction error—a prediction error which has been directly assigned a high precision value—is considered very reliable, and so the model is updated on the basis of that error without much attenuation. But imprecise prediction error is treated as noisy and unreliable, and is therefore more likely to be attenuated if any updating occurs at all. The model therefore also needs to maintain a higher-order model of precision, in order that imprecise error can be treated as such; predictions are treated as more reliable than prediction error under conditions of sensory uncertainty.

It's possible to explain action according to this same basic computational architecture. Whereas perception is thought to function through an alteration of prior beliefs on the basis of prediction error, action is construed as the minimisation of prediction error through an alteration of the way that sensory states are sampled. That is to say that the world is sampled such that some set of predictions come out true (but this is not merely a corroborative but also a disambiguating process, as in the case of the visual saccade, for example, Friston et al. 2012a). In the simplest case the way this might occur is that the model predicts some set of proprioceptive states corresponding to the movement of the body, and then realises those proprioceptive states by the propagation of prediction through classical reflex arcs. Where prediction error is encountered (that is, at the stage when the predicted outcome of the movement does not yet obtain), this error is assigned a very low precision, and so resampled in such a way that the predictions come out true. But the same basic principle can be posited for more complex actions and higher levels of planning, so that the predictive-processing theorist can explain my walk to the cafeteria in terms of the visual, auditory, tactile, and other sensory states that I predict to encounter during the course of that action. It is the minimisation of prediction error that drives the action, but

2 What I mean by representation here should be spelled out. The extent to which PP requires representations, and of what kind, is controversial (see Clark 2016, §6.6, and Hohwy In Press, for diverging views). So I shall here assume a weak and inclusive notion: for these purposes the term "representation" describes an isomorphism between one or more physical particulars and a cause or structured set of causes within the world, so that the particular(s) carry information about their causes in a specific manner (Dretske 1981). That is a very weak notion, but it can be supplemented with further conditions, such as a history of playing a particular role for a system that consumes the representation (e.g., Millikan 1984). Similarly, given the possibly wide scope of PP (section 3.3), it seems prudent to assume a pluralism about representation, on which the content of representations may be fixed by different kinds of facts in different kinds of representational system (Shea 2013).

it is minimised by resampling, rather than updating. This way of distinguishing between perception and action within PP mirrors a distinction in the philosophy of science between theory-revision and experiment (see Hohwy 2013, p. 43, for a nice illustration of this).

Everything that I have explained so far is relatively uncontroversial, except for the way that some of the finer details should be elucidated, for example, whether the hierarchy should be conceived in terms of spatiotemporal or computational depth. But there is a more significant disagreement over the scope of the explanation (Sims 2016). That is to say that not everybody who endorses the general mode of explanation can agree on how much it is supposed to be an explanation for. On this matter, four very general positions can be distinguished (see Table 1). I should note at this point that I do not mean to imply that these positions are consistently held by any particular researchers, except perhaps implicitly. This taxonomy is supposed to be of heuristic value, in mapping out the conceptual possibilities. On the first position (minimal predictive processing), PP is only to be construed as a theory of any number of *perceptual* processes—perhaps visual perception, for example. This is a rare position in the philosophical literature, however, since one thing that seems to be broadly agreed on is that one of the reasons that PP is so interesting is that it can give a unified explanation of perception and action in terms of the same mechanism. This position is also not vulnerable to the dark room problem, and for these reasons it is not relevant for the purposes of my exposition. On the second position (mixed predictive processing), PP is to be interpreted as a theory of just some neurocognitive processes of both perceptual and motor kinds. That means that the explanatory burden for neurocognitive function is shared amongst both PP and other kinds of models in a "mixed" theory. For example, one may suggest that there are mechanisms involved in action which are "model-free," and do not require a representation of the causes of sensory states. One may insist, for example, that neurocognitive function includes "complex admixtures of strategies including the canny use of bodily form and various 'representation-lean' ploys." (Clark 2013b, p. 8) On the third position (maximal predictive processing), PP is to be construed as a complete theory of all neurocognitive function. That means that all neurocognitive function can be explained in terms of the minimisation of prediction error. A proponent of this position will insist on the "preposterous" nature of the hypothesis: "it leaves no other job for the brain to do than minimise free energy[3]—so that everything mental must come down to this principle." (Hohwy 2015, pp. 8-9)[4] On the fourth, and boldest, position, the mechanism described in PP is not only to be understood as a complete theory of neurocognitive function but of adaptive behaviour in general; all adaptive behaviour is understood in terms of the minimisation of surprise. It seems clear that Karl J. Friston intends the "free energy principle" to be taken in this way:

> Most treatments of self-organization in theoretical biology have addressed the peculiar resistance of biological systems to the dispersive effects of fluctuations in their environment by appealing to statistical thermodynamics and information theory. Recent formulations try to explain adaptive behavior in terms of minimizing an upper (free energy) bound on the surprise (negative log-likelihood) of sensory samples. This minimization usefully connects the imperative for biological systems to maintain their sensory states within physiological bounds, with an intuitive understanding of adaptive behavior in terms of active inference about the causes of those states. (Friston 2013, p. 1)

For all of these positions (except the minimal position) there is a conceptual difficulty with the explanation of action that has been called "the dark room problem." (Friston et al. 2012b). That problem is an apparent consequence of the explanation of action—and therefore motivation as well—on the sole basis of the minimisation of surprise.

3 Free energy is the long term average of prediction-error, and therefore an upper bound on surprise over the long term.

4 It's probable that Hohwy holds a stronger position than this—for example, it seems that he may be willing to countenance an interpretation of natural selection in terms of surprise minimisation (see in particular Hohwy 2015, p. 10)—but what he says here exemplifies well the commitments of the maximal predictive-processing theorist in general.

Table 1: Kinds of positions with respect to the scope of PP.

Position	Scope
Minimal predictive processing	Some perceptual processes
Mixed predictive processing	Some perceptual and motor processes
Maximal predictive processing	All neurocognitive processes
Free energy principle	All biological processes, on multiple timescales

2 Three Aspects of the Dark-room Problem

The way that the dark room problem puts pressure on PP and its extension to action can be captured in the following line of reasoning. First, the critic claims that the explanation of action on the basis of prediction-error minimisation entails that the agent is always acting to minimise surprise. Second, it is inferred that this basic principle means that the ideal surprise-minimising agent should be expected to seek out a place where it can be free of surprising and unanticipated stimuli. And this would be an environment that is free of any stimuli whatsoever; this is the "dark room" that gives the problem its name. But then, the critic goes on, this is an absurd consequence; this kind of behaviour would spell extinction for any agent which carried it out. That is because an environment without any stimuli is also an environment without any nourishment or opportunity to reproduce. A creature which behaved in this way would not survive, and it would not pass on its genes to offspring. It follows by *modus tollens*, then, that action must be driven by processes which are not surprise-minimizing. This is how the problem is spelled out in its most basic terms. But in fact as we shall see it is better conceived as a way to articulate a cluster of related concerns, rather than a unitary argument with a single conclusion. I will be arguing that there are three distinct difficulties that are raised here. The first, the negative problem, concerns the apparent result that surprise-minimisation entails that the agent seeks to rid itself of stimuli altogether; the second, the positive problem, concerns the presumed poverty of behaviour that could be produced by mere surprise-minimisation (the charge is that it fails to account for rich repertoires of exploratory, exploitative, and playful behaviour); the third, the problem of triviality, is a sceptical concern about whether the extension of scope means sapping PP of any empirical content, rendering it trivial. What unifies these problems under the "dark room" rubric is that they originate in the concerns about how to model motivation within the framework; this is an issue which is revealed starkly in the dark room scenario. The reason that the problem of empirical content is also related to the dark room scenario is that it is attempts to address the positive problem in terms of evolutionarily selected "deep" priors which gives rise to the charge of triviality; this will be made clear in section 3.3.

2.1 The Negative Problem

The negative problem is the aspect that is the most initially intuitive. It is the identification of surprise-minimisation with stimuli-minimisation in general. There is a sense in which all stimuli are minimally surprising, if one takes the baseline for stimulation to be the state of the organism prior to the stimuli. That is to say that any stimuli will be a change in what the organism's current state is, and be surprising in virtue of this difference, however minimal. On top of this, it appears to be that a consistently applied PP framework will place surprise-minimisation as the sole principle that drives action. This need not be the case—a contrary example would be the mixed models of section 3.1—but this is certainly the case for the more ambitious readings of PP on which it is supposed to suffice as a unified theory of all the processes underlying perception and action.

So with these two pieces in place, the negative problem states that insofar as all stimuli are minimally surprising, and insofar as a consistently applied PP must place surprise-minimisation as the sole principle driving action, then it seems that the surprise-minimising agent ought to minimise stimuli in general. But that seems wrong, or at least at odds with what we know about living things: they don't

seek to minimise stimuli in general. So something has gone awry. Either the analysis of all stimuli as minimally surprising is incorrect, or it's wrong to say that all action is the minimisation of surprise.[5] Here are two examples of the dark room problem thus characterised in the literature. The first is in a paper by Schwartenbeck and collaborators, who aim to give a formal treatment of the issue, and who are one of the first to explicitly distinguish the negative and positive problems: "Should we not, in accordance with the principle, prefer living in a highly predictable and un-stimulating environment where we could minimize our long-term surprise?" (Schwartenbeck et al. 2013) And another more recent example, from Clark's recent and comprehensive book-length treatment of the philosophical issues associated with PP:

> The hapless prediction-driven organism, the worry goes, should simply seek out states that are easily predicted, such as an empty darkened room in which to spend the remainder of its increasingly hungry, thirsty, and depressing days. This is the so-called 'Darkened Room Puzzle'. (Clark 2016, p. 262)

That is the baseline concern that one can take away from the dark room problem: that minimising surprise entails minimising stimuli. One may deal with this problem by rejecting either of the two premises that lead to that result: either that all stimuli are surprising or that all action is surprise minimisation. On the first strategy, we are owed an explanation of why some stimuli are surprising and why others are not; on the second, we are owed an account of the other mechanisms that drive action, and how they are related to PP.

2.2 The Positive Problem

Typically, then, the PP theorist rejects the premise that all stimuli are minimally surprising. She rejects this on the basis that surprise minimisation always occurs relative to a model which assigns probabilities over possible sensory states, based on a set of (Bayesian) prior beliefs about the structured causes which produce those sensory states. This hierarchical generative model predicts the states that the agent will encounter, and it's on this basis that the least surprising state is in fact not a 'blank slate', as the negative problem assumes, but rather the kind of environment that the agent already expects to encounter. For prediction-error minimising agents, action is driven by prediction error relative to a set of predictions about the optimum states for the agent to be in. And this, the PP theorist will say, is by no means the darkened room of the negative problem.

Now, the positive problem has to do with the role that is left for uniquely motivational states to play in PP, once their traditional role has been usurped by "prior beliefs" or "predictions" about the agent's sensory states, and which drive action on the PP framework. It may be that, by positing the influence of a model of causes of sensory states, the PP theorist can show how the negative problem is mistaken. She can do so by pointing out that the sensory states that the agent expects to be in are species specific and are specified on the basis of an agent's prior adaptation to a particular ecological niche. But this seems to imply that we are left without the classical distinction between representations with different directions of fit, because the way that action works within PP is by minimising surprise with respect to prior beliefs rather than maximising utility with respect to the agent's desires. Thus we see Pezzulo and colleagues claim that:

> […] our scheme for behavioural control is based on Bayesian inference and does not call on reward prediction errors for learning or inference. One advantage of this is that the concept of rewards is replaced by the realization of prior preferences. This means that epistemic value and pragmatic

5 Klein (in press) notes that this problem is recognised early on by Mumford (Mumford 1992). But in fact the problem is extant even *earlier* on; even Freud (Freud 1950 [1895]), who inherits the Fechnerian (Fecher 1873) conception of cognition as the minimisation of quantity, is compelled to posit the "reality principle" in order to address very similar issues.

> value (e.g., utility or reward functions) have the same currency and can be accommodated within the same (information hungry) Bayesian scheme […] (Pezzulo et al. 2015, p. 27)

However, this may render a large number of observable behaviours inexplicable – those behaviours that are playful, exploratory, and exploitative. In other words, reducing the states driving action to states that predict future states of affairs means that we can no longer make a cogent distinction between what is probable for an agent and what is of value for it, and we need the concept of value to explain why there are unlikely or uncertain states that are nonetheless valuable or desirable on the part of the agent. It's fairly intuitive to judge that states with high utility for an agent are not always those with high prior probability, and putative examples of this dissociation are becoming ever more widespread in the critical literature:

> Should the first amphibian out of water dive back in? If a wolf eats deer not because he is hungry, but because he is attracted to the equilibrium state of his ancestors, would a sudden bonanza of deer inspire him to eat only the amount to which he is accustomed? Should a person immersed in the "statistical bath" of poverty her entire life refuse a winning lottery ticket, since this would necessitate transitioning from a state of high equilibrium to a rare one? (Gershman and Daw 2012, p. 306)

In other words, even if the dark room scenario does not obtain (agents don't aim to minimise all stimuli) we should still expect an agent who only minimises surprise with respect to a model of the causes of sensory states to lack many of the playful, exploratory, and exploitative behaviours that we observe in the natural world—and perhaps more pertinently, in ourselves.[6] For example, it is not necessarily adaptive for an agent to consume resources in amounts predicted by past consumption—especially if past consumption was in meagre amounts. We would expect that agent, if adaptive, to *exploit* any sudden availability of resources. The criticism that surprise-minimisation does not predict such behavior remains live even after the negative problem is resolved by appeal to a model which relativises surprise minimisation to a particular set of prior beliefs.

2.3 The Triviality Problem

The triviality problem has to do with the empirical content of PP. Stated baldly, it is a concern that PP, when pitched at a sufficiently wide scope, may turn out to lack empirical content. This is a problem that is most often associated with the widest-scoped version of PP, the free energy principle. The free energy principle becomes relevant at this point because one way in which the positive problem may be overcome is to expand the scope of the model in order to include evolutionary influences. That would offer a possible way of addressing the issue because it can explain the existence of the problematic behaviours in terms of priors that are evolutionarily specified. It does not require that these priors be learned from the environment. It can do so because evolution is itself construed as the minimisation of surprise at phylogenetic timescales.

Although the triviality issue is very often raised informally at conferences, it is less widespread in print.[7] But I can provide two loci at which the issue is raised explicitly by critics. The first is in the original discussion of the dark room problem itself, in the dialogue between Friston, Thornton, and Clark. In responding to the notion that the negative problem is dissolved by appeal to the model of the causes of sensory states, Thornton makes the charge that: "If we allow unlimited rein over the interpretations [i.e., models] agents are assumed to apply, the dark room problem can be eliminated. But the hypothesis then seems to be stating something that is true by definition." (Friston et al. 2012b, p. 1)

6 Of course, there are various ways in the literature to deal with this concern, see especially section 3.2 below.

7 Though more so with respect to Bayesian models of cognition in general, cf., Bowers and Davis 2012.

Indeed, this sometimes seems to be the case, and a tautological reading of the free energy principle is even sometimes endorsed quite explicitly. Look at what Friston himself says in the same article: "The tautology here is deliberate […] Like adaptive fitness, the free-energy formulation is not a mechanism or magic recipe for life; it is just a characterization of biological systems that exist." (Friston et al. 2012b, p. 2) The concern here is therefore that the characterisation of all action as surprise-minimisation reduces to a definition or tautology, and does not constitute an empirical hypothesis that generates substantive predictions or explains causes in the world.[8]

One way to interpret this is a concern about falsifiability. That would be to say that the free energy principle is not falsifiable, because it is compatible with every state of affairs. It therefore lacks empirical content. Although the falsifiability interpretation of the triviality problem is a common critical refrain, is has been noted that the objection lacks some force (Hohwy 2015, p. 14-15). Firstly, it is a widespread view that falsifiability is not sufficient nor even necessary for something to be a genuine scientific explanation. Secondly, the relationship between biological function and natural selection seems similarly definitional or conceptual in this way, but almost nobody denies that the theory of natural selection is an empirical theory that describes a causal process (Ruse 2008, pp. 44-45).

So with the natural selection analogy in mind, one way to deal with the triviality problem has been to argue that the value of PP will come out of its pragmatic value in constraining more local empirical hypotheses that describe mechanisms more limited in scope. And it may not only act as a general constraint (perhaps in the same way that the laws of physics (controversially) constrain explanations of particular systems), but also perhaps be suggestive of explanations in subordinate domains like systems biology and abnormal psychology.

That then leads us to the second place where this concern is raised at length in the literature, and to the more subtle version of the triviality problem that it constitutes, in a soon-to-be-published treatment by Klein (Klein in press). Klein notes that even if it is the case that PP and the attendant free energy principle admits of pragmatic value in producing hypotheses, the fact that its proponents quite openly admit the tautological nature of the wider scoped hypotheses is problematic:

> Appeal to apparent tautologies should trouble you. For whatever tautologies do, they don't explain why things happen. At best, they give us reason to believe that something is the case. But philosophy of science has moved away from epistemic conceptions of explanation and towards ontic ones […] Good explanations detail a causal story, and it is not obvious that [the free energy principle] does so. (Klein in press)

That leaves no empirical content for the theory itself – it is rather what Klein, after McMullin (McMullin 1985), calls a "Galilean idealisation." Like the frictionless plane, the content of the free energy principle would on this view be literally false, though perhaps useful in formulating empirical hypotheses if taken with a grain of salt. It remains to be seen what consequences this has for the way we should think about the theory, both in an epistemic (does it make endorsement of it less justified?) and ontological (what entities and processes does it imply?) sense.

Such is the dialectic that leads us through the three aspects of the dark room problem. The initial intuition is that replacing conventional motivational imperatives with the imperative to minimise surprise entails absurd consequences—that the best strategy for surprise minimisation would be the minimisation of sensation in general. That produces the appearance of a dilemma, with two corresponding ways to reply: either sensation is not minimally surprising or not all motivation is surprise-minimisation. The latter is relatively undesirable, since it undermines the unificatory appeal of PP. But the former is easily followed, given that PP includes the notion of a model of causes of sensory

8 An anonymous reviewer suggests that this is a straw man. But it is a view clearly held by critics: "[Thornton:] we can certainly view the process by which agents adapt to their environments as a process by which they reduce their surprise. The problem is we can also view it the other way around, seeing the situation in terms of agents reducing their surprise by adapting to the environment." (Friston et al. 2012b, p. 1)

states within which there are states that are less surprising than an absence of states altogether. But even with the model taken into account, one may doubt whether the imperative of surprise-minimisation can produce complex behaviours like play and exploration. Last, there also seems to be a problem of triviality for wider-scoped free energy principle that is also related to the dark room problem. That emerges out of the appearance of tautology in the claim that all adaptive behaviour is free energy minimisation, and challenges the ability of free energy theorists to provide substantive empirical hypotheses regarding specific mechanisms.

3 Three Replies to the Dark-Room Problem

3.1 Mixed Predictive Processing

On first gloss, it seems that the issues associated with the dark room problem pushes us towards a view that is mixed. A view that is mixed is a view that makes room for cognitive processes that are not surprise-minimisation. (Clark 2016) is the most thorough working out of a view like this that exists in the philosophical literature, though it is (Pezzulo et al. 2015) that have given a more thorough mechanistic account; I will base my discussion around both of these. However, I should begin my explanation of these mixed predictive processing theories by saying something about the distinction between model-free and model-based processes, since one way to develop a mixed view of predictive processing is to have model-free processes play the role of motivating action, and thereby giving an answer to the (negative) dark room problem that takes the horn of the dilemma on which action is driven by processes *other* than surprise-minimisation of the kind posited in PP.[9]

The distinction between model-based and model-free processes originates in the study of reinforcement learning, where it is used to distinguish between a process that learns the value of available options by trial and error, and without a model of the causal structure of the environment—this is model-free—and learning that assigns value on the basis of a model of how rewarding events in the world are statistically related to other events—this is model-based (Gläscher et al. 2010). PP would fall unambiguously under the "model-based" category of learning processes. Model-free processes are attractive in the context of the dark room problem because they may be construed as offering a set of imperatives to action which could be mixed with the standard PP mechanism of surprise minimisation in order to yield imperatives to action that look genuinely "motivational," and which therefore entail action that defeats the dark room problem. For example, it may be that there is a mechanism underlying action that causes an agent to indiscriminately seek out and consume sources as reward on the basis of availability, and drives action in this way. Ainslie's (Ainslie 2001) behavioural findings of hyperbolic discounting could be indicative of such a mechanism. He has found in studies both in animal models and human participants that the way rewards are valued increases steeply as they approach in time, and that smaller but more immediate rewards tend to be consumed in preference to temporally distant but larger rewards in decision-making tasks. It may be that whatever mechanism produces this effect works in independence from any kind of model of the causes of reward, that is, it is model-free (cf., Clark 2016, pp. 252-256).

If this is the case, then a mixed theorist can give an answer to the dark problem which deals with all three aspects at a single stroke. The negative problem is clearly no issue, since the model-free mechanisms that drive action are not minimising surprise, they are reward-seeking and therefore attempt to bring the agent into contact with the appropriate stimuli. And cutting off the dark room problem this early in the dialectic also means that the other two issues (the positive problem and the triviality problem) do not come up, since those problems result from the sole appeal to a model in order to deal

9 An anonymous referee has advised me that model-free processes can be understood within the context of predictive processing (e.g., Pezzulo et al. 2015; also Clark 2016, § 8.6), and thereby constitute a complement to predictive processing rather than a competitor. That's true. But it's not necessary to do so, and to construe model-free processes in this way just means that the model is not genuinely mixed; it collapses into the maximal view.

with the negative problem; this answer takes the other horn of the dilemma, which doesn't lead to those two issues.

However, it is unlikely that many PP theorists are going to want to take this path. That is because it undermines one of the most attractive features of PP: the way in which it serves to unify the mechanisms underlying perception and action within a single theory. Indeed, one might observe that it solves the problems associated with PP by ceasing to be a PP theory; it fails to unify perception and action in the way that is extolled in the philosophical literature:

> [PP] is a proposal that has already been applied to a large—and ever-increasing—variety of phenomena. It thus serves as a powerful illustration of the potential of some such story to tackle a wide range of issues, illuminating perception, action, reason, emotion, experience, understanding other agents, and the nature and origins of various pathologies and breakdowns. (Clark 2016, p. 10)

This may be one reason why authors have formulated mixed views on which the model-based processes *themselves* play an arbitrating role. That is to say that it is the models themselves which determine when model-free processes drive action, and when learning is instead contextualised within a model of the causes of sensory states. Again, Andy Clark holds a view like this: he thinks that "[…] a kind of meta-model […] would be used to determine and deploy whatever [model-based or model-free] resource is best in the current situation, toggling between them when the need arises." (Clark 2016, p. 253) This is an attractive view for other reasons, as well. There is evidence to suggest that model-free learning processes do not exist in isolation from those that are model-based, but rather that they are highly integrated (Daw et al. 2011).

Clark's proposal, more specifically, is that model-free and model-based processes need to be understood as situated along a scale where the latter kind of learning is dominated by top-down influence within the generative hierarchical model and the former is dominated by bottom-up sensory influences. A mechanism of this sort is outlined more formally in Pezzulo et al. (Pezzulo et al. 2015). They envision the relationship between model-based and model-free in terms of a hierarchy where higher-levels within the model contextualise lower levels, and that learning is to be considered "model-free" when the higher levels fail to contextualise those lower. What determines whether contextualisation occurs or not are assignments of precision within the model; when prediction errors are assigned higher precision values then they drive action in ways that are less contextualised, because the higher levels of the model exert less of an influence.

Notice, however, that it seems to be that that *this* kind of mixed PP view entails that all learning is minimally inferential and model-based, because there is no learning that is entirely independent of the meta-model. Certainly, Pezzulo et al. (Pezzulo et al. 2015, p. 32) seem to recognise this: "Strictly speaking […] habitual behaviour is not completely model free in that it continues to depend on the (simplest) type of predictive model, of the kind 'because there is a stimulus, I expect a response'." So a mixed-view, when it is properly elaborated, appears to in fact be a species of "maximal predictive processing" theory. That is because after all is said and done, surprise minimisation nonetheless remains the sole imperative driving action, even in putatively "model-free" modes of learning. Therefore, whether or not this view is successful in addressing the dark room problem depends on whether these maximal views are so successful. Let's consider that now.

3.2 Maximal Predictive Processing

Maximal predictive processing is the view that prediction-error minimisation is *all* that the brain ever does; all neurocognitive function can be explained in terms of the minimisation of surprise. So we aim to explain all cognition, and thereby all mental phenomena, in terms of prediction-error minimisation. The maximal theorist claims that the mammalian brain works (and only works) by minimising

prediction error, but is agnostic on the question of whether or not other biological entities or systems function in this way.

In facing up to the negative aspect of the dark room problem, the maximal PP-theorist chooses to take the horn of the dilemma on which not all stimuli are minimally surprising. The way this works is to demonstrate that the minimisation of surprise is carried out with respect to a hierarchical generative model of the causal structure of the world. This model assigns probabilities to sensory states—that is, it generates predictions—such that the agent anticipates that it will be in states that reflect the ecological niche to which it is adapted, which is that within which it can harvest reward and pass on its genes. Given that ecological niche is species-specific, it appears that there needs to be some kind of story given here about the origins of the priors which specify that niche. In other words, there must be a relation between the neural and surprise-minimising morphology of the agent and the non-neural but niche-specifying morphology of the agent, such that the non-neural morphology can play an appropriate role within the model without itself being surprise-minimising. If non-neural morphology is in fact directly (and not vicariously) surprise-minimising, then this view collapses into the much stronger free energy principle which I discuss in section 3.3.

Now, there are a number of ways in which this general strategy may be pursued. One is to say that the relevant morphological traits are themselves represented within the model, and that this allows them to play a role in prediction-error minimisation without themselves minimising prediction error. With this in mind, a first attempt at such an account of morphological representation within the surprise-minimising brain might focus upon the interoceptive prediction of the internal milieu (Craig 2003). Interoceptive systems monitor the physiological states of the body such as "[…] those relating to heart rate, glucose levels, build-up of carbon dioxide in the bloodstream, temperature, inflammation, and so on." (Barrett and Simmons 2015, p. 419) The prediction of these sensory states leads us to perceive them as feelings about those states. So, for example, we might interoceptively perceive dehydration as thirst. There are influential attempts in the literature to account for interoception within the scope of PP (Seth 2013; Barrett and Simmons 2015). Within the PP framework the homeostatic states are those that are predicted, and deviation from those predicted states (deviation from interoceptive states associated with satiation, for instance) will lead the creature to take action in order to bring itself back into line with those states.

That seems a satisfactory first pass at how the basic PP story may be extended to take the morphology of the agent into account. When it comes to the negative aspect of the dark room problem, this affords the following answer. The agent does not stay in the dark room because doing this would lead it to occupy sensory states that are surprising, states that are interoceptively perceived as hunger and thirst. Therefore, staying in this impoverished environment does not in fact minimise surprise, but rather elicits it. In other words, staying in that environment is a very poor strategy for the minimisation of surprise. A much better strategy is to actually get out and exploit richer environments for nourishment so that the interoceptive states can be brought back into line with prior expectations. The bottom line is that on this view we should not expect the dark room scenario to obtain. That is because the dark room problem does not take into account interoceptive sources of surprise, but only exteroceptive sources like vision and hearing. Furthermore, we could conceivably extend this undeveloped account to encompass those exteroceptive senses. For instance, the mammalian eye is structured in such a way that it is receptive to changing patterns of light. It is not unconceivable that part of the model reflects this morphological trait, eliciting surprise when such stimuli are absent. In other words, an absence of visual stimulation would itself be surprising.

Now, the answer to the negative aspect of the dark room problem that is given here is one which generates the positive problem. The stipulation of a model which specifies the expected states (states that are not the dark room) is subject to the issue that motivation is solely driven by prior beliefs, leaving no room for pro-attitudes as traditionally conceived. It is thought that this crowding out of pro-attitudes by prior beliefs would lead to an impoverished behavioural repertoire that would fail to

include the ubiquitous tendencies towards play, exploration, exploitation, and behaviours do not seem to be best conceived as the minimisation of surprise. There are two quite general answers to be given at this point.

The first, as set out in Schwartenbeck et al. (Schwartenbeck et al. 2013), has to do with the way that prediction-error minimisation is formalised within PP. More specifically, when an action policy is selected amongst alternatives, the fact of uncertainty about outcomes will dictate that an agent attempts to visit many varied states with equal probability. That is because it will not be clear for the agent which states actually have the highest utility (construed in terms of prior beliefs). So there will be a shift between occupying the least surprising states and many novel states, depending on the level of uncertainty: "[…] when the differences in the expected utilities of outcomes become less differentiable, agents will try to visit several states and not just the state that has highest utility." (Schwartenbeck et al. 2013, p. 3) There will be a context-sensitive weighting of these two different kinds (exploitative and exploratory) kinds of strategy, where this weighting is influenced by the estimation of uncertainty through the assignment of precision as described in the first section of this paper.

In fact, this may seem to mirror the distinction between model-free and model-based modes of learning, as they are understood by both Pezzulo et al. (Pezzulo et al. 2015) and Clark (Clark 2016). That is because for them that distinction also appears to be a trade-off between two kinds of strategy that is arbitrated by the dynamics of precision assignment. Here, the agent would switch between exploratory and exploitative modes of engagement with the environment on the basis of how reliable their information is considered to be vis-à-vis the states that are least surprising (have the highest "value"). When such information is assigned very high precision values, then the agent engages in exploitative behaviours because there are states that are unambiguously more valuable than other states. But when such information is assigned low precision values, then the agent engages in exploratory behaviours because it is not sure which state will be most valuable (probable).

The other answer to be given is that the minimisation of prediction error takes place within the context of a hierarchical model, which means that the minimisation of surprise is an optimisation process that occurs over very many levels of spatiotemporal generality. With this in mind, it may well be the case that intuitive appeals to putative counter-examples where there are unlikely events that have very high utility do not sufficiently take into account deeper imperatives, or a balance between those imperatives and others in the hierarchy. For example, in response to the question of Gershman and Daw (Gershman and Daw 2012, p. 306), "[s]hould a person immersed in the "statistical bath" of poverty her entire life refuse a winning lottery ticket[?]", we might respond that although it is indeed true that for someone with a long history of poverty the state of sudden riches would be surprising, the deeper prior belief that compels the person towards keeping themselves fed or to acquire resources means that they will try to get themselves out of that situation of poverty if given such a chance. They won't refuse the ticket. So perhaps we can account for apparent counter-examples of this kind by appeal to distinctions between prior beliefs at different levels of hierarchical depth, or different levels of spatiotemporal generality. The deeper those beliefs are, the more likely they are to look like states with high utility rather than high prior probability.

There is a vexing question that comes up at this point of appeal to evolutionarily selected "deep" prior beliefs. One may first suppose that some of these deep priors need not be genetically innate but can be extracted from the environment itself. These would be priors that are extracted from highly consistent regularities within the environment, for example, the inability of two solid objects to simultaneously occupy the same space (cf., Hohwy et al. 2008, p. 692). However, it is unlikely that all such deep priors can be accounted for in this way. That is because many such priors will be idiosyncratic to whatever species the agent belongs to. Since the regularities that are extracted from the environment are presumably *in* the environment, they must remain constant across species. So if the deep prior in question is idiosyncratic to species (e.g., some prior or set of priors that produces a behavioral disposition to seek out dark environments in troglofauna, for instance), then it appears that it cannot

be learned from the environment. It must be innate. So the question I have in mind can be posed as follows.

We think that we know that brains perform many of their tasks in virtue of their performing active inference. Now it appears that in order for us to be able to explain motivation within a pure active-inference framework we need to posit innate priors. These innate priors are determined by the total morphology of the organism, insofar as this morphology is isomorphic with an optimum trajectory or a set of good-enough trajectories through state space. So the question is this: how does the non-neural morphology play the right role in active inference? This can't just be through its representation in active inference, because then there's no reason those innate priors don't just update when they encounter prediction error—why they are recalcitrant. Having the priors themselves be fixed morphological traits or processes (metabolism, for instance) addresses this issue, but then we have a problem about the computational interaction between neural and non-neural morphology. How is such interaction to be explained?

I will be arguing that the free energy principle, as developed and applied by Friston and his collaborators, provides one kind of answer to such questions. This is a kind of PP that is expanded in scope in order that it constitutes an explanation of biological adaptation in general, and on various timescales. In order to give a full reply to the dark room problem, the maximal-PP theorist is obliged to go further and either: i) embed PP within the wider scoped free energy principle; or ii) give an alternative account. I am open to the idea that there is an alternative account available, but in the rest of this paper I will be exploring (i).

3.3 The Free Energy Principle

It looks as though a satisfactory answer to questions about how we came to have the priors that we have can be given by embedding the maximal PP story within the wider scoped free energy principle. The free energy principle gives a surprise-minimisation account of biological processes in general, which affords us a way to explain the origin of the prior beliefs that are relevant to answering the dark room problem and the way in which those prior beliefs are related to morphological facts about the agent. But to see why this is so it's first necessary to set out the theory in sufficient detail.

We can start by explaining its initial motivation. The free energy principle is usually motivated by a much more general reflection on a putative distinction between biological and non-biological self-organising systems (e.g., Friston and Stephan 2007, §2.2). An example of the former kind might be a bird, or a bacterium. An example of the latter kind might be a snowflake, or a hurricane. Both of these kinds of complex system exhibit self-organisation; that means that they both spontaneously arrange themselves into an ordered pattern or structure without the intervention of an outside agent (Ashby 1962). However, Friston and Stephan (Friston and Stephan 2007) note a qualitative difference between them. The difference is that biological systems are *adaptive*. In the case of a snowflake, for instance, it will cross a phase boundary and melt with the change of temperature. But the "[…] key aspect of biological systems is that they act upon the environment to change their position within it, or relation to it, in a way that precludes extremes of temperature, pressure or other external fields." (Friston and Stephan 2007, p. 422)

The distinction may not be as stark as these authors suggest. After all, given sufficiently rapid and intense changes in temperature, biological systems also dissipate into the environment. That is to say that systems like snowflakes are not unique in this respect. Conversely, one may give potential examples of self-organising systems that act on their environment but that are non-biological. Aggregates of biological systems (like societies) act on their environments in some way, but it would be controversial to label these biological in the same way as their constituents are. But critique of this kind lacks propriety; Friston and Stephan are not doing conceptual analysis, they are suggesting constraints on the

behaviour of biological systems for heuristic purposes. Their question is this: how is it possible for a biological system to avoid dissipation? For this purpose, their loose distinction is sufficient.

One way to understand that capacity is in terms of an exchange of energy between the organism and its environment. This reflects the traditional biophysical understanding of biological systems as energetically open systems: they take in energy and matter in a low-entropy form as nutrition (construed broadly) and excrete it back into the environment in the form of relatively high-entropy waste. This allows us to reconcile the increase of complexity and order in living systems with the second law of thermodynamics, which states that entropy is always increasing in closed systems. The biological system is an open system, which allows it complexity and order at the expense of its surrounding environment (Schrödinger 1944).

Another way to understand this capacity of regulating the relationship to environment is in terms of information—it is to understand the capacity as that of moving around within a particular set of sensory states to which the system is suited. That set of states is implicitly specified by the phenotype of the system, because the system is already evolutionarily adapted to some specific environmental niche. A ferrophilic bacterium, for example, is adapted to a solution which contains specific levels of iron and oxygen. As such its phenotype will bear some substantive relation—perhaps representational (Shea 2012)—to this niche, and must alter its relationship to the environment in line with that relation. If the relation is construed in terms of representation, for instance, then it must regulate its relationship to the environment so that the propriety-conditions of that representation are satisfied.

These two ways of understanding biological systems—thermodynamic and informational—are complementary. Even though the environment is always becoming more and more disordered, it nonetheless behaves in a regular and lawful way, and the exploitation of this regularity makes it possible for the biological system to embed that regularity into its physical structure: "organisms could maintain configurational order, if they transcribed physical laws governing their environment into their structure." (Friston and Stephan 2007, p. 422)

How do biological systems manage to do this? On the free energy principle, the task is construed as a problem of Bayesian inference. The inference in question occurs across a boundary that segregates the internal states of the system and its external environment—this boundary is called a Markov blanket. Markov blankets consist of two kinds of state: sensory states and active states. The prototypical example of some such boundary is the cell wall. The task of the biological system is to infer the causes that act on it from the outside, with access only to the sensory states in the Markov blanket. The way that it does so can be modelled with the very same formalisms that govern active inference and belief updating in PP as applied to neurocognitive function. That is to say that the system approximates a model of the causes of its sensory states in two ways: by updating its internal states where those states fail to correspond to sensory states, and by acting on its environment in order to change the way that the outside causes generate sensory states.

These very abstract considerations can be illustrated more concretely with reference to the example of circadian rhythmicity (Bechtel 2011; Sheredos 2012). Circadian rhythms are periodic cycles which regulate other processes (metabolic, behavioural, genetic, and so on) on a roughly 24 hour period. These are sensitive to external cues (so-called Zeitgebers) in calibrating the clock, but the rhythm is endogenously produced, which means that its periodicity will remain in effect even in the absence of any Zeitgebers. Sheredos (Sheredos 2012) has argued that the circadian rhythms of cyanobacteria are cognitive in the minimal sense specified by the free energy principle. That is to say that the systems which perform circadian rhythmicity perform prediction-error minimisation. The circadian rhythm in the cyanobacterium regulates two metabolic processes that are chemically incompatible: photosynthesis (day-time) and nitrogen fixation (night-time). Roughly, the endogenous tendency of the system to a default period and phase of rhythm can be construed as the priors, and the sensitivity to Zeitgebers can be construed as producing prediction error in the system. In the cyanobacterium, the function of signalling prediction error is realised by high levels of phosphate, which are both produced

during photosynthesis and play a central role in transforming the protein which regulates the circadian cycle. So if the bacterium is unexpectedly performing photosynthesis at a time when it predicts there should be low levels of ambient light, the relatively high levels of phosphate will phosphorylate the protein regulating the circadian cycle, and this will recalibrate the clock. This feedback mechanism instantiates the hierarchical and bidirectional feedback mechanism that the free energy principle (and PP) describes, but it does so within a non-neural system. According to the free energy principle, all adaptive behaviour is like this.

The free energy principle may also be considered to apply over longer time scales (Friston 2013; Friston et al. 2015; Hobson and Friston 2016). One may distinguish functional (metabolism, learning, and inference), developmental, and even phylogenetic time scales in this regard. Natural selection takes place over very many generations of individual phenotypes; persistent self-organisation emerges from a fluctuating environment with a general tendency towards disorder, and at the expense of the order in that environment. Hobson and Friston (Hobson and Friston 2016) therefore suggest that natural selection can be explained under the same mathematical formalism as in PP when applied to brain function. Namely, natural selection is to be construed as a process wherein a genotype models the causal regularities (in this case, selection pressures) that impinge upon it from without, and in doing so embeds this causal structure into the phenotype across multiple generations. Each generation is understood to be a Bayesian "update" on the basis of prediction error that corresponds to selection pressures that the phenotype is not yet adapted to:

> […] in natural selection, each new generation corresponds to a Bayesian update, converting a prior distribution over phenotypic characteristics into a posterior distribution. […] this means that evolution is the process of predicting which phenotypes are best adapted to their econiche. (Hobson and Friston 2016, p. 247)

Again a specific example will be helpful. Circadian rhythmicity has a genetic component, as described in (Bechtel 2011). One could construe the endogenous tendency to rhythmicity as the "prior" which is updated on the basis of prediction error that results if Zeitgebers are out of sync with that prior. For instance, if a cyanobacterium is introduced into an environment where the day-night cycle is different (e.g., a different time zone), then its periodicity will be updated on this basis (the phase may well remain the same, if the length of the day itself is the same). However, we may ask questions about how it is that the periodicity has this endogeneity in the first place. The obvious answer is that it is genetic, but this just labels the problem and leaves the mechanism obscure. Bechtel (Bechtel 2011, p. 145) has shown that—in much the same way as the circadian rhythm itself—the genetic basis of the rhythm can be understood as instantiating a feedback mechanism that is regulated on the basis of error. The empirical research which is the basis of his discussion targets the circadian rhythm in fruit flies (Hardin et al. 1990). That research demonstrates that the genetic basis (the gene *per*) of the prior is down-regulated by high levels of the protein (PER) that it expresses; when there is a buildup of PER, this acts on the gene to prevent further generation of the protein. The period of this process is circadian: it occurs over a roughly 24-hour period.

With the foregoing example in mind, here's how the introduction of the free energy principle aids us in answering the questions about priors driving action. The first thing to point out is that the mechanism of free energy minimisation is not arbitrarily limited to instantiation in a *neural* realisation base; any realisation base that is sufficiently complex will be enough. The example of the circadian rhythms suggests that a chemical basis may be sufficient to realise active inference, and that a neural realisation is probably not necessary. The second thing to note is that the functional continuity between various otherwise distinct systems (genetic, cognitive, and behavioural) means that one can locate the origins of priors in different systems and across different timescales; one need not have a single centre of cognition which somehow respects facts about the morphology of the agent by means

of representing those facts in some way. Within this context, the relevant non-neural facts about the agent that make it so that the agent moves into and around a specific environmental niche are not themselves "dumb" or computationally inert, but are built into the computational machinery of active inference itself. So, in fact, there is no hard interface between the brain and the morphological facts that it requires access to in order to drive action in the right ways. So, in principle, both the origins of prior beliefs can be given along with a constraint on the account of how they input into processes of learning, inference, and action.

On the basis of these lines of reasoning, it seems to me that maximal PP that is grounded within the free energy principle is well-equipped to handle the concerns associated with the dark room problem. But the problem of triviality is yet significant. Clark gives voice to these concerns when he suggests that to excessively widen the scope of PP threatens "[…] to over-intellectualize large swatches of adaptive response in both human and non-human animals." (Clark 2013b, p. 8) But of course, the danger of the problem depends on what exactly this over-intellectualisation amounts to. If it commits us to saying that bacteria or genomes have attention, or imagine, or suffer from schizophrenia, then of course this seems like a debauched extension of anthropic notions into domains where they do not belong. But if we are simply placing functional requirements on those simpler systems, then this anxiety seems out of place.

Perhaps on this point we need to bite the bullet of Klein's suggestion – perhaps the free energy principle is a Galilean idealisation. However, something similar may be true of the laws of physics (Cartwright 1986), and these are nonetheless considered to be a significant advancement in scientific knowledge and highly valuable in constraining scientific models in more specific contexts. The same may be true of the free energy principle, which could serve to constrain theorising about specific mechanisms in cases like that of the circadian rhythm, neurocognitive function, and perhaps in future even social and aggregate entities (cf., Friston and Frith 2015). In that case its criteria for endorsement would be largely pragmatic.

4 Conclusion

The dark room problem is both plural and significant. I've tried to show here how the various concerns that constitute the problem emerge in the dialectic between PP-theorists and PP-critics. The initial puzzle—the negative problem—is an intuitive and naïve one. The question can be phrased like this: if action is just the minimisation of surprise, then why don't we try to minimise all stimuli? The answer to this question must either devalue the role of surprise-minimisation or explain why not all stimuli are surprising. Mixed PP views take the first horn of this dilemma by specifying mechanisms underlying action which do not work via surprise minimisation. One way to do so, I argued, is to appeal to "model-free" learning processes. These are reinforcement-learning schemes that do not require any representation of the way that events are statistically related to one another; they learn the value of different actions through trial and error.

Another way to construct a mixed PP view is to have both model-free and model-based processes integrated within a 'meta-model': within the predictive-processing architecture itself. Then, I argued, this just collapses into a maximal predictive processing view—that is the view that predictive processing is all that the brain ever does, and so all neurocognitive function must be explained in terms of surprise minimisation. If this view is endorsed, then the PP-theorist is taking the horn of the dilemma on which not all stimuli are surprising; that is because some are assigned a high probability within a model of the causes of sensory states. Then the maximal-PP theorist is obliged to respond to the positive aspect of the dark room problem: how does surprise-minimisation account for behavioural repertoires which include exploration and exploitation?

I suggested that the maximal-PP theorist can give two related answers. The first, following a suggestion by Schwartenbeck et al. (Schwartenbeck et al. 2013), is that exploratory and exploitative be-

haviours will be selected according to a trade-off that is driven by the dynamics of precision assignment. When beliefs about which states are "valuable" are imprecise, then the agent will try to occupy all of them (and find novel ones) in the exploratory mode; when beliefs about such states are precise, then the agent will just occupy those which are most valuable, in the exploitative mode. Second, appeals to prior beliefs which are deeper in the hierarchy can help explain why some states appear to have low probability but high value: it is because they entail deeper states that *do* have a high probability (winning a lottery entails having access to resources).

However, this raises puzzles about the origins of "deep" priors as well as how genetic information might interface with priors that are active in learning and inference in ontogenetic time. The free energy principle provides some suggestions here, though there is still much to be done in this regard. The example of circadian rhythms demonstrates how functional continuity can be established between free energy minimisation in both phylogenetic and ontogenetic time, thereby suggesting a relatively robust account of the way in which tendencies to particular kinds of action can originate in evolutionary processes. However, this raises questions about whether expanding Bayesian active inference to so wide a scope does not sap the free energy principle of any substantive empirical content. I think it is possible that this may be the case. But if it is, we may well go on to ask what such triviality amounts to if the account is both explanatory and of use as a heuristic. Certainly it must not be trivial in any sense that should worry us. But that is a challenge that may be taken up by the critics of PP in future.

References

Ainslie, G. (2001). *Breakdown of will.* Cambridge: Cambridge University Press.

Ashby, W. R. (1962). Principles of the self-organizing system. In H. H. von Forster & G. W. Zopf (Eds.) *Principles of self-organization: Transactions of the university of Illinois symposium* (pp. 255-278). London: Pergamon Press.

Barrett, L. F. & Simmons, W. K. (2015). Interoceptive predictions in the brain. *Nature Reviews Neuroscience, 16*, 419-429.

Bechtel, W. (2011). Representing time of day in circadian clocks. In A. Newen, B. Bartels & E.-M. Jung (Eds.) *Knowledge and representation* (pp. 129-162). Paderborn: Mentis.

Boring, E. G. (1930). A new ambiguous figure. *American Journal of Psychology, 42*, 444-445.

Bowers, J. S. & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin, 138*, 389-414.

Cartwright, N. (1986). *How the laws of physics lie.* Oxford: Clarendon Press.

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

———(2013b). The many faces of precision (Replies to commentaries on "Whatever next?"). *Frontiers in Psychology, 4* (270).

——— (2016). *Surfing uncertainty: Prediction, action, and the embodied mind.* New York: Oxford University Press.

Craig, A. D. (2003). Interoception: The sense of the physiological condition of the body. *Current Opinion in Neurobiology, 13*, 500-505.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P. & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron, 69*, 1204-1215.

Dretske, F. I. (1981). *Knowledge and the flow of information.* Cambridge, MA: MIT Press.

Fecher, G. T. (1873). *Einige Ideen zur Schöpfungs- und Entwickelungsgeschichte der Organismen.* Leipzig: Breitkopt & Härtel.

Freud, S. (1950 [1895]). Project for a scientific psychology. In J. Strachey (Ed.) *The standard edition of the complete psychological works of Sigmund Freud* (pp. 281-391). London: The Hogarth Press and the Institute of Psychoanalysis.

Friston, K. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences, 13*, 293-301.

———(2013). Life as we know it. *Journal of the Royal Society Interface, 10*.

Friston, K. & Frith, C. (2015). A duet for one. *Consciousness and Cognition, 36*, 390–405. https://dx.doi.org/10.1016/j.concog.2014.12.003.

Friston, K. J. & Stephan, K. E. (2007). Free-energy and the brain. *Synthese, 159*, 417-458.

Friston, K., Adams, R. A., Perrinet, L. & Breakspear, M. (2012a). Perceptions as hypotheses: Saccades as experiments. *Frontiers in Psychology, 3* (151).

Friston, K., Thornton, C. & Clark, A. (2012b). Free-energy minimization and the dark-room problem. *Frontiers in Psychology, 3* (130).

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

Gershman, S. J. & Daw, N. D. (2012). Perception, action and utility: The tangled skein. In M. I. Rabinovich, K. J. Friston & P. Verona (Eds.) *Principles of brain dynamics: Global state interactions* (pp. 293-312). Cambridge: MIT Press.

Gläscher, J., Daw, N., Dayan, P. & O'Doherty, J. P. (2010). States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron, 66*, 585-595.

Hardin, P. E., Hall, J. C. & Roshbash, M. (1990). Feedback of the Drosophila period gene product on circadian cycling of its messenger RNA levels. *Nature, 343*, 536-540.

Hobson, J. A. & Friston, K. J. (2016). A response to our theatre critics. *Journal of Consciousness Studies, 23*, 245-254.

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

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

———(In Press). The predictive processing hypothesis and 4e cognition. In A. Newen, L. Bruin & S. Gallagher (Eds.) *The Oxford handbook of cognition: Embodied, embedded, enactive and extended*. Oxford: Oxford University Press.

Hohwy, J., Roepstorff, A. & Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition, 108*, 687-701.

Klein, C. (in press). What do predictive coders want? *Synthese*. https://dx.doi.org/10.1007/s11229-016-1250-6.

McMullin, E. (1985). Galilean idealization. *Studies in the History and Philosophy of Science Part A, 16*, 247-273.

Millikan, R. G. (1984). *Language, thought, and other biological categories: New foundations for realism*. Cambridge, MA: MIT Press.

Mumford, D. (1992). On the computational architecture of the neocortex. *Biological Cybernetics, 66*, 241-251.

Pezzulo, G., Rigoli, F. & Friston, K. (2015). Active inference, homeostatic regulation and adaptive behavioural control. *Progress in Neurobiology, 134*, 17-35.

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

Ruse, M. (2008). Darwinian evolutionary theory: Its structure and its mechanism. In M. Ruse (Ed.) *The Oxford handbook of philosophy of biology* (pp. 34-63). Oxford: Oxford University Press.

Schrödinger, E. (1944). *What is life? The physical aspect of the living cell*. Cambridge: Cambridge University Press.

Schwartenbeck, P., FitzGerald, T., Dolan, R. J. & Friston, K. (2013). Exploration, novelty, surprise, and free energy minimization. *Frontiers in Psychology, 4* (710).

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences, 17*, 565-573.

Shea, N. (2012). Inherited representations are read in development. *British Journal for the Philosophy of Science, 64*, 1-31.

———(2013). Naturalising representational content. *Philosophy Compass, 8*, 496-509.

Sheredos, B. (2012). Reductio ad bacterium: The ubiquity of Bayesian "brains" and the goals of cognitive science. *Frontiers in Psychology, 3* (498).

Sims, A. (2016). A problem of scope for the free energy principle as a theory of cognition. *Philosophical Psychology, 29*, 967-980.

Vance, J. (2015). Review of the predictive mind. *Notre Dame Philosophical Reviews*.