How to Knit Your Own Markov Blanket:

Resisting the Second Law with Metamorphic Minds

Andy Clark

Hohwy (Hohwy 2016, Hohwy 2017) argues there is a tension between the free energy principle and leading depictions of mind as embodied, enactive, and extended (so-called 'EEE[1] cognition'). The tension is traced to the importance, in free energy formulations, of a conception of mind and agency that depends upon the presence of a 'Markov blanket' demarcating the agent from the surrounding world. In what follows I show that the Markov blanket considerations do not, in fact, lead to the kinds of tension that Hohwy depicts. On the contrary, they actively favour the EEE story. This is because the Markov property, as exemplified in biological agents, picks out neither a unique nor a stationary boundary. It is this multiplicity and mutability– rather than the absence of agent-environment boundaries as such- that EEE cognition celebrates.

Keywords

Active inference | Autopoiesis | EEE cognition | Embodied cognition | Evil demon | Extended mind | Free energy minimization | Markov blanket | Prediction error minimization | Process ontology

Acknowledgements

Thanks to Karl Friston and two anonymous referees for extremely helpful comments on an earlier draft, to John Dupré for invaluable discussions concerning the extended mind and process ontologies, and to Giovanna Colombetti and Joel Krueger for participating in and making possible the 'Breaking Boundaries' symposium at the University of Exeter, where some of these ideas started to take shape. This paper was drafted during a period of sabbatical leave (Autumn 2016) kindly granted by the University of Edinburgh, and completed as part of ERC Advanced Grant XSPECT - DLV-692739.

"My cousin has great changes coming – one day he'll wake with wings"
Cousin Caterpillar (The Incredible String Band)

1 The Markov Blanket Conception of Mind

Markov blankets are named after Andrey Markov (1856-1922), a mathematician whose seminal work explored abstract systems that remember their past trajectories only insofar as they store a single (current) value. In such systems (Markov chains – see Norris 1998) the next state depends only on the value of the current state. This is the so-called Markov *property*. For this reason, such systems are sometimes said to be 'memoryless'.

Now consider a complex system composed of many interacting nodes (variables). Pearl (Pearl 1988) introduced the term 'Markov blanket' to describe the set of nodes such that, for some given node X, the behavior of X could be fully predicted just by knowing the states of those other nodes. The states of those neighbouring nodes thus fix (statistically, not causally) the state of the target node conditionally independently of all the other states of the system, forming a 'Markov blanket' that shields

1 Many of the core ideas of EEE cognition are lately referred to using the even larger grouping '4E Cognition' (see e.g. Newen et al. in press) rather than EEE cognition. This adds 'embedded', in the sense defended by (Rupert 2009), to the E-pantheon. I have used the 'EEE' branding partly because that is the label used by Hohwy (Hohwy 2016, Hohwy 2017) in the 'evil demon' papers to which the current work responds. 'EEE Cognition' also provides the most apt label for the contested issues, as it is claims associated with embodied, enactive, and extended (rather than merely 'embedded') cognition that are the locus of our residual disagreements concerning the conceptual implications of predictive processing.

Clark, A. (2017). How to Knit Your Own Markov Blanket: Resisting the Second Law with Metamorphic Minds.
In T. Metzinger & W. Wiese (Eds.). *Philosophy and Predictive Processing:* 3. Frankfurt am Main: MIND Group. doi: 10.15502/9783958573031

the target node from the rest of the activity in the system. The practical upshot is that, to predict the state of the target node, all you ever need to know are the states of the nodes that form its Markov blanket[2]. The Markov blanket comprises the so-called 'parents' and 'children' of the blanketed node or nodes, corresponding to the most proximal actors upon the node (the parents) and the most proximal 'acted-upons' (the children), along with whatever else acts upon the children (the parents of the children). Markov blankets can be redundant, in that a target node may be enclosed within many Markov blankets. Markov blanketed organizations may nest within larger Markov blanketed organization (see figure 1), a property that will be important for the treatment that follows. Finally, a Markov *boundary* (Pearl 1988) exists when a Markov blanket for a node has no proper subset that is also a Markov blanket for that node – it is thus the most minimal (or 'non-redundant') Markov blanket for the node.

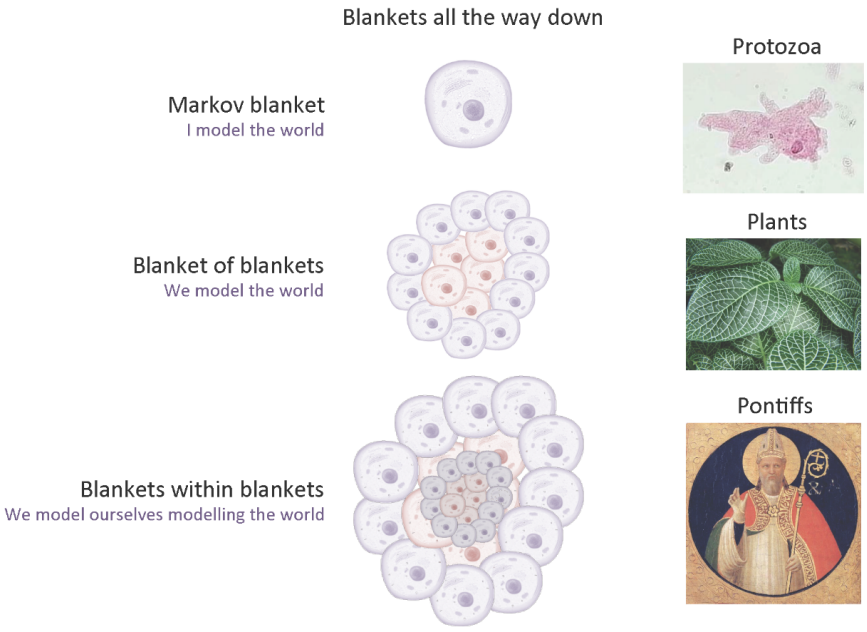

Figure 1: Blankets of blankets of blankets. Image by Karl Friston (by permission).

Hohwy's (Hohwy 2016, Hohwy 2017) arguments for a form of 'neurocentric seclusion' depict the states of our biological sensory systems as determining a Markov blanket that (he claims) defines the boundaries of the mind. Thus we read that:

> the mind begins where sensory input is delivered through exteroceptive, proprioceptive and interoceptive receptors and ends where proprioceptive predictions are delivered, mainly in the spinal cord. (Hohwy 2016, p. 276)

Adding in a footnote (footnote 14) that:

> In more technical terms (see Friston 2013), the sensory input and active output at this boundary forms a so-called Markov blanket (Pearl 1988) such that observation of the states of these parts of the system, together with observation of the prior expectations of the system in principle will allow prediction of the behavior of the system as such. Causes beyond this blanket, such as bodily states

2 This does not imply that, just by knowing the states of the blanket, a theorist could perfectly predict the evolution of the target (contained) nodes. This is because the blanketed system may have its own intrinsic dynamics. Thanks to Wanja Wiese for suggesting this clarification.

> or external states, are rendered uninformative once the states of the blanket are known. (Hohwy 2016, p. 283; citation style adapted)

The mind, if this is correct, is firmly bounded by the sensory flows associated with exteroceptive, interoceptive, and proprioceptive signaling. Step beyond those sensory flows and (Hohwy argues) the Markov property bites, rendering all that led up to those sensory stimulations statistically otiose as a predictor of systemic response. That also means that the mind, thus conceived, has no access to states of the world beyond that provided by the evidence available at the Markov blanket. Such minds "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" (Hohwy 2017, sect. 3). Hence they might be fooled by a clever Cartesian demon, or by the keepers of the Matrix.

2 The Markov Blanket Conception of Agents

Hohwy (Hohwy 2016, Hohwy 2017) also offers the Markov blanket construct as a formalization of some important claims concerning the nature and definition of agents. The considerations here weave together the Markov blanket considerations and the free energy principle (FEP) described in (Friston 2010), (Friston and Stephan 2007) and elsewhere.

FEP (more on which shortly) leverages the simple truth that agents that exist do so because they are able to persist, appearing to resist – for periods of time - the second law of thermodynamics that states that entropy (disorder) increases over time[3]. Biological agents are able to resist because the second law applies only to isolated (or closed) systems. By exchanging matter and energy with the environment, such systems are able to preserve their own integrity and order. They do so, of course, only by increasing disorder elsewhere (thus 'obeying' the second law). We thus enter the realm of living or adaptive systems – systems that actively seek out, and work to bring about, the conditions that are necessary for their own survival[4].

Such agents are, in an important sense, defined by the particular way they resist disorder. A specific type of living agent simply *is* a set of states that maintain themselves within certain bounds – the bounds that describe the conditions necessary for their own survival. Such bounds include, for human agents, acceptable ranges of body temperature, the production of glucose to power the metabolism that enables foraging, and so on. Only while a whole host of such states remain within set or tolerable ranges does the agent (qua that very living being) exist. This adaptive tautology implies that, in a very broad sense, every living creature must visit and revisit the set of states that define it as the creature it is. In this sense, a living organization is said to be (at least locally) 'ergodic' – see (Friston 2013, sect. 4).

This is where FEP gets invited onto the stage. FEP states that living organisms that persist must minimize free energy in their exchanges with the environment. The 'free energy' in question here is an information-theoretic isomorph of thermodynamic free energy, which is a measure of the energy available to do useful work. Useful work, in the information-theoretic story, involves fitting a model to a domain, so reducing information-theoretic free energy is improving the model. The free energy is then a bound on the long-term average 'surprisal' (Tribus 1961) associated with environmental engagements, where this names the implausibility of some sensory state given a model of the world. Entropy, in this information-theoretic rendition, is the long-term average of surprisal. Reducing information-theoretic free energy thus amounts to improving the model so as to reduce (long-term average) surprisal. Organisms that minimize long-term average surprisal will, by definition, appear to resist the second law of thermodynamics. They will take steps to avoid environmental states and encounters that would cause them to undergo catastrophic ('surprisal-ing') phase-transitions. They

3 More accurately, the second law tells us that entropy in an isolated system will either stay the same or increase over time. It can remain constant only in certain idealized (e.g. 'steady-state') cases. But whenever energy is transferred or transformed, entropy increases.

4 This formulation suggests – correctly in my view – strong links with the notion of 'autopoesis' (Varela et al. 1974). See also section 3 following.

achieve this by being structured in ways that in effect 'exchange entropy' with the environment, allowing them to self-organize so as to avoid, for a while at least, the kinds of catastrophic encounters that would cause them to cease to exist.

Hohwy links the FEP to the Markov blanket considerations by identifying the free energy minimizing agent or model with the internal states screened off by the Markov blanket. It is in this sense that, as Hohwy (Hohwy 2017, sect. 2) puts it "the internal, blanketed states constitute the model". It is important to note that the free energy minimizing 'model' here is that which is bounded by sensory and active (action-causing) states. This model is then said to be self-evidencing in the sense of (Hempel 1965). Self-evidencing is typically exemplified by cases such as the following:

> the velocity of recession of a galaxy explains the redshift of its characteristic spectrum, even if the observation of that shift is an essential part of the scientist's evidence that the galaxy is indeed receding at that the specified velocity. (Lipton 2001, pp. 44–5)

But in the case at hand, the very existence of a whole living system (a plant or an animal) provides evidence for itself considered as a surprisal-minimizing model. In other words, the goodness of the system as a means of exchanging entropy with the environment so as to persist in the face of the second law[5] is (self) evidenced by its own existence.

Notice that nothing in this Markov blanket conception of biological agents requires those agents' brains or control systems to engage in online prediction error minimization at all[6]. Hohwy's treatment (Hohwy 2017) thus covers two issues that – though deeply related - are also importantly distinct. The first concerns the status of living systems as self-evidencing free energy minimizing systems. The second concerns the specific vision of the human brain as implementing a process of prediction error minimization.

Thus consider a very simple free-energy minimizing life-form, such as a single-celled organism capable of survival-enhancing chemotaxis. Such a life-form may respond to environmental perturbations using a variety of tricks and ploys, none of which require it to engage in a process in which incoming sensory stimulations are met with attempts to generate the incoming signal 'from the top down' using stored knowledge about the world. Such a being, though living and perfectly able to resist the second law by exchanging entropy with its environment, could be operating in a purely 'feed-forward' manner, responding to detected chemical gradients in ways not nuanced by any form of top-down predictive flow. Talk of such a being 'predicting' such-and-such, or 'minimizing prediction error with respect to such and such' is either simply false or merely short-hand for what is really a rather different claim – the claim that the creature is structured so as to favour the kinds of environment necessary for its own persistence.

To describe this whole simple (reactive, feed-forward) creature as a 'model' of its world, though common in this literature, can also seem somewhat strained. An intelligent agent might harbour an explicit model of the layout of her own house, and use it to drive various kinds of 'offline' reasoning (such as counting the windows while thousands of miles away). An embodied agent might also use parts of her own body as a model, for example by counting the windows using her fingers. These are core and familiar uses of the notion of a model[7].

5 Or more precisely, in the face of the 'fluctuation theorem'. This applies to far-from-equilibrium systems (such as living beings), and has the second law as a special case. See (Friston and Stephan 2007).

6 Notice that the mere fact that some creature (a simple feed-forward robot, for example) is not engaging in active online prediction error minimization in no way renders the appeal to a Markov blanket unexplanatory with respect to that creature. The discovery of a Markov blanket indicates the presence of some kind of boundary responsible for those statistical independencies. The crucial thing to notice, however, is that those boundaries are often both malleable (over time) and multiple (at a given time), as we shall see.

7 For example, in everyday use the word 'model' might be used to mean a scaled down replica such as a toy model of a boat - but not the boat itself. If I said 'the boat is a model of the sea' most people would take me to be speaking metaphorically. Yet in the technical sense used by Friston and others, the boat is quite literally a model of the dynamics of the sea. It is wise to keep these differences in mind when assessing the claim that the free energy perspective makes agents into surprise minimizing 'models' of their worlds.

The matter is complicated, however, by the self-evidencing that is inherent in free energy or surprisal minimization itself. To see why, notice that negative surprisal is also the 'Bayesian model evidence' (see Friston 2013) associated with some specific Markov blanket and its internal states. Surprisal is the negative log probability of sensory states and action given that Markov blanket. In turn, this is the Bayesian model evidence for the Markov blanket itself. In this (mathematically quite deep) sense minimizing free energy or surprisal is the same as maximizing model evidence for the existence of the Markov blanket. The free energy principle, considered as an imperative for biological self-organization, thus necessarily entails some form of self-evidencing[8].

The upshot is that any adaptive system (any system that persists in the face of a changing environment) must display self-evidencing, and might properly be described as self-evidencing a certain 'model of the world'. This is because any system that successfully 'deals with' its environment, so as temporarily to resist the second law, counts as modelling its environment in this somewhat technical sense. Nevertheless, models (organisms or systems) that are self-evidencing in this specific sense need not rely upon top-down predictions to structure and inform their exchanges with the wider world. Predictive processing thus constitutes a biologically plausible process theory that may or may not be implemented in any given biological system. For example, there may be simple systems (such as bacteria and viruses) that minimize their surprisal through a genetically pre-configured response to sensory perturbations. By the same token, there may be other more complicated systems that resist increases in free energy or entropy in part by relying upon some form of predictive processing[9].

Hohwy (Hohwy 2017, sect. 3) also asks "what makes internal models internal?" and once again answers by appealing to Markov blankets and self-evidencing. The suggestion seems to be that the real internalist commitment is not to cognitive processing being in some important sense 'brain-bound' (in the sense of Clark 2008). Rather, since the organism itself is now (at least in the specific mathematical sense just described) the model, the notion of the 'inner model' collapses into the claim that there exists a Markov blanket organization separating the whole acting organism from the wider world. But whatever its other merits, this is not a plausible reconstruction of the notion of internalism at issue in the literature on EEE cognition. For EEE theorists do not seek to deny the existence or importance of systemic boundaries blanketing the organism from the wider world. Instead, such theorists (see Clark 2003, Clark 2008) stress the multiplicity, flexibility, and transformability of those boundaries, and the way the choice of what boundaries to stress reflects the explanatory interests and projects of the theorist. This is the main issue to be pursued in the rest of this treatment.

Putting all this together, every biological system is treated (by Hohwy) as a self-evidencing 'model' of the survival-relevant web of interacting hidden causes partitioned on the other side of a Markov blanket constituted by a set of sensory and active states – a blanket that also defines it as the very system (creature) it is. The system tracks the wider world only via the changing states of the Markov blanket, so that "[t]he nodes of the internal model have access to the states of the blanket […] but the model only represents the external causes vicariously" (Hohwy 2017, sect. 3). This also provides Hohwy with the opening for his interestingly twisted form of 'evil demon' scepticism, as we shall shortly see.

3 EEE Cognition Revisited

Summarizing his own recommended approach, Hohwy comments that:

> Many of those [EEE – embodied, extended, enactive] approaches seek to obliterate the Markov blanket, and attempt to make perception less inferential, more like action and more connected to

8 Thanks to Karl Friston (personal communication) for extremely useful discussion of this issue.

9 Among those systems, those equipped with generative models that enable them actively and systematically to anticipate how the world will alter in response to their own possible future actions plausibly constitute the sub-class most demanding of the full predictive processing interpretation – see (Pezzulo et al. 2015).

> the world. The free energy approach instead makes action inferential and conceives it as just a matter of internal processing within the Markov blanket. (Hohwy 2017, sect. 7)

What is right about the EEE stories, Hohwy immediately suggests, is just the strong emphasis on causal interactions between mind and world – interactions mediated via "the causal interface of the Markov blanket" and resulting in some kind of "reciprocal mirroring of mind and world".

I have argued elsewhere (Clark in press) that EEE approaches are best seen as opposed to substantial notions of mind-world mirroring – the kinds of mirroring that depicts minds as harbouring rich enough internal recapitulations of reality to enable us to do most of our cognitive work using traditional brain-bound inner models rather than by making the most of the opportunities provided by body, world, and action. But suppose the free-energy minimizing 'model' that does the mirroring is in fact the whole embodied, active organism. We can, if we wish, describe this as an organism's having "transcribed physical laws governing their environment into their structure" or speak of systems that "embed [the laws that govern environmental unfoldings] into their anatomy" (both quotes from Friston and Stephan 2007, p. 422). But the notion of mirroring itself is now hugely emaciated, effectively reduced to that of (non-accidentally) doing whatever it takes to ensure persistence in the face of the second law of thermodynamics. Every simple trick and ploy that has been celebrated, in the EEE literature, as a means of securing adaptive success thus counts (relative to this undemanding conception) as the organisms 'mirroring' their environment[10]. Now, the organism 'mirrors' the environment much as the shape of the boat (recall note 6 above) 'mirrors' the dynamics of the ocean. At this point, some theorists may prefer to drop the appeal to mirroring *tout court*.

Moreover, the emphasis on causal interactions, once all that is taken into account, seems entirely of a piece with even the most radical versions of EEE. For example, citing Wiener's Cybernetics (Wiener 1961), Maturana and Varela's autopoesis (Varela et al. 1974), Chiel and Beer's neuroethology (Chiel and Beer 1997), and Clark's situatedness (Clark 1997), a leading group of situated roboticists write that:

> In this view, adaptive behavior can best be understood within the context of the (biomechanics of the) body, the (structure of the organism's) environment, and the continuous exchange of signals/energy between the nervous system, the body, and the environment. Hence the appropriate question to ask is not what the neural basis of adaptive behavior is, but what the contributions of all components of the coupled system to adaptive behavior and their mutual interactions are (Mohan et al. 2013, p. 17)

The free energy minimizing story, as we briefly saw, delivers a perfect and principled fit with just these kinds of consideration. It depicts the whole embodied organism, appropriately coupled with the larger environment, as the system relative to which free energy (and surprise/prediction error) is minimized. To that extent, it strikes me as a story that counts in favour of core tenets of the EEE conception: not merely one that merely preserves the best of EEE while avoiding mistakes and excesses.

But rather than dwell on these issues, it may be helpful to look a little harder at Hohwy's conception of the EEE project itself. Many EEE approaches, Hohwy commented (op cit, sect. 7) "seek to obliterate the Markov blanket". This, it seems to me, is Hohwy's core reservation. It marks the crucial spot at which EEE approaches, in Hohwy's view, go wrong. But even granted the wide diversity of work that falls under the EEE umbrella, I am not persuaded that, on the whole, they seek to obliterate the Markov blanket at all. Instead, they aim only to reveal something that is highly compatible with the larger story on offer viz that the Markov property, as exemplified in biological agents, picks out neither a unique nor a stationary boundary. It is this multiplicity and mutability – rather than the absence of

10 For useful surveys of many of those tricks and ploys, see (Clark 1997, Pfeifer and Bongard 2006).

agent-environment boundaries as such - that EEE cognition celebrates, as we'll see in more detail in section 4 following.

Indeed, a functional analogue of the Markov blanket construct already plays a key role in seminal work on the 'enactive' approach that stresses the importance of 'autopoiesis' (Varela et al. 1974) – a form of organization in which the constituent processes actively produce the components needed to maintain themselves, and hence maintain the organization itself. Illustrating this, Thompson (Thompson 2007) notes that:

> A cell stands out of a molecular soup by creating the boundaries that set it apart from that which it is not. Metabolic processes within the cell determine these boundaries. In this way the cell emerges as a figure out of a chemical background. Should this process of self-production be interrupted, the cellular components…gradually diffuse back into a molecular soup. (Varela et al. 1974, p. 99)

It would be harder to write a more elegant and compelling description of the importance, for living forms that resist the second law, of a Markov blanket organization than this. Any autopoietic system will, by definition "embody a circular process of self-generation [that] continually re-creates the difference between itself and everything else" (Thompson 2007, p. 99). Compare Friston:

> For example, the surface of a cell may constitute a Markov blanket separating intracellular and extracellular states. On the other hand, a candle flame cannot possess a Markov blanket, because any pattern of molecular interactions is destroyed almost instantaneously by the flux of gas molecules from its surface. (Friston 2013, sect. 2)

The Markov blanket considerations offer a formal and statistical (but, importantly, not intrinsically causal[11]) window onto the space of autopoietic organizational forms. The enactive 'E' is thus potentially in perfect harmony with the FEP.

Perhaps there is something in the embodied 'E' that will cause a problem? Here, it helps to consider a concrete example. In a series of experiments, Havas and colleagues (Havas et al. 2010) used Botox to induce facial rigidity during an 'emotion language processing' task. Subjects were shown sentences such as "your closest friend has just been hospitalized" and asked to push a button when they had finished reading the sentence. The study found that Botox-induced rigidity hindered the speed of processing of emotion language, concluding that "involuntary facial expressions may play a causal role in the processing of emotional language". In a follow-up study, Neal and Chartrand (Neal and Chartrand 2011) showed the reverse effect, demonstrating that improving facial feedback[12] enhanced the speed of processing. Botox-induced facial rigidity thus slowed reading, while gel-induced enhancement speeded it up. At a minimum, such results suggest that the processing of emotion language involves causal-functional loops that run through our own involuntary facial expressions. Loops running through the actual production of these facial expressions (not merely the issuing of neural commands that would normally result in those expressions) thus look to be both functional and integral to the normal processing of such stimuli.

This is a very typical piece of research in 'embodied cognition' (for many more such examples, see Clark 2008). It is also highly compatible with the predictive processing (PP) story itself. It seems very plausible that the role (in emotion processing) of the facial movements is to influence the flow of top-down prediction that attempts to 'explain away' the facial sensory signal. Our own facial expressions (like our own visceral states – see Seth 2013, Pezzulo 2014) simply constitute further sensory evidence

11 The Markov property is defined by statistical, rather than causal, relations between activity in the various inner and outer nodes. It is this fact, ultimately, that opens up a space of interesting and important questions concerning the active creation of new Markov blanket organizations as part of a life-cycle or as the result of lifetime learning and contingent external influences. We turn to these matters in sections 5-7 below.

12 Feedback signals are enhanced when muscle contractions meet resistance, so they used a gel to amplify that effect.

that helps drive the Bayesian machine more rapidly towards conclusions – here, conclusions concerning the meanings of the words on the screen.

Someone might still ask, of course, if this functionally important loop though facial expression is part of the cognitive process itself, or is merely input to that process? Otherwise put, is the facial expression partly *constitutive* of the cognitive process or is its role merely causal – just input to the true cognitive process? Here, the appeal to Markov blankets seems of little help. For we could just as easily ask 'does the actual facial expression lie within or beyond the Markov blanket that marks the boundaries of the mind and the mental?' The point to notice is that translating the question into these terms goes no way at all towards providing an answer! This is because (as we shall see in more detail below) a single adaptive system will typically comprise multiple blanketed organizations – blankets of blankets of blankets. Which ones (if any) should count as tracking, at a given moment, the machinery of mind is precisely the question at issue.

Such debates have (for better or for worse) been central to the philosophical and scientific debate concerning the 'extended mind'[13] – see, for example, a similar debate (Clark 2007) concerning the role of actual physical gestures in the processes of thinking and reasoning. Notice that the 'causal versus constitutive' question makes sense even if all parties are agreed that the inner neural regime is treating the facial states as additional evidence in a Bayesian inference. The question stands – and all the old arguments, pro and con apply - even if the neural contributions are all understood through the lens of Markov blankets, predictions and prediction error minimization.

4 Extended Predictive Minds

Hohwy (Hohwy 2016) does not agree. There, he offers three reasons to reject the 'extended mind' claim concerning processes and operations supported by familiar bio-external hardware such as notebooks and smart-phones. The first is that:

> it is far from clear that notebooks and smartphones actually play any part of the functional role set out by PEM. (Hohwy 2016, p. 270)

But what exactly is required for an item to play *part of* that functional role? It seems clear enough that a reliable non-biological systemic element could play a role in enabling a system to minimize free-energy and resist the second law – examples might include cochlear implants and spectacles. The prediction-error minimization (PEM) story is, however, more specific and refers to the specific message-passing scheme depicted in work on hierarchical predictive coding. Deploying that rather specific kind of process-schema is not required, however (as we saw earlier) in order to count as part of an integrated system that resists the second law. A simple free-energy or surprisal minimizing system need not necessarily engage in anything that resembles the top-down use of stored information to predict the current or evolving sensory flux. By extension, not every *proper part* of an integrated free-energy minimizing system (e.g. a cognitive agent) that *does* implement such an online prediction error minimizing process need itself be directly involved in that process. It is plausible, for example, that gross morphology (e.g. the shape of my hand) serves to minimize free energy in my embodied exchanges with the world, while not implementing any part of such an online process. By the same token, a system that minimizes free energy using online prediction error minimizing techniques (a 'PEM system') could be part of a larger free energy minimizing whole that includes multiple sub-systems that do not work that way.

Moreover, even if we decided - for whatever reason - that any genuinely cognitive process must be in some way entangled with prediction error minimization (a very strong claim for which no justification has been provided) that would still fall short of ruling out the notebook. It would fall short

13 See e.g. (Adams and Aizawa 2001, Clark 2008), and many of the exchanges in (Menary 2010).

because the timescale of entanglement would remain to be determined. Thus recall that Clark and Chalmers (Clark and Chalmers 1998), in their flagship treatment of these issues, argued only for the extension of long-term standing (so-called 'dispositional') beliefs – for example, your standing belief that Missouri is one of the United States of America. Most likely, you do not go around rehearsing this sentence even though you count, at all times, as believing it. Our claim was that for this kind of belief, long-term bio-external storage could be on a par with long-term bio-internal storage. The entries in Otto's infamous notebook play (we argued) this very role, under-writing patterns in potential behaviour (e.g. how Otto would answer certain questions if asked) apt for one who holds that belief. But for the notebook entry to actually drive a piece of behaviour in the here-and-now, that information needs to be activated or accessed, potentially as part of a process that involves prediction error minimization along the way. That means that the making use of the bio-external encoding may call upon whatever processes we deem (for whatever reason) essential to here-and-now cognizing.

To illustrate this, let's adapt an example treated at greater length in (Clark 2005). Imagine that one option for long-term bio-storage, in some alien life-form, is to create a stable 'bit-mapped image' – a kind of 'pixel-level' retinal screenshot, tagged with some information to enable later retrieval. Let's further suppose that this bit-mapped image is stored in some relatively inert fashion, e.g. using an alien-biological equivalent of flash memory. When these alien creatures want to recall that event, they may choose (if they wish – we may assume they command a fluid reconstructive bio-memory also) to retrieve the full bit-mapped image. When they do so, the information in the stored image still needs to be interpreted, and hence is now dealt with in much the same way as incoming sensory information.

The bit-mapped 'memory' is thus every bit as stable and 'inert' as the notebook used by Otto in Clark's and Chalmers' (Clark and Chalmers 1998) infamous thought-experiment. But in the aliens, the process of encoding and accessing the stored trace is fully biologically supported. Notice that this is not to deny that memory in us humans is reconstructive and dynamic. The point here is just that it was not conceptually necessary for that to be the case. Had cognitive science found that a few elements of the human bio-memory system were not reconstructive, we would not have concluded that those elements failed to form part of human cognizing. The imaginary aliens are meant to help reveal this fact. The extended mind theorist argues that, in the alien case, we would have no hesitation in counting the bit-mapped encodings as part and parcel of the alien life form's cognition. But if so, then the fact that something very similar happens in other beings, such as ourselves, using a combination of biological and bio-external (e.g. notebook-based) ought not – she argues - be treated in any different way. For the functional upshot is relevantly similar. So it can only be skin-and skull prejudice that stops us extending the same courtesy to many of our own bio-external resources.

Might the relative inertness of a resource itself be a problem for the claim that it forms part of a cognitive economy that enables a creature temporarily to resist the second law? I don't think so. For a relative inert systemic element may nonetheless help a larger system resist the second law. As a clear example, consider birds that swallow stones and grit to help them digest their food. These so-called 'gizzard stones' (Solomon et al. 2002, p. 664) are relatively inert but function to help grind the food, rendering it digestable and thus helping the birds stay alive. Likewise for bit-mapped images and trusty notebooks – they may be relatively inert when not in use yet function, within the larger system, so as to help enable apt adaptive response.

Such, in a nutshell, is the core argument for the 'extended mind'. I will not here rehearse the layers upon layers of further argument that might ensue (again, see Clark 2005, for the full story). Instead, let's also imagine that the alien life-form's brain is, in all other respects, a predictive processing device. Moreover, let's assume that at the time of encoding and at the time of use, the use of the bit-mapping faculty is selected and orchestrated by the core PP principles of precision-weighted prediction error minimization. Thus the choice of when to use the bit-mapping faculty is itself responsive to systemic best-guessing about what means of storage (and what kind of current retrieval) will reduce the greatest amount of sensory uncertainty in the long term. In other words, the prediction-free strategy is deli-

cately embedded in a prediction-rich economy. It would be wrong (I submit) to argue that the bit-map storage faculty is not part of the overall cognitive economy of that creature. The moral is that not every *part* of the full cognitive economy needs itself to display the full PEM profile. And once this is allowed (as it surely must be) for the imaginary, bio-internal case of the alien life-form, I see no principled reason to disallow the bio-external analogues available to creatures such as ourselves. Thus we should reject Hohwy's assertion that:

> The challenge is to specify the role of notebooks or smartphones, or any other thing, such that it clearly plays an appropriate prediction error minimization role. (Hohwy 2016, p. 270)

We should reject this because the very most that could reasonably be required is that such a resource be appropriately embedded within some larger prediction error minimizing economy. Alternatively, it might be suggested that the 'embedding' scenario I described, since it appeals to the long-term error-minimizing effects of the use of the notebook or bit-mapping, shows how very easy it is to achieve the right degree of integration to count as 'playing part of the PEM functional role'. Either way, the prima facie case against the notebook is defeated.

Hohwy's second worry is that cases such as the notebook raise questions concerning the evidentiary boundary for the self-evidencing model. The thought is that:

> This boundary should make it clear that prediction error is minimized for a system including the external object to which cognition is extended, and with respect to hidden causes outside this extended boundary. (Hohwy 2016, p. 270)

In the case of the notebook, let's assume it includes an entry with an address that matters for doing my job, hence getting paid, hence tending to eat and persist. This, in any civilized society, ought to be a caricature - but you get the idea. The address picks out a place beyond the boundary of the bio-Andy-plus-notebook system. So the relevant 'hidden cause' lies beyond the boundary, just as Hohwy (op cit, p. 270) insists. Is the notebook 'part of the model providing evidence for itself'? I suggest that it is – or rather, I suggest that this is one place we might draw a defensible (though non-unique – more on that soon) boundary. Again, we can motivate this by considering a different, more purely biological, case. Consider the spider and the web. The web, once created, is an organism-external structure that forms a proper part of the free-energy minimizing system. This is the system responsible for adaptive fitness, and it is in exactly this sense that the web is depicted as part of an 'extended phenotype' (see Dawkins 1982) or even as an 'external physiological organ' (Turner 2009).

In just the same way, it seems to me, we should count the notebook as an organism-external structure that is part of a larger free energy minimizing system that, by exchanging entropy with the even larger environment, provides evidence for itself. Hohwy (op cit) also insists that evidentiary boundaries be defined so as to demarcate systems that minimize average prediction error 'in the long run'. This raises complex issues concerning change within the life-cycle, and we return to these in section 5.

Is there, within this larger free energy minimizing whole, a smaller whole relative to which the notebook is external? Surely there is. Similarly, my biological body is heavily reliant on the activity of various colonies of microbes. Indeed, up to 90% of the cells comprising my body are said to be 'microbial symbionts'. Each individual microbe is a free energy minimizing whole, complete with its own Markov blanket, and I would easily survive the loss of one or many of these cells. The lesson (which we return to below) is that complex creatures are composed of layer upon layer of Markov blankets - which layers we choose to emphasize can only be fixed by our local explanatory interests and purposes.

How might we apply this lesson to the matter of locating the mind? Here, the key observation (going back at least to Haugeland 1998) is that we should not simply assume that metabolic boundaries and cognitive boundaries always and everywhere coincide. Just as the boundaries of the liver cell are

not those of the liver, and the boundaries of the digestive system are not those of the immune system, so the bounds of metabolism and bio-sensing need not be those most relevant for sensing (more generally understood) or thinking. Which boundaries we choose (hence which capacities we identify as those of the agent herself) will depend on our wider explanatory purposes.

Hohwy's third and final worry is that:

> There is something unattractive about both acknowledging that an external object (such as a notebook) is represented in the mind's model of the world and insisting at the same time that that object is itself part of some of the mind's mental states. It is unattractive because it means the object is both beyond one evidentiary boundary and within a further evidentiary boundary. (Hohwy 2016, p. 270)

To see why this is not a problem, compare the case of an agent who uses a sensory aid such as spectacles or a blind-person's cane. When in use, the spectacles or cane mark clear (though non-unique) evidentiary limits. My brain minimizes prediction error relative to a flow of evidence that would not be available without the spectacles/cane. But suppose I lose the spectacles or the cane. Now they lie outside the evidentiary limits, and I must minimize prediction relative to an impoverished flow of evidence.

Hohwy goes on to argue that:

> This is not an inconceivable state of affairs but it […] requires that we posit two overlapping yet intimately linked EE [Explanatory-Evidentiary] -circles with different evidentiary boundaries. If there are two EE-circles, then the input to each of the circles will be evidence for the existence of two distinct yet overlapping agents. This may be considered an argument for extended cognition but at the unattractive cost of proliferating the number of agents centered on a particular organism. (Hohwy 2016, p. 270)

But what this really shows, I suggest, is something subtly but importantly different. What it really shows is that the notion of a single persisting agent should not be identified with a stationary set of Explanatory-Evidentiary (EE) boundaries at all. Instead, we should adopt a *process ontology* relative to which a persisting agent can be identified with a rolling process that builds and re-builds its own evidentiary boundaries on the fly. Such agents 'knit their own Markov blankets' in ways that can change over time, without the agent thereby ceasing to exist. It is to this – admittedly challenging – project that I next turn.

5 The Multiplicity and Malleability of Markov Blankets

Consider the metamorphic insects[14]. These insects undergo a dramatic change of form as part of their standard lifecycle. Where metamorphosis occurs, the young life-stages do not look or behave in anything like the same way as the adult or mature life-stages. They may eat radically different foods, and locomote in totally different ways. A familiar example is the transformation of a caterpillar into a butterfly. Caterpillars crawl, eat leaves and do not mate. Butterflies fly, seek out nectar, and mate with other butterflies. In the amphibian world, the small swimming tadpole becomes a large jumping frog. Such examples can seem exotic. But in fact, metamorphic insects alone account for at least 40% of the world's total animal populations. As an evolutionary strategy, metamorphosis works - it is not a rare or exceptional solution to the problem of adaptive success. Moreover, even non-metamorphic animals such as cats, dogs, and humans, exhibit striking differences between mature and immature forms, with the very young looking and behaving quite differently to the older forms. While at the extreme end

14 The brief sketch that follows is based on (Jabr 2012), and the Encyclopædia Britannica entry at https://www.britannica.com/science/metamorphosis.

of the spectrum lie the so-called hypermetamorphic insects that exhibit a whole series of dramatic changes across the lifespan.

Metamorphosis poses an interesting puzzle[15] for Hohwy's picture of the links between persisting agents, free energy, and the 'self-evidencing model'. The phase-transitions that occur as part of the developmental trajectories characteristic of metamorphic animals are dramatic enough to count as failures of the earlier stage life-form to continue to harvest evidence for its own existence. Yet it seems wrong to think that the transitions constitute breakdowns or failures in the organism's war against the second law of thermodynamics. The caterpillar does not lose the war when it transforms into a butterfly. On the contrary, the act of transformation is itself an essential part of the on-going project of exchanging entropy with the environment so as to persist in the face of the second law. For example, it is conjectured that metamorphosis has adaptive value because it allows younger and older forms to share the same territory without consuming the same resources or being exposed to the same predators.

A natural response to this prima facie puzzle is to point to the genetically determined nature of the phase-transitions (hence the succession of differently blanketed organizational forms) themselves. The genes that control metamorphosis are reasonably well-understood, and can be selectively blocked so that (for example) the caterpillar never turns into a pupa and hence never undergoes the caterpillar-to-butterfly phase transition – see (Bayer et al. 2003). It is reasonable, then, to think of the overall life-cycle as an evolved, self-evidencing, free-energy minimizing strategy. The life-cycle is self-evidencing insofar as the very existence of the linked stages (caterpillar, pupa, butterfly) provides evidence for the 'model' that is the metamorphic agent, where that agent is not identified with a specific morphology (which would correspond merely to one stage of the lifecycle) but with the temporally extended whole[16].

This, I submit, is the correct way to think about metamorphic beings in the free energy framework. It is also a revealing platform from which to re-consider Hohwy's worries concerning embodied cognition and the extended mind. Thus consider the wide range of bodily and sensory augmentations that already characterize many human lifespans. These include the use of spectacles to improve vision, and the wearing of clothes to help tolerate heat and cold. For some of us, they include the use of pacemakers, cochlear implants, and prosthetic limbs. In the near-future they may include the use of techniques such as refractive lens exchange to deliver not just restored but augmented vision – for example, by providing infra-red (IR) sensitivity. If my bio-typical lens is replaced with an artificial one that augments my visual repertoire, and its outputs properly integrated in downstream neural processing, it would seem strangely unmotivated to insist that 'my' true evidentiary boundaries remain those of the bare (IR-insensitive) biological system. Closer to home, we can easily imagine IR sensory evidence being made available via a mediating wearable technology such as Google Glass.

To be sure, we could identify a Markov blanket wherever the new technologies interface with the old bio-systems. In the case of ordinary spectacles or the IR-enabling Google-glass, this would line up with a known long-term boundary of interest. In the case of the lens replacement, perhaps only surgeons would consider the interface (between the new lens and the post-retinal wiring) a boundary of interest. But the lesson is that complex living beings are composed of layer upon layer of Markov blankets, reaching at least all the way down to cellular organelles and macro-molecules like DNA (deoxyribonucleic acid). Different explanatory purposes drive us to highlight some of these blankets (of blankets) at the expense of others. But no blanket or set of blankets is privileged in and of itself. Nor does the temporality of any specific Markov blanket organization seem intrinsically privileged. In the

15 I was led to consider the intriguing case of the metamorphic insects by (Friston and Stephan 2007, p. 435), who note the prima facie puzzle posed by their dramatic phase-transitions, that nonetheless occur within well-defined within developmental trajectories.

16 This again raises interesting question concerning ergodicity – here, the tendency of the blanketed systems to visit and re-visit the same sets of states over time. For we now confront a time-series in which there are profound changes in the states that are re-visited during different 'chunks' of the life-cycle. This is reminiscent, in the context of work on the 'extended mind', of the arrival of a new technology whose operations become so deeply integrated as to be called upon again and again in (some part of the) temporally-subsequent lifespan.

case of the metamorphic insects there is dramatic change across time, so that no stable Markov blanket organization characterizes the biological agent across the life-span. Instead, what seems to have been selected is an ongoing process that delivers different organizational forms (different blankets of blankets) at different times. As one of these forms gives way to another, there is some kind of preservation of systemic integrity – the biological being thus resists dissipation and death. But it would be a mistake to look for a single form or blanket of blankets that persists throughout the lifespan. Instead, we should see the shifting forms (and the shifting mosaics of Markov blankets) as themselves the means – the process – by which long-term surprisal is minimized.

The extended mind claim is that, considered as cognitive agents, human beings are 'mentally metamorphic' – as we move through life, the sets of tools, strategies, and devices (neural, bodily, and bio-external) that constitute us as the mindful beings we are undergoes dramatic alteration. Very young human life-forms do not count, as part of their cognitive apparatus, the use of pens, papers, and notebooks. But that can change over time, as the capacities they make available become more and more deeply integrated with bio-native cognitive operations. As we move through life, different bits of the encountered and humanly designed world become repeatedly and deeply incorporated into our individual cognitive routines, persisting or decaying according to need, use and the vagaries of our enabling socio-technological cocoon. Importantly, the predictive processing architecture itself provides a powerful mechanism enabling the flexible, repeated integration of capacities and operations made available by the use of reliable bio-external resources (see Clark in press, sect. 8, and Clark 2016, ch. 8). I won't rehearse those considerations here, since it is in any case evident – merely from observing our human capacity to become fluent users of a lifetime succession of new tools and technologies – that such fluidity is a crucial part of our heritage[17].

6 Blanket-Weaving Blankets of Blankets

Can we reconcile Hohwy's vision of agents as self-evidencing, free-energy-minimizing systems with the EEE emphasis on the multiplicity and malleability of the Markov blanket organizations themselves? The best way to do so, it seems to me, would be to embrace a process (rather than a state-based) ontology, perhaps of the form described and defended in (Dupré 2012, Dupré 2014)[18]. State-based ontologies, Dupré notes, view entities as things and often ask what changes those 'things' can and cannot tolerate. Process–based ontologies for biology, by contrast, take change as given and focus on the way some processes of change (in the case of living systems) constitute a powerful means of temporarily resisting the second law.

Dupré notes that the life-cycles of many organisms – including humans – includes multiple very different forms, and asks (Dupré 2014, p. 33) "why assume there must be anything common to every stage beyond their participation in a continuous process?" The metamorphic insects (section 5 above) raise this issue in an especially dramatic way. But it applies to every animal that undergoes developmental change. From such a perspective "a cat is a pathway from zygote to kitten to mature animal to death" (Dupré 2014, p. 33). As this pathway unfolds, nothing need be common to all the temporal parts except for their participation in the process. Nonetheless there seems no reason to think that a temporally extended process cannot itself be a free-energy minimizing system that might even be identified with some self-evidencing agent of interest. For it is the process-based succession of processes that must, ultimately, account for the temporary resistance of living organisms to the second law.

17 But notice also that the fluidity in question is itself plausibly at least partially determined by cultural innovations such as reading and writing, since these enable the easy transmission of the staged training regimes required to master new tools and techniques. See (Heyes 2012).

18 See also (Dupré and O'Malley 2009), and for a general introduction, (Seibt 2016).

At this point, I suspect that Hohwy will wish to emphasize the difference between an organism with a genetically (or at any rate, epigenetically[19]) determined pathway from one form to the next, and the case of a human agent who comes across a new tool or technology such as a smartphone and slowly incorporates it into their cognitive routines. Should we say that the caterpillar-butterfly process counts as a unity that resists entropy and maximizes evidence for itself over the long run, whereas the human-smartphone unity is too fleeting and too arbitrary to count? Such a response is perhaps suggested by the comment that:

> Whereas prediction error can be minimized transiently by systems with all sorts of objects included (*e.g.* shooting the tiger with a gun), on average and over the long run, it is most likely that the model providing evidence for itself is just the traditional, un-extended biological organism. (Hohwy 2016, p. 270)

But notice, first, that we do not have to settle for a single self-evidencing model. Within the human agent, there will be many self-evidencing 'models' including single cells and, as Dupré (Dupré 2014) nicely notes, every member of the successive colonies of bacteria that inhabit the human gut and help us digest our food. Should we nonetheless assume that where there is a persisting human person there is a unique, unchanging self-evidencing system bounded by a unique, unchanging Markov blanket? There is no reason to think so. The sets of unfolding processes that constitute the infant human are not the same as the sets of processes that constitute the adult form, and the infant brain itself is very different from the adult. So even once we fix on some Markov blanket organization of interest, that organization will itself be realized as a process undergoing constant change. Any Markov blanket bounded organizational form can be changed, re-deployed, or re-configured through interaction with the inner and outer environment.

Most importantly of all, we need to recognize learning itself as a key transformative process. It is a process (or set of processes) that forms part of both our biological and cultural heritage, and that – I would argue – delivers a succession of differently constituted mental and sensory organizations as part of its normal operation. That same process (or set of processes) can, we saw, re-configure effective systemic boundaries in many ways. The person equipped with the IR-lens will soon learn to use that information fluidly for the control of perceptuo-motor routines, as will the person who uses an add-on device such as Google Glass. In each case, the effect of learning is to generate a 'motor-informational weave' (Clark in press) that alters the effective bounds of sensing and action. When the well-woven equipment (the notebook, a software package) plays a role in the storage and transformation of information, we may speak of 'extended cognizing' rather than merely 'extended sensing' – but in each case, the key features are the same. The fact that we can still, if we wish, define a set of boundaries (a Markov blanket) that falls within these larger wholes is unsurprising. For even within the basic biological whole, there are many other boundaries we could choose. And whatever set of boundaries we happen to focus upon, they will themselves be subject to change in virtue of their realization by temporally-evolving processes rather than stable states.

Creatures like us, I conclude, are Nature's experts at knitting their own Markov blankets. Courtesy of biology, culture and learning we are 'natural-born Cyborgs' (Clark 2003) – self-organizing processes that constantly re-invent themselves, repeatedly re-defining their own cognitive, bodily, and sensory forms.

19 For example, (Dupré 2014, p. 34) notes that "the growing meristem of a plant is typically an opportunistic growth process capable of producing a variety of structures – leaves, flowers, roots – in response to the environment it encounters."

7 Demonic Couplings

I have tried to show that the Markov blanket organizations most relevant to the study of mind and life are multiple and malleable. But multiple, malleable organizations are organizations nonetheless. That means that for a given explanatory interest, at a given moment in time (a window within the evolving process), there will indeed be a Markov blanket on one side of which lie the interacting worldly causes that are (in Hohwy's very broad sense) 'modelled' by the temporarily persisting system. Suppose in addition that the system in question is one that constructs some kind of conscious experience of its world. This awareness, Hohwy argues, must involve a form of inference conditioned upon the sensory evidence, concerning the world beyond the blanket. This, opens the door to scepticism, since as long as the energetic exchanges across the sensory boundaries are held constant the agent cannot know whether the world beyond is one way or another. Given her priors and the sensory evidence, she will construct the same world regardless. This is the 'evil demon' (or Matrix) scenario pursued in (Hohwy 2016) and revealingly refined in (Hohwy 2017). Is there something about this sceptical scenario that puts pressure on EEE cognition?

In (Clark in press) I argue that there is not. I claim that what EEE cognition rejects is not scepticism so much as the 'richly re-constructive' vision of mind. A richly reconstructive vision depicts the mind as an inner arena populated by representational forms sufficiently rich and stable to enable us to do all the 'real' cognitive work in an effectively 'offline' manner, relegating sensing and action to simply inputting a problem specification and outputting a solution - one that action merely implements. The alternative is to recognize that action, as well as gross bodily form and reliable environmental features, can form part of the solution itself – as we see in many well-studied examples (see Clark 2008) ranging from the use of fingers in counting, to gestures while speaking, to yellow sticky-notes posted on the wall. The core EEE insight, I believe, is that circular (perceptuo-motor) causality in a richly structured and often repeatedly self-structured environment is itself the core adaptive strategy, and that the neural contribution is best seen as just one part of this wider body-and-world-involving mosaic.

As far as I can tell, Hohwy (Hohwy 2017) accepts that broad vision. But he remains concerned to stress the apparent insulation of conscious experience behind a veil of sensing and acting. If all that this means is that, in principle, as long as my priors and the flow of energies across my sensory surfaces remain fixed, I cannot know whether or not I am a Matrix-world 'brain-in-a-highly-intelligent-vat', I am happy to agree. Notice that being thus en-vatted would in no way alter the fact that the problem-solving strategies I deploy make great use of extra-neural opportunities. The vat-being, thus construed, can robustly rely on the use of her fingers for counting, and she can post yellow sticky notes aplenty on the surfaces of her sensed office. Mildly cognitively impaired Otto, in vat-world, likewise makes use of a trusty notebook that allows him to behave in many of the ways characteristic of having a multitude of standing beliefs that are not enshrined in his neural apparatus in the bio-typical manner. As we move deeper and deeper into the 21st century, vat-world too may be increasingly home to enhanced and augmented agents, who have extended their bio-typical sensory boundaries using new devices such as IR-sensitive replacement lenses. All this is true, it seems to me, regardless of the fact that, in vat-world, these cognitive extensions and boundary-changing alterations are implemented in vat-controlled software rather than whatever hardware ultimately - or so we presume - comprises the physical universe we inhabit

In a revealing twist on his original (2016) demon-description, Hohwy (Hohwy 2017) further shows that to the extent that all this is true, the prediction-error minimizing agent is in fact manipulating the demon (the intelligent vat) by giving action commands whose fulfilment requires the demon to compliantly alter the apparent flows of energy across her (apparent) sensory surfaces.

What shall we make of this? In Hohwy's well-chosen words:

> [N]o 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. (Hohwy 2017, sect. 6)

This is a lovely observation. The inevitable entrainment of the demon/vat by the agent shows (I would argue) that the shape and structure of the experienced world – the world as presented in experience – is nothing over and above the shape and structure of an open-ended set of possible encounters, shaped by human needs and the affordances of body and world. Whether all that is ultimately implemented in the standard or some non-standard (e.g. envatted) way is unimportant. This is just another way of showing that the demon thought experiment puts no pressure on the core claims of EEE cognition, at least as I would construct them. Within vat-world, all the interesting and unexpected patterns of reliance upon bodily and extra-neural sources of stability and structure celebrated by EEE cognition remain in force. More radically still, Chalmers (Chalmers 2005) and Clark (Clark 2005) argue that that the very reality of arms, legs, notebooks, and sticky-notes is unaffected by these variant implementations, since 'realness' should never have been predicated upon the correctness of some specific deep physics in the first place. If we go this route, the premise of the thought experiment is itself undermined – 'all' the demon has done is altered some implementation details while preserving our perfectly real world. But I mention this only in passing since (as we just saw) it is not necessary to endorse this more radical view in order to move beyond the image of neurocentric seclusion, even for the envatted brain.

Finally – more for fun than as an attempt at further argument – suppose we step outside the realm of constructed experience and ask how the vat-agent exchanges entropy with the environment so as to persist (as some kind of evolving process) in the face of the second law? The answer, from the agent's perspective, implicates multiple perception-action loops that exploit extra-neural resources of varying kinds. But those extra-neural resources are now held steady only – I presume – by the expenditure of a great deal of demonic energy. If we deny this, then the truth of the demon scenario would simultaneously imply the falsity of the free-energy principle itself. Assuming this is not the intention, we must imagine that the demon is busily exchanging entropy with some larger demon-world, so as to persist long enough to be suitably entrained by the vat-agent. Such a demon would herself be well-placed to conform to the core principles of EEE cognition, increasing the efficiency of her own demonic stratagems by making maximal use of sources of order and structure beyond the bounds of the demon-brain itself.

8 Conclusions: From States to Processes

I have tried to show that EEE approaches, even those that embrace 'extended minds', do not thereby seek to obliterate Markov blankets from their depictions of mind and agency. To do so would, as Hohwy and Friston correctly stress, disable the controlled exchange of entropy and expose the agent or cognitive system to immediate fatal dissipation – bad news even for an extended mind. In (Clark in press) I argue that EEE approaches are best seen as rejecting 'richly reconstructive' visions of mind. But that is really only half the story. The other half of the story (pursued today) concerns the way

embodied minds and embodied agents repeatedly re-configure their own boundaries in ways that promote adaptive success. Such repeated reconfiguration does not always result in the destruction of the older blanketed organizations, which may simply become further nested. But it may sometimes – as in the case of the metamorphic insects – do so. Either way, it results in the creation of new systems of explanatory interest. Such systems remain locally ergodic. They visit and re-visit states that their predecessors did not, while remaining stages of a single life-cycle.

Dramatic change in the metamorphic insects is genetically controlled, and thus counts as 'expected' in some very broad evolutionary sense. But many biological agents are, in the same broad sense, 'expected' to engage in lifetime learning, altering their effective environments and (in our case) creating tools and technologies that may shift the sensory boundaries themselves. That same process can usher into being new individualistic organizational forms, such as the bio-being+smartphone, or new collective forms, such as a nation-state. Learning and neural plasticity thus open the doors for new or successor blanketed systems consequent upon cultural and technological (including bio-engineered) innovations.

The perspective I am recommending may seem challenging insofar as it invites us to contemplate agents whose sensory boundaries are not fixed and whose cognitive architectures may extend, in temporally varying ways, beyond the biological brain. But perhaps this should not surprise us unduly. For there already exists a potent inner equivalent in the 'predictive processing' account of attention and variable precision-weighting itself. Attentional mechanisms, that story suggests, alter patterns of inner (neural) effective connectivity so as to enforce information flows that are highly specialized for the task at hand (Clark 2016, Clark in press). Attention, if this is correct, itself imposes a kind of transient organizational form, with its own distinctive Markov blanket organization (marked by temporary conditional statistical independencies), upon the brain. Attentional mechanisms may thus be seen as driving the formation and dissolution of a short-lived Markov partitioning within the neural economy itself[20], temporarily insulating some aspects of on-board processing from others according to the changing demands of task and context. These transient neuronal ensembles then recruit (and may also be recruited by) shifting coalitions of bodily and worldly elements, resulting in the repeated construction of temporary task-specific devices that span brain, body, and world. In this way, the ebb and flow of neural influence is matched by an ebb and flow of bodily and worldly influence. It is the progressive generation and maintenance of these nested transient partitionings, swept up in the circular causal dynamics that bind perception and action, that enables living beings to persist and minimize free energy across their lifespan.

Hohwy's insightful probing has thus revealed something deeply important. It has forced us to recognize that the picture of biological agents as free-energy-minimizing systems requires something closer to a process-based (rather than a static or state-based) ontology. If the free-energy minimizing system is really a free-energy minimizing process, much that is otherwise puzzling falls into place. For processes are by their very nature on-going, and can repeatedly generate new forms, boundaries, and constituent structures as they continue to exchange entropy with whatever counts (at a given stage) as the wider world.

20 A more familiar case may be the wake-sleep cycle itself, which regularly creates a new set of partitions (a new transient Markov blanket) between sensory systems and the world. This possibility is noted by Karl Friston in comments reported on the Frith blog at: http://frithmind.org/social-minds/2014/05/12/under-the-markov-blanket/

References

Adams, F. & Aizawa, K. (2001). The bounds of cognition. *Philosophical Psychology, 14* (1), 43–64.

Bayer, C., Zhou, X., Zhou, B., Riddiford, L. M. & von Kalm, L. (2003). Evolution of the Drosophila broad locus: The Manduca sexta broad Z4 isoform has biological activity in Drosophila. *Development Genes and Evolution, 213* (10), 471–476.

Chalmers, D. (2005). The Matrix as metaphysics. In C. Grau (Ed.) *Philosophers explore The Matrix*. New York: Oxford University Press.

Chiel, H. J. & Beer, R. D. (1997). The brain has a body: Adaptive behavior emerges from interactions of nervous system, body and environment. *Trends in Neurosciences, 20* (12), 553–557.

Clark, A. (1997). *Being there: Putting brain, body, and world together again.* Cambridge, MA: MIT Press.

Clark, A. (2003). *Natural born cyborgs: Minds, technologies, and the future of human intelligence.* New York: Oxford University Press.

Clark, A. (2005). Intrinsic content, active memory and the extended mind. *Analysis, 65* (285), 1–11. https://dx.doi.org/10.1111/j.1467-8284.2005.00514.x.

——— (2007). Curing cognitive hiccups: A defense of the extended mind. *The Journal of Philosophy, 104* (4), 163–192.

——— (2008). *Supersizing the mind: Embodiment, action, and cognitive extension.* New York: Oxford University Press.

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

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

Clark, A. & Chalmers, D. (1998). The extended mind. *Analysis, 58* (1), 7–19.

Dawkins, R. (1982). *The extended phenotype.* Oxford: Oxford University Press.

Dupré, J. (2012). *Processes of life: Essays in the philosophy of biology.* New York: Oxford University Press.

——— (2014). A process ontology for biology. *Physiology News, 100*, 33–34.

Dupré, J. & O'Malley, M. A. (2009). Varieties of living things: Life at the intersection of lineage and metabolism. *Philosophy & Theory in Biology, 1*, e003.

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.

——— (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.

Haugeland, J. (1998). Having thought: Essays in the metaphysics of mind. In J. Haugeland (Eds.) *Mind embodied and embedded* (pp. 207–240). Cambridge, MA: Harvard University Press.

Havas, D. A., Glenberg, A. M., Gutowski, K. A., Lucarelli, M. J. & Davidson, R. J. (2010). Cosmetic use of botulinum toxin-A affects processing of emotional language. *Psychological Science, 21* (7), 895–900.

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

Heyes, C. (2012). Grist and mills: On the cultural origins of cultural learning. *Philosophical Transactions of the Royal Society B, Biological Sciences, 367* (1599), 2181–2191. https://dx.doi.org/10.1098/rstb.2012.0120.

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

——— (2017). How to entrain your evil demon. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Jabr, F. (2012). How did insect metamorphosis evolve? *Scientific American Online*. https://www.scientificamerican.com/article/insect-metamorphosis-evolution/.

Lipton, P. (2001). What good is an explanation? In G. Hon & S. S. Rakover (Eds.) *Explanation: Theoretical approaches and applications* (pp. 43–59). Dordrecht: Springer Netherlands.

Menary, R. (Ed.) (2010). *The extended mind.* Cambridge, MA: MIT Press.

Mohan, V., Morasso, P., Sandini, G. & Kasderidis, S. (2013). Inference through embodied simulation in cognitive robots. *Cognitive Computation, 5* (3), 355–382.

Neal, D. T. & Chartrand, T. L. (2011). Embodied emotion perception: Amplifying and dampening facial feedback modulates emotion perception accuracy. *Social Psychological and Personality Science, 2* (6), 673–678.

Newen, A., de Bruin, L. & Gallagher, S. (in press). *Oxford handbook of 4E cognition.* New York: Oxford University Press.

Norris, J. R. (1998). *Markov chains.* Cambridge: Cambridge University Press.

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

Pezzulo, G. (2014). Why do you fear the bogeyman? An embodied predictive coding model of perceptual inference. *Cognitive, Affective, & Behavioral Neuroscience, 14* (3), 902–911.

Pfeifer, R. & Bongard, J. (2006). *How the body shapes the way we think: A new view of intelligence.* Cambridge, MA: MIT Press.

Rupert, R. D. (2009). *Cognitive systems and the extended mind.* New York: Oxford University Press.

Seibt, J. (2016). Process philosophy. In E. N. Zalta (Ed.) *The Stanford encyclopedia of philosophy* Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2016/entries/process-philosophy/.

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences, 17* (11), 565–573.

Solomon, E. P., Berg, L. R. & Martin, D. W. (2002). *Biology. Sixth edition.* Stamford, CT: Thompson Learning.

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

Tribus, M. (1961). *Thermostatics and thermodynamics: An introduction to energy, information and states of matter, with engineering applications.* Van Nostrand.

Turner, J. S. (2009). *The extended organism: The physiology of animal-built structures.* Cambridge, MA: Harvard University Press.

Varela, F. G., Maturana, H. R. & Uribe, R. (1974). Autopoiesis: The organization of living systems, its characterization and a model. *Biosystems, 5* (4), 187–196.

Wiener, N. (1961). *Cybernetics: Or control and communication in the animal and the machine.* Cambridge, MA: MIT Press.