How to Entrain Your Evil Demon

Jakob Hohwy

The notion that the brain is a prediction error minimizer entails, via the notion of Markov blankets and self-evidencing, a form of global scepticism — an inability to rule out evil demon scenarios. This type of scepticism is viewed by some as a sign of a fatally flawed conception of mind and cognition. Here I discuss whether this scepticism is ameliorated by acknowledging the role of action in the most ambitious approach to prediction error minimization, namely under the free energy principle. I argue that the scepticism remains but that the role of action in the free energy principle constrains the demon's work. This yields new insights about the free energy principle, epistemology, and the place of mind in nature.

Keywords

Active inference | Agency | Approximate bayesian inference | Coupled oscillation | Epistemic value | Evil demon | Exact inference | Free energy principle | Functional role semantics | Internal models | Internalism | Interventionism | Markov blanket | Perceptual inference | Perceptual learning | Prediction error minimization | Scepticism | Self-evidencing | Variational bayesian inference

Acknowledgements

Thanks to Regina Fabry and Thomas Metzinger for a discussion that lead to this paper; thanks to two anonymous referees and the editors of this volume.

1 Prediction Error Minimization and the Free Energy Principle

An emerging, unified theory of brain function seeks to explain all aspects of mind and cognition as the upshots of prediction error minimization (Friston 2003; Friston 2010; Hohwy 2010; Clark 2013; Hohwy 2013; Clark 2016b). The idea is that the brain is a model of its environment, which garners evidence for itself by explaining away sensory input. This happens in a process of approximate Bayesian inference, where hypotheses about sensory input are generated from the model, and the predictions of these hypotheses tested against the actual input.

The extent to which these predictions are correct determines the accuracy of the model. Essentially, the better the model is at minimizing the error in its predictions (the prediction error) the more evidence it accumulates for itself. In the course of maximizing evidence for itself, the model infers the causes of its sensory input. To illustrate, it seems reasonable to say that if I hypothesize that the current sound is caused by an approaching train, predict that soon the sound will therefore get louder and then recede, and then confirm this prediction, then I have correctly inferred something about the causes of my evidence. This is enshrined in perceptual inference and perceptual learning where an internal model of hidden causes is optimized through prediction error minimization.

Exact inference would correspond to following Bayes' rule for updating perceptual beliefs however it is unlikely the brain engages in exact inference since for realistic perceptual settings inverting the model that generates the predictions to infer the hidden causes presents an intractable computational problem. Instead, there is some reason to think that the brain can engage in approximate inference

Hohwy, J. (2017). How to Entrain Your Evil Demon.
In T. Metzinger & W. Wiese (Eds.). *Philosophy and Predictive Processing:* 2. Frankfurt am Main: MIND Group. doi: 10.15502/9783958573048

(Friston 2003; Friston 2005). In exact inference, prediction error is minimized over the long term perspective, namely as the model becomes increasingly better. This leads to the idea that a system that minimizes prediction error on average and over the long term is likely to approximate the outcomes of exact Bayesian inference. Approximate inference (especially variational Bayesian inference) does not present intractable problems and importantly can be executed by systematically varying internal model parameters for a system that just has access to its own internal states and the states of its sensory organs. The move to approximate inference is attractive as it overcomes formal obstacles for conceiving the brain as an inferential system. Moreover, there is a good case that it is biologically plausible since it speaks to the overall architecture of the brain in terms of not just relatively sparse and focused forward (bottom-up) connectivity but also hitherto poorly understood, more diffuse and copious backwards (top-down) connectivity; perceptual inference and learning through prediction error minimization in the brain is also arguably consistent with the different types of plasticity operating at different time scales and different hierarchical levels in the brain (Mumford 1992; Friston 2005).

The free energy principle (Friston 2010) sets notions of perception and action in just the long term perspective needed for prediction error minimization. The principle begins with the observation that persisting organisms maintain themselves in a limited set of states, rather than dispersing through all states. This is a statistical notion because it allows a description of the organism in terms of a probability density, or model, which identifies the states in which it is most probable to find it. The principle then states that organisms manage to maintain themselves in their expected states by minimizing their free energy, which (given assumptions about the shape of the probability densities) is the long-term average prediction error, given their model (for an introduction to the free energy principle, see Hohwy 2015).

Crucially, the free energy principle imbues organisms with agency, such that they can act to maintain themselves in their expected states. This happens through prediction error minimization: predictions are held stable rather than revised in the light of immediate prediction error, and action moves the organism around to change the input to the senses until the predictions come true. Since prediction error minimization in the long run approximates Bayesian inference, action can be said to be an inferential process in the same sense as perception is inferential. Hence, action is labeled 'active inference' (Friston and Stephan 2007; Friston 2010). Active inference increases the accuracy of internal models and perceptual inference optimizes these models. Active inference is a powerful addition to the inferential framework because it allows the agent to, firstly, increase the epistemic value of their model (e.g., confirming that it is really a train approaching by opening the window facing the tracks to hear the sound better), and secondly, to occupy one's expected states (e.g., predicting that one is travelling on the train, noting the ensuing prediction error, and moving around in the environment until one is in fact in a train and the prediction error is minimized) (see Hohwy 2013, Clark 2015, Clark 2016a, for extensive discussion of the role of active inference in a prediction error minimization framework).

This chapter explores the free energy principle through a very particular, epistemological lens. It seems that the principle entails global scepticism. This could be viewed as a significant epistemological problem, and I will discuss that towards the end of the paper. But entailment of scepticism is also at times viewed as a symptom of a fatally flawed, old-fashioned model of the mind, at odds with contemporary embodied, enactive and extended (EEE) models of mind. These models make action central to mind and cognition, and in some cases imply that the inclusion of action removes the scepticism. This in turn gives rise to the hypothesis that the free energy principle, with its strong focus on action, could avoid scepticism and be an underlying theoretical framework for EEE models. Here, I argue that, action notwithstanding, scepticism remains for the free energy principle and indeed the addition of action does nothing to remove the inferential, internalist aspect of the prediction error minimization approach. The more constructive part of the chapter explores how active inference nevertheless does change the epistemic landscape and how this is reflected in the representational properties of the

internal prediction-generating model. This seems to retain the inferential and internalist aspects of prediction error minimization but is able to accommodate to some degree several EEE insights.

The chapter begins by introducing the notions of Markov blankets and self-evidencing, which gives a principled and relatively clear understanding of inferentialism and internalism within a prediction error minimization scheme. It then moves on to explain how scepticism is entailed by such schemes and how adding action does not make scepticism go away. The chapter then moves on to a discussion of the interesting way in which action does change our conception of the epistemic status of agents and what this tells us about internal models. Finally, the chapter attempts to assess what this all means for our overall understanding of the free energy principle.

2 Markov Blankets and Self-Evidencing.

From the previous section's brief description of prediction error minimization under the free energy principle, a somewhat unusual conception of a biological agent emerges on which the agent simply is a model, which is engaged in prediction error minimization through action and through optimization of model parameters. It may seem incongruous to equate 'agent' with 'model', but, on the free energy principle, models are the things that do the acting, based on their representation of the world, and which (therefore) persist through time. It seems reasonable to label as an 'agent', at least in a rudimentary sense, an acting, representing, persisting thing. As we will see, this equation of agent with model in turn has epistemological and theoretical implications.

One way of describing an agent as a model begins with the causal nets terms of a Markov blanket (Pearl 1988; Friston 2013; Hohwy 2016b), where a Markov blanket for a node in a causal net is the node's parents, children and parents of its children (Fig. 1).

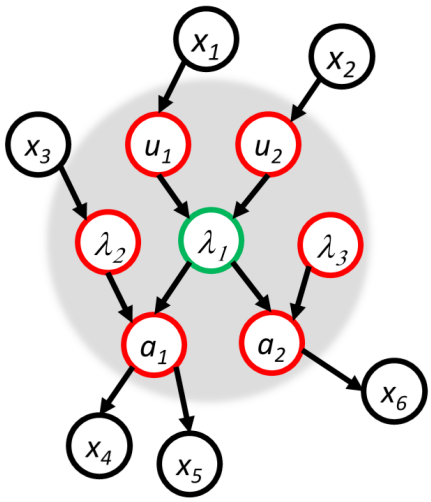

Figure 1. Markov blanket. The behavior of the green node λ_1 is known once the red nodes of the blanket (parents u_i, children a_i and parents of children $\lambda_{2,3}$) are observed; states x are external to the blanket and do not need to be observed to know the internal states.

The behavior of the blanketed node (node λ_1 in Fig 1) is in principle predictable just from observation of the states of the blanket, without knowing anything about the states of the external nodes beyond the blanket (that is, every node in the network is conditionally independent of λ_1 when conditioned on the nodes of the blanket).

The free energy agent maps onto the Markov blanket in the following way. The internal, blanketed states constitute the model. The children of the model are the active states that drive action through prediction error minimization in active inference, and the sensory states are the parents of the model,

driving inference. If the system minimizes free energy — or the long term average prediction error — then the hidden causes beyond the blanket are inferred.

Delineating the agent in terms of a Markov blanket means we can conceive succinctly of the behaviour that makes something an agent, in terms of a model that is self-evidencing. The notion of self-evidencing comes from Hempel's discussions of scientific explanation and captures the idea, firstly, that a hypothesis is supported to the extent it can explain away evidence and, secondly, that the "information or assumption that [the evidence] occurs forms an indispensable part of the only available evidential support for [the hypothesis]" (Hempel 1965, pp. 372-4). Above we said that the agent garners evidence for its model to the extent it can encounter and explain away sensory input (in much the way a scientist might garner evidence for their theory). But here there is not one thing, the agent, which provides evidence for another, the model: the model changes its sensory and active states to increase evidence for itself. These processes within the Markov blanket happen at the subpersonal level — there is no agent *making* the changes to the model. Hence, the notion of self-evidencing makes better sense of the initially incongruous idea that the model is the agent. The more the internal and active states change to anticipate the changes in sensory states, the more the organism evidences itself and thereby manages to persist (Hohwy 2016b).

Ultimately, this still somewhat challenging equation of model with agency may need to be fleshed out by connecting the notion of self-evidencing to the notion of self-organisation. For present purposes, we note that self-evidencing relative to a Markov blanket, which defines the borders of the agent, is a core ingredient of a prediction error minimization account.

3 Introducing the Evil Demon

Crucially, self-evidencing means we can understand the formation of a well-evidenced model, in terms of the existence of its Markov blanket: if the Markov blanket breaks down, the model is destroyed (there literally ceases to be evidence for its existence), and the agent disappears. The model infers the presumably distal hidden causes of its sensory input, and thereby garners evidence for itself, and this entire self-evidencing process can be understood just in terms of the causal behaviour of the Markov blanket and the internal nodes, without giving the model any kind of direct access to the hidden causes (even though naturally no-one but a solipsist would assume there exist no hidden causes whatsoever). The nodes of the internal model have access to the states of the blanket, and can gauge the prediction error but the model only represents the hidden external causes vicariously. The model builds up rich, detailed, and context-sensitive associations amongst the sensory states, and between sensory and active states, and learns which of these many possible associations tend to keep prediction error low in the long run.

This raises a familiar sceptical spectre: the model will not be able to distinguish between possibilities where similar flows of sensory input are caused by two very different causal processes, beyond the blanket. In the first possibility, the causes are as we suppose them to be, the familiar people, houses, trains, trees, fruits, etc. that we perceive in everyday life. In the second possibility, versions of which are familiar from much philosophy going back to Descartes' *Meditations*, the sensory input is caused by an evil demon (or evil scientist) with total control of the sensory states of the agent.

On the demon scenario, it seems the very same states that were assumed to represent people, houses, trains, trees, fruits, etc. are now not representing those things. They are all misrepresentations because really the hidden causes now belong to the cunning machinery and states of the evil demon. The prediction error minimization scheme therefore entails scepticism. There is never any justification for any perceptual belief that p because the evidence for those beliefs cannot exclude the possibility that not-p. It is difficult to see how one could both be committed to the prediction error minimization framework and yet prevent this descent into scepticism (for earlier versions of probabilistic schemes and the entailed scepticism, see Eliasmith 2000, Usher 2001, Grush 2003).

Many different kinds of philosophical frameworks entail scepticism, and it is not surprising that such an empiricist kind of framework as the prediction error scheme entails scepticism. A kindred framework, for example, is Thomas Metzinger's phenomenal self-model framework. Metzinger notes how our phenomenal, perceptual experience of the world seems certain to us but then notes that "[e] pistemologically speaking, however, the subjective experience of certainty transported by … phenomenal content is unjustified. [P]henomenal content supervenes on internal and contemporaneous properties of the human brain. A disembodied brain in a vat, therefore, could in principle realize precisely the same degree of experiential certainty going along with our ordinary conscious experience of embodiment" (Metzinger 2004, p. 310). Metzinger is not here primarily interested in the epistemic threat from scepticism but rather in what the advent of scepticism can teach us about a particular conception of mind and consciousness. Though I shall discuss the epistemic ramifications of scepticism later in this paper, the initial focus is the same: skepticism as a testing ground for understanding and assessing competing models of the mind.

In the contemporary debate, entailment of scepticism is at times viewed as symptomatic of a poorly conceived model of the mind that operates with a representation-hungry internalist machine sandwiched between worldly input and bodily action (Hurley 1998). The reasoning is straightforward. Skepticism arises for systems with insuperable sensory veils, such that there is only, and at best, indirect access to the external world, namely by internal representations of external states of affairs. Due to its heavy inferential nature and invocation of Markov blankets, it may appear that the prediction error minimization model is the epitome of scepticism-inducing cognitive science (Anderson and Chemero 2013). Thompson and Cosmelli , likewise, see the evil demon scenario as an important testing ground for the debate whether perception is 'Brainbound' or 'Enactive' (Thompson and Cosmelli 2011). They argue that, if we look at the explanatory aspects of enactive theories, and seriously consider what it would take to maintain an evil demon (or brain in a vat) scenario, then we should accept that the scenario is not possible. Cosmelli and Thompson use this in an argument against brainbound conceptions of the mind, including the type that the prediction error minimization account seems committed to (Thompson 2007, p. 242); I return briefly to their views below.

The alternative is a model that sees mind as non-representational, extended, situated, embodied or enactive (Hurley 1998; Noë 2004; Gallagher 2005; Thompson 2007; Clark 2008; Hutto and Myin 2013). The reasoning with respect to scepticism, on behalf of such EEE views, is that skepticism will not arise if the sensory veil is obliterated and that it is obliterated by having the mind be extended into or embedded into the environment, by not operating with representations, by making the body essential to the operation of the mind, or by having fluid and direct enactive commerce with the world.

The advantage of these alternative models is not merely that they seem to avoid scepticism — though this is not ignored altogether it is rarely the focus of EEE theories — but that they build into the core conception of mind the fact that minds belong to agents with bodies that live and act in environments with other agents. It is seen as misguided to build models of the mind that leave all that matters (body, world, action, others) as mere incidental afterthoughts to the rarified internal, representational workings of the mind.

There are some nice questions about how exactly EEE models of the mind avoid scepticism. These models seem so revisionary about our internal, mental workings that traditional notions of knowledge, belief and justification threaten to become obsolete. Scepticism might be off the table but it may be that basic epistemic conceptions go missing too. For example, if there is no principled distinction between belief and what the belief is about, then it is difficult to see how there can be beliefs at all. It would be a pyrrhic victory if scepticism is conquered but at the price of the very conceptions of knowledge, belief and justification.

Here, I will consider a possible way out of this quandary. Specifically, I will consider if the addition of *action* to the prediction error minimization scheme is sufficient to overcome scepticism while both respecting some insights of the alternative EEE models of the mind and retaining the virtues of

prediction error minimization as a unified explanation of perception, knowledge, and action. There are several attractions to considering this way out of the quandary. First, action in the shape of active inference is a key part of the prediction error minimization scheme conceived under the free energy principle, so it would be odd to mount a defense of prediction error minimization schemes without giving action a central role. The importance of action is recognised clearly by both main philosophical treatments of the scheme, which repeatedly appeal to active inference (Hohwy 2013; Clark 2016b). Second, action is a linchpin of EEE approaches, which overwhelmingly appeal to action in arguments against the staid, passive model of the brain as merely engaged in computational operations over internal representations (see references above). To be clear, in this chapter, I focus on action as a treatment for scepticism within the prediction error minimization scheme. This is an interesting exercise because, as we shall see, it will tell us something about the scheme and its epistemological consequences. This means I am not here offering a full treatment of the connection between EEE accounts and prediction error minimization. That is a task for another occasion (for a first treatment, see Hohwy 2016b, and for others Bruineberg and Rietveld 2014, Clark 2015, Bruineberg 2016, Kirchoff 2016). However, the discussion offered here is, I believe, an important testing ground for some of these debates.

Before moving on to discussing whether the addition of active inference changes the game with respect to scepticism, I will briefly discuss, and then set aside, an argument that denies a role for scepticism for some of these debates. Clark (Clark 2016a, Sec. 3) argues that scepticism and evil demon scenarios are irrelevant to assessing the compatibility of prediction error minimization schemes and (some) EEE approaches. He points out that scepticism is unlikely to disappear even on embodied cognition type views, or, in other words that the emergence of scepticism is a red herring in these debates, something we should not focus on. I have sympathy for this position: it agrees with me that the prediction error minimization scheme entails scepticism though it disagrees concerning the significance of this for assessing embodied cognition. However, as we saw above, Clark's position is at odds with others who are advocating EEE approaches: they see scepticism as a clear symptom of misguided, internalist, non-EEE theories.

The reason Clark can reasonably assert that scepticism is a red herring is that, for him "the claim that lies at the heart of recent work on the embodied mind [is that it] fundamentally rejects […] the richly reconstructive model of perception" (Clark 2016a, p. 12). That is, the embodied mind debate concerns "the question whether apt actions are always and everywhere computed by using sensing to get enough information into the system to allow it to plot its response by exploring an internally represented recapitulation of the distal world" (Clark 2016a, p. 11). I think he is right that the issue of whether internal models are rich and reconstructive is orthogonal to the issue of whether scepticism is entailed by the view or not. What this should tell us, however, is not that the issue of scepticism is irrelevant *tout court* to all kinds of EEE views but rather that it is irrelevant specifically to the issue of rich and reconstructive models.

Notice that Clark invokes active inference in his own conception of the embodied mind and his argument that it is consistent with the prediction error minimization scheme. He says "The appearance of conflict [between the embodied mind and prediction error mimimization schemes] arises from ambiguities in the notions of inference and seclusion themselves. For these notions may seem to imply the presence of a rich inner recapitulation of the distal environment, with a consequent downgrading of the role of action and upgrading of the role of reasoning defined over that inner model" (Clark 2016a, p. 11). The difference between Clark and others, like (Thompson and Cosmelli 2011) and (Anderson and Chemero 2013) is then that the type of action-oriented embodied mind that Clark defends is not threatened by the scheme's entailment of scepticism.

This leads me to a clarificatory point about the notion of internalism. It is tempting to say that any account of perception and cognition that operates with internal models must in some sense be internalist. But the natural next question is what makes internal models internal? I think a natural default has been to answer that internal models are internal because they are housed in the brain. But this

cannot be right. The notion of internal models belongs with machine learning and computational science and as such cannot be necessarily wedded to biological organs. A better answer is provided by the notion of Markov blankets and self-evidencing through approximation to Bayesian inference. Here there is a principled distinction between the internal, known causes as they are inferred by the model and the external, hidden causes on the other side of the Markov blanket. This seems a clear way to define internalism as a view of the mind according to which perceptual and cognitive processing all happen within the internal model, or, equivalently, within the Markov blanket. This is then what non-internalist views must deny (for further discussion of where to place the blanket, and how much it can encompass, see Hohwy 2016b).

Notice that this definition of internalism makes Clark an internalist (as emphasized by Anderson and Chemero 2013) and the main difference between Clark's and my own interpretations of prediction error minimization schemes then seems to be whether internal models are rich and reconstructive, and whether they downgrade the role for action. That discussion is for another occasion. This chapter focuses on the issue of skepticism.

4 Active Inference and Sensory Veil Skepticism

Active inference, as we saw in previous sections, builds on the idea that sensory input can become better predictable through making changes to the active states of the Markov blanket. The system is able to learn that certain active states are associated with certain sensory states. For example, there may be a learnt association between my finger actively exerting force on the button of the remote control and a change in sensory states as the TV (television) channel changes. If my expected state is that I am in fact watching the news rather than the soap, then I might prioritise the actually false hypothesis that my finger is pressing the button. The hypothesis generates a prediction of sensory input, which is fleshed out partly in terms of expected bodily states, for example that there will be a certain proprioceptive input related to arm and finger movements. This prediction engenders a prediction error, which is minimized as the arm and finger are entrained via reflex arcs and movement actually occurs (see Hohwy 2016a, for more discussion of active inference).

Adding active inference to the prediction error minimization story does not however make that story any less inferential. Action is inference in the same way as perception is inference, namely by approximating Bayes through prediction error minimization. The processing, from the point of view of the system, is still told entirely in terms of self-evidencing within the Markov blanket: it is mainly a matter of hierarchical statistical associations between patterns of states of the nodes of the blanket and the internal states. In perceptual inference, the internal model might build up expectations for how sensory states unfold over time, for example, between sensory states u_1 and u_2 (see Fig. 1). In active inference, there will also be learnt associations between states of the blanket, for example between active state a_1 and sensory state u_2, given internal states λ_1.

This means that even though action is part of the story we can still in principle understand all that the brain does simply in terms of approximate inference and while 'throwing away the world' beyond the Markov blanket (i.e., all the x nodes in Fig. 1). Naturally, 'throwing away the world' should be taken in an explanatory rather than literal sense. For the purposes of explaining mind and cognition, we just need to know how the Markov blanket system — the brain in this case — constructs these probabilistic mappings among its own states (namely by the message passing underlying perceptual and active inference in the cortical hierarchy). We do not need to know what the worldly states are, even though, obviously, they exist and play a causal role via the interface of the blanket in causing the internal states to be the way they are.

Given this internalist perspective, there is no reason to think that adding action to the prediction error scheme will make any immediate difference to the issue of scepticism. Scepticism arose as a consequence of the fact that prediction error minimization occurs within the sensory veil, and active

inference does not change this fundamental aspect — active inference does not break down the Markov blanket. In fact, the existence of the Markov blanket is evidenced by action.

The free energy principle's marriage of perception with action thus has little prospect of avoiding the traditional internalist picture of mind. This is an intriguing observation because, in other respects, there is much in common between the free energy principle and the EEE approaches to cognition (Friston 2011; Bruineberg and Rietveld 2014; Bruineberg 2016; Clark 2016a; Hohwy 2016b; Kirchoff 2016). However, at the same time, the free energy principle seems at odds with those approaches since it absorbs action within the Markov blanket. The free energy principle is unlikely to be the scheme that will be the perfect partner for existing EEE approaches to the mind. It may be that a new type of understanding is needed which can combine — synthesise — insights from both the EEE approaches and more traditional internalist approaches under the free energy principle.

5 Epistemic Advantages of Action

In the setting of the free energy principle, action cannot help evade scepticism — and the sandwich model of the mind — in the way some EEE approaches might have hoped for. This is not to say that there are no epistemic advantages of action. For example, it is a commonplace that we can act to explore the world and find out how it works. There exist in the literature considerations about whether action can deal with scepticism. These considerations happen within a traditional, inferential framework rather than by obliterating the sensory veil through EEE approaches. Consider here Reichenbach's cubical world (Reichenbach 1938, § 14; Fig. 2), which speaks to the kind of Markov blanket scenario we are considering.

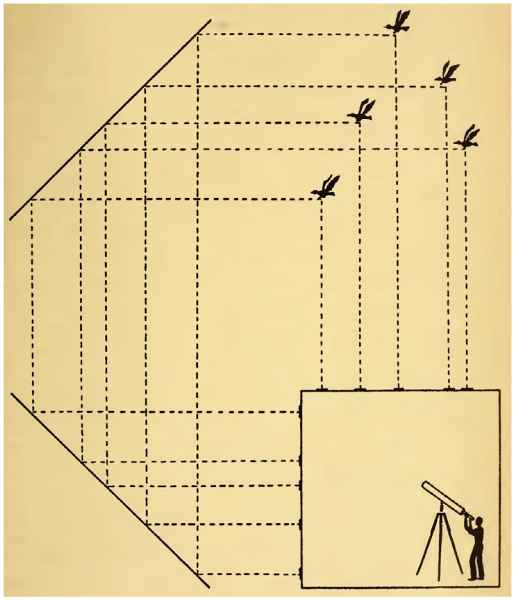

Figure 2. Reichenbach's cubical world (Reichenbach 1938, p. 117). The observer in the cubical world must infer the hidden causes (flocks of birds) on the basis of their reflections cast on the semi-transparent walls of the world, some directly, some cast indirectly via mirrors.

The observer is locked in a room with translucent walls (conceive of this as the Markov blanket) on one of which the reflections of five birds are projected. At the same time, a set-up of mirrors projects the reflections of the five birds to another wall, giving a total of ten reflections. The observer must infer the causes of their reflection input without leaving the cubical world. The question is whether the observer can select between the hypotheses that there are five birds and that there are ten birds, and,

further, whether anything in the situation could convince the observer that there are any birds out there at all.

Reichenbach argues that the coincidence between the movement of the two sets of reflections is evidence that there is a common cause (i.e., five birds), and that if there is a common cause it must be beyond the cube. Sober (Sober 2011) argues that this is not efficient, since it presupposes something not given in the evidence, namely that common cause is more likely than coincidence given the observation of coincidence. This prior seems to presuppose too much.

Sober then appeals to interventionist approaches to causal inference. According to interventionism, it is possible to distinguish common causes from coincidence by intervening — acting — on the relevant states of affairs (by 'surgically' varying and holding fixed certain random variables). By acting to hold one of two coinciding variables constant one may discover they are not related as cause and effect, namely if the behavior of the other is not changing after the intervention. This is some evidence that they have a common cause, evidence which is unlikely to be obtained without intervention. Though this is epistemic progress through action, it does not however exclude the possibility that this common cause is somehow internal to the cubical world (or Markov blanket), so it cannot speak to external world scepticism; for example, Sober discusses the possibility that the common cause is a prior intention. Neither can it rule out the evil demon possibility, which is indeed a common cause of sensory input.

Notice here that the intervention is on the hidden causes in the world, conducted via our active states (e.g., brain states entraining our limbs so we move the hand to scare the birds) and detected via our sense organs (e.g., noticing whether there is a change in all the reflections). All the perceptual and active inferential work is done within the Markov blanket.

These kinds of examples show that action can have epistemic value. There are certain questions in causal inference that cannot easily be answered without action. This role of action is part and parcel of active inference and the free energy principle. This corresponds to how active inference was described above; action can minimize prediction error in two ways, to maximize utility (help us occupy expected, low prediction error states) and to maximize epistemic value in the ways discussed in the interventionist framework (Hohwy 2013, Ch. 4, Friston et al. 2015). What we are considering here is active inference for epistemic value about the external causes of sensory input and their relations amongst each other, to our sensory organs, and from our active states.

Even if actions can make such inroads on some epistemic issues for agents considered in terms of Markov blankets, it is no solace to the EEE approaches since it remains a wholly internal, inferential approach. And of course none of this speaks to the evil demon scenario where the issue is not whether the external world exists but whether the external world harbours a demon or not. It is difficult to see how any intervention could increase the probability that the sensory input is not caused by an evil demon since the demon, we must assume, will make sure the sensory input does not reveal its own existence.

6 How Action Entrains the Environment

I have argued that action cannot be used to extricate the prediction error minimization scheme from the sceptical scenario: action does not obliterate the sensory veil nor can it be used to favour directly the non-sceptical hypothesis.

Consider however what happens during active inference. For example, the internal states predict that the proprioceptive input is of the type that occurs when the arm raised. Since the arm, let us assume, is not currently raised, this prediction leads to prediction error. The proprioceptive prediction, at the active states of the Markov blanket, triggers reflex arcs to move the limbs around until the prediction is satisfied at the sensory states of the blanket, registering the occurrence of the expected proprioceptive input. Here hidden states external to the Markov blanket (the limbs) are causally af-

fected by states internal to the blanket, and in turn cause changes to the blanket and its internal states. This portrays the central dynamic nature, or circular causality, of the free energy principle, see Fig. 3 (notice that formally speaking causal nets apply to acyclic graphs, not cyclic ones as depicted in Fig. 3.; there are ways around this, for example by appealing to dynamic Bayes nets).

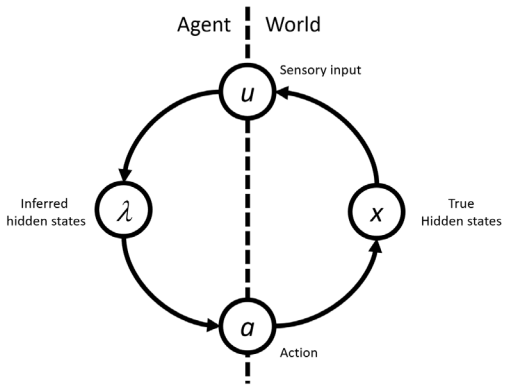

Figure 3. A simplified version of the Markov blanket shown in Fig. 1 but with depiction of circular causality induced by active inference. The dotted line marks the mind-world divide or Markov blanket; u are the sensory states of the Markov blanket, a are the active states and λ the internal states; x are the true hidden states of the world.

In this scenario, no evidence available to the agent distinguishes between the hypothesis that they really have an arm and the hypothesis that an evil demon is deceiving them to think they have an arm. Note that the internal states of the agent are part of a dynamic causal chain that modulates the states of the hidden causes. This holds whether the external states harbor real arms and other familiar things, or a demon. We are assuming the world contains either of these, and that they causally impact the agent's sensory states. In the demon world, this means the demon's states causally entrain the agent's internal states. But equally, through active inference, the agent's states causally entrain the demon's states. If the demon is not entrained like this, then the agent's prediction errors would not be minimized in active inference. (I am setting aside the 'lucky demon' scenario where the causal link from a to x is cut and yet the demon luckily guesses what sensory input the agent predicts; I am also setting aside the solipsistic scenario where there is a direct causal link from a to u; similarly, I am setting aside the case of an 'active Laplacian demon' who manipulates u on the basis of perfect knowledge of the laws of nature and power to set the initial conditions of the world).

Put slightly differently and assuming our demon is an agent with its own Markov blanket, just as the demon's input to the agent works like a learning signal for the agent's internal models (through prediction error) so the agent's output to the demon works like a learning signal for the demon's internal states. The demon thus infers the internal states of the envatted agent, and it acts to minimize its own prediction error. The demon's internal states must in turn have causal power in the world so that they can affect the agent's sensory states in the predicted ways. In very basic terms, here the agent and the demon begin to 'oscillate' together — they are essentially locked together in a system of coupled oscillation. As Friston points out, this is basically a version of the pair of pendulum clocks described in the 1600s by Huygens' (Huygens 1967, p. 185), which when hung from a beam in the right kind of set-up will eventually begin to swing synchronously even if separated by each their own Markov blanket (Friston 2013; Fig. 4).

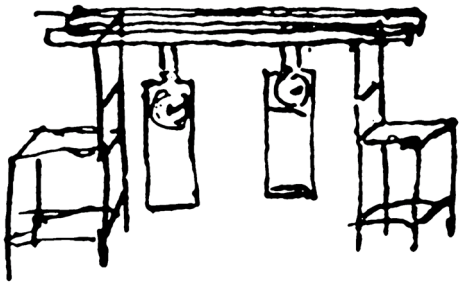

Figure 4. Huygens' sketch of coupled oscillation between two pendulum clocks.

Coupling the agent and the demon like this levels the playing field between them. They are tied in a causal dance of give and take. It may be that the demon started things off by giving the agent its basic expectations (as evolution arguably does in the non-demon case) but this is not tantamount to knowing a priori what active inference the agent will engage in since the agent is a self-organising, noisy system with some autonomy, individual learning history, and ability to vary parameters in approximate Bayesian inference. For example, the agent may experiment with the balance between perceptual and active inference in unanticipated ways; sometimes it may optimize the prediction error bound on surprise in a sustained manner before engaging action, other times it may jump to action before perceptual inference has been fully optimized; similarly, at times the agent may act on the expectation that exploration of the free energy landscape is the best way to keep long term error low (Hohwy 2013, Chs. 4, 7). Hence the demon can do nothing but follow suit to the agent's attempts to change the evidence through action to fit its predictions. The only alternative avenue for a recalcitrant demon is to let the prediction error of the agent increase. In the short run, this will just cause the agent to optimize its internal model before it acts, rather than jump to active inference. In the long run, this choice is however only open if the demon wants the agent to perish since, as the free energy principle sets out, the existence of the agent depends on continually acquiring evidence for itself.

Above we encountered the discussion of scepticism by the enactivists (Thompson and Cosmelli 2011). Their point was that for a demon to actually operate a brain-in-a-vat it seems nomologically necessary that the brain is coupled with a body, or something that in some respect is functionally equivalent to a normal body, in an environment conducive to the brain's autonomous homeostatic functioning. There is something right about this, in the sense that external causes, harboured in the demon and its world, must exist for the agent to exist. More, Cosmelli and Thompson appeal to the dynamics of bodily existence, which makes it imperative for the demon to provide a body-like environment. In this sense, their argument provides a more rudimentary, body-bound version of the more general appeal to active inference that I am rehearsing here. Cosmelli and Thompson argue that their approach entails that the mind is enactive and embodied rather than brain-bound, even on the demon scenario; they use this to propose that scepticism in some sense is self-undermining. However, as argued above, if there is a Markov blanket, then this appeal to the necessary existence of body-like external causes does not make the sceptical scenario self-undermining. It is a given that there are external causes including those that keep the body and brain of the agent alive, but the evil-demon scepticism remains regardless.

Consider now what this kind of causal coupling, which I have argued follows from the free energy principle, entails about the epistemic and representational state of the agent. The predictions of the agent are based on its internal hierarchical model of the external world. This model carries information about statistical relations and causal interactions amongst the modeled causes at several interlocked time-scales. For example, it will represent that arms can be raised to greet friends or hail cabs except when one is very tired or something heavy is placed on the arm, and so on. This representation is

specified in the hierarchy by having several levels of inferred, interacting causes that when convolved in the right way best minimize long term prediction error and thus explain away the sensory input.

If we consider this model as a large 'theory' specifying all these statistical and causal relations, then nothing much is lost by stripping away the names of properties (i.e., the property names 'arms', 'friends', 'cabs' etc.) from the statistical and causal information and instead operating with existential quantification over the properties. This leaves behind a statistical/causal functional role that can be used to implicitly define what those properties are. For example, for there to be an 'arm' is for there to be something that has the role of probably being raised when 'friends' are nearby, and which tends to interact causally with 'tiredness' and 'heavy objects', where 'friends', 'tiredness' and 'heavy object' are in turn interdefined by their own partly overlapping statistical/causal roles (see Hohwy 2013, Ch. 8, and references therein for this view, which is inspired by standard functional role semantics). Conceiving the internal model like this reflects that what makes prediction error minimization succeed is only information about the causal-statistical properties of the input, rather than about the intrinsic aspects of the objects themselves.

Conceived like this, there is no difference between the internal representations of the agent in the demon world and in the non-demon world. The statistical/causal roles harboured in the internal model of the agent are they same, and both are blind to the difference between real arms, friends and cabs and the simulacra of these in the demon scenario. These statistical/causal roles are however what gives the internal states their content, and it seems their satisfaction conditions are the same in the two worlds. The demon world will have the action-induced statistical/causal properties, and the definite descriptions defining those properties will quantify over some of the demon's hidden causes.

From the point of view of the free energy principle, there is a sense in which the agent need not care about the fact that their internal model fails to distinguish the demon and non-demon possibility. All the agent cares about is self-evidencing — maintaining itself in its expected states — and as long as prediction error is kept low it succeeds in doing this. Since what the agent needs to help keep prediction error low is statistical/causal information, the functional role conception of their internal model is apt. Ontologically, that is, there may be an evidence-transcendent difference, for example in terms of arms vs. non-arm demon properties that can have no effect on statistical/causal relations (this echoes discussion of Matrix-type scenarios in Chalmers 2005, but set in the context of functional role semantics and active inference).

Epistemically, not only can the agent keep the prediction error low, they must also be getting something right about the causal/statistical structure of the hidden causes, even those states that partly originate in a demon. This epistemic success follows from the entraining of the hidden causes through active inference. If prediction error is minimized through action, then there must be something out there such that (modulo the overall level of prediction error minimization and irreducible noise) it stands in the modelled causal/statistical relations to each other and the agent.

If the story is told without appeal to active inference, then the world need not be entrained to the agent's internal states. An entirely passive perceptual system may be able to minimize prediction error over some time scale but must more promiscuously change its internal model in whatever way will explain away the sensory data. In such a system, which does not bend the world to its expectations, organisms (if any should endure at all) are more short-lived and there is more scope for false, hallucinatory models.

Of course, the causal/statistical epistemic success of an agent engaging in active inference is consistent with considerable accompanying ignorance. The agent's unintentional social demon cognition may fall short; for example, in the demon world there is a deeply hidden cause that is not modelled successful, namely some of the demon's own mental states (such as the demon's desire to keep the agent alive). Similarly, in the non-demon world, there are probably levels of natural law governing deep seated regularities, which are nevertheless as yet undiscovered or even undiscoverable for us.

7 Concluding Remarks

Even if an evil demon is causing an agent's sensory input, the agent's actions entrain the demon to cause less surprising sensory input. This leaves the skeptical scenario in place but dulls some of its epistemic sting. The demon world and the non-demon world must have overlapping statistical/causal structure, given the agent's internal model, and the agent can know this structure even if the intrinsic structure and some more deeply hidden causes remain unknown.

As far as competing models of the mind goes, we saw that skeptical scenarios are viewed as a symptom of undesirable internalism that ignores the role of action, embodiment, and extension of mental states into the external world. Some embodied, extended and enactive (EEE) approaches on the other hand avoid the sceptical scenario but arguably at the price of undermining familiar notions of knowledge, belief and justification.

A prediction error minimization scheme is wedded to self-evidencing in the context of Markov blankets and thereby to a principled way of distinguishing mind from world in terms of internal inference of the hidden causes located externally to the Markov blanket. Such a scheme can accommodate action, through the free energy principle. As argued in this chapter, however, including action cannot change the scheme's staunchly internalist nature.

Accommodating action does have an interesting consequence, of relevance to the demon scenario. The actions of a creature who manages to persist over time must entrain the causes in its world (if any such causes exist), whether these are the familiar causes we expect to exist in the world, or whether these are causes harboured in an evil demon. We also saw that since active inference must entrain the causes of the external world, a free energy minimizing system can know at least some of the statistical/causal structure of the world, whether the demon scenario is true or not. This all adds up to an advantage for the free energy approach to mind. It is an account of the mind that can make reasonable and interesting, albeit limited, inroads on the evil demon scenario and, all the while, provide a unified if internalist and inferentialist explanation of perception and action.

This advantage for the free energy approach is achieved by going in what seems to be the opposite direction from EEE approaches. Many of those approaches seek to obliterate the Markov blanket, and attempt to make perception less inferential, more like action and more connected to the world. The free energy approach instead makes action inferential and conceives it as just a matter of internal processing within the Markov blanket.

It would be a mistake to view this as entirely anathema to the EEE approaches. The internalist and inferentialist view does not conceive of the mind or the brain as causally insulated from the world around it. Indeed, this view must conceive of the mind and the world as causally linked, through the causal interface of the Markov blanket, since perception is inference on the causes of sensory input and active inference entrains these causes to fit the internal model, leading to coupled oscillation and reciprocal mirroring of mind and world.

Similarly, under the free energy principle, agents are conceived in terms of the states of the world that they tend to occupy, under the assumption that their Markov blanket will begin to disintegrate, and the agent begin to disperse, if action ceases to minimize prediction error. The mind of an agent is thus just another set of causes in the overall causal nexus, albeit a self-organising type of cause. So the free energy principle delivers a dual perspective that seems to supersede some of the existing debates: epistemic insulation, exemplified with the sceptical demon scenario, goes hand in hand with more embodied, situated causal integration or oscillation (Hohwy 2013, p. 228). In this light, the free energy principle seems to offer progress — a sort of synthesis of internalism and EEE — in our debate about mind and world: it provides a unified perspective on the epistemic and causal status of the mind.

References

Anderson, M. & Chemero, A. (2013). The problem with brain GUTs: Conflation of different senses of "prediction" threatens metaphysical disaster. *Behavioral & Brain Sciences, 36*, 204–205.

Bruineberg, J., Kiverstein, J. & Rietveld, E. (2016). The anticipating brain is not a scientist: The free-energy principle from an ecological enactive perspective. *Synthese*. https://dx.doi.org/10.1007/s11229-016-1239-1.

Bruineberg, J. & Rietveld, E. (2014). Self-organization, free energy minimization, and optimal grip on a field of affordances. *Frontiers in Human Neuroscience, 8*. https://dx.doi.org/10.3389/fnhum.2014.00599.

Chalmers, D. J. (2005). The matrix as metaphysics. In C. Grau (Ed.) *Philosophers Explore the Matrix*, Oxford: Oxford University Press.

Clark, A. (2008). *Supersizing the mind: Embodiment, action, and cognitive extension.* Oxford University Press, USA.

——— (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral & Brain Sciences, 36* (3), 181–204.

——— (2015). In T. K. Metzinger & J. M. Windt (Eds.) *Embodied prediction*. *Open MIND:* 7(T). Frankfurt am Main: MIND Group. https://dx.doi.org/10.15502/9783958570115.

——— (2016a). Busting out: Predictive brains, embodied minds, and the puzzle of the evidentiary veil. *Noûs*. https://dx.doi.org/10.1111/nous.12140.

———(2016b). *Surfing uncertainty.* New York: Oxford University Press.

Eliasmith, C. (2000). *How neurons mean: A neurocomputational theory of representational content.* Ph.D., Washington University in St. Louis.

Friston, K. (2003). Learning and inference in the brain. *Neural Networks, 16* (9), 1325-1352.

Friston, K. J. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society London: Biological Sciences, 369* (1456), 815–836.

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

———(2011). Embodied inference: Or "I think therefore I am, if I am what I think". *The Implications of Embodiment.* W. Wolfgang Tschacher and C. Bergomi. Sussex, Imprint Academic.

———(2013). Life as we know it. *Journal of The Royal Society: Interface, 10* (86). https://dx.doi.org/10.1098/rsif.2013.0475.

Friston, K. & Stephan, K. (2007). Free energy and the brain. *Synthese, 159* (3), 417-458.

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

Gallagher, S. (2005). *How the body shapes the mind.* Oxford: Oxford University Press.

Grush, R. (2003). In defense of some 'Cartesian' assumptions concerning the brain and its operation. *Biology and Philosophy, 18* (1), 53–93. https://dx.doi.org/10.1023/A:1023344808741.

Hempel, C. G. (1965). *Aspects of scientific explanation and other essays in the philosophy of science.* New York: Free Press.

Hohwy, J. (2010). In W. Christensen, E. Schier & J. Sutton (Eds.) *The hypothesis testing brain: Some philosophical applications* (pp. 135–144). Macquarie Centre for Cognitive Science. https://dx.doi.org/10.5096/ASCS200922.

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

——— (2015). In T. Metzinger & J. M. Windt (Eds.) *The neural organ explains the mind* (pp. 1-23). Frankfurt am Main: MIND Group. https://dx.doi.org/10.15502/9783958570016.

———(2016a). Prediction, agency, and body ownership. In: *The Pragmatic Turn: Toward Action-Oriented Views in Cognitive Science,* ed. A. K. Engel, K. J. Friston, and D. Kragic. Strüngmann Forum Reports, vol. 18, J. Lupp, series editor. Cambridge, MA: MIT Press.

———(2016b). The self-evidencing brain. *Noûs, 50* (2), 259-285. https://dx.doi.org/10.1111/nous.12062.

Hurley, S.L. (1998). *Consciousness in action.* Harvard University Press.

Hutto, D. & Myin, E. (2013). *Radicalizing enactivism: Basic minds without content.* Cambridge, Mass.: MIT Press.

Huygens, C. (1967). *Œuvres complétes.* Amsterdam: Swets & Zeitlinger.

Kirchoff, M. (2016). Autopoiesis, free energy, and the life–Mind continuity thesis. *Synthese*. https://dx.doi.org/10.1007/s11229-016-1100-6.

Metzinger, T. (2004). *Being no one: The self-model theory of subjectivity.* MIT Press (MA).

Mumford, D. (1992). On the computational architecture of the neocortex. II. The role of cortico-cortical loops. *Biological Cybernetics, 66*, 241–251. https://dx.doi.org/10.1007/BF00198477.

Noë, A. (2004). *Action in perception*. Cambridge, MA: MIT Press.

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

Reichenbach, H. (1938). *Experience and prediction—An analysis of the foundations and structure of knowledge*. Chicago: University of Chicago Press.

Sober, E. (2011). Reichenbach's cubical universe and the problem of the external world. *Synthese, 181* (1), 3–21. https://dx.doi.org/10.1007/s11229-009-9593-x.

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

Thompson, E. & Cosmelli, D. (2011). Brain in a vat or body in a world? Brainbound versus enactive views of experience. *Philosophical Topics, 39*, 163-180.

Usher, M. (2001). A statistical referential theory of content: Using information theory to account for misrepresentation. *Mind & Language, 16* (3), 311–334. https://dx.doi.org/10.1111/1468-0017.00172.