Extended predictive minds: do Markov Blankets matter?

Abstract:

The extended mind thesis claims that a subject’s mind sometimes encompasses the environmental props the subject interacts with while solving cognitive tasks. Recently, the debate over the extended mind has been focused on Markov Blankets: the statistical boundaries separating biological systems from the environment. Here, I argue such a focus is mistaken, because Markov Blankets neither adjudicate, nor help us adjudicate, whether the extended mind thesis is true. To do so, I briefly introduce Markov Blankets and the free energy principle in section 2. I then turn from exposition to criticism. In section 3, I argue that using Markov Blankets to determine whether the mind extends either begs the question against the extended mind or provides us an answer based on circular reasoning. In section 4, I consider whether Markov Blankets help us perspicuously frame the debate over the extended mind, answering in the negative. This is because resorting to Markov Blankets to determine whether the mind extends yields extensionally inadequate conclusions which violate the parity principle. In section 5, I argue that resorting to Markov Blankets makes internalism about the mind vacuously true, preventing any substantial inquiry over the extended mind. A brief concluding paragraph follows.

Keywords: Extended Mind, Free-energy Principle, Markov Blankets, Active Inference

1 - Introduction

Vehicle externalism (also known as the extended mind thesis) claims that a subject’s “thinking machinery”[1] sometimes includes the environmental props the subject interacts with while solving cognitive tasks (Clark and Chalmers 1998; Hurley 2010). Importantly, vehicle externalism makes no claim concerning the nature of the thinking machinery; it is only a claim concerning its *physical constituents* (vehicles). Vehicle externalism is thus compatible with different accounts of mentality, including computationalism (Clark 2008), ecological psychology (Chemero 2009), enactivism (Di Paolo 2009), dynamicism (Palermos 2014) and more. As a consequence, how vehicle externalism should be articulated and whether or not it is true are intensely debated topics (Kiverstein 2018; Rowlands *et al*. 2020).

[1] Here, I use the phrase “vehicle externalism” to stay neutral on the distinction between extended *cognition*, extended *mind* and extended *consciousness*. This is because I’m interested in vehicle externalism *per se*, rather than any particular form it might assume - so, I needed a “catch all” term. Similarly, I use “thinking machinery” as a catch the system that is supposed to extend, be it a cognitive, conscious or mental system.

The recent popularity of "predictive" approaches to the mind, especially Friston's free-energy principle (FEP e.g. Friston 2010), generated a wave of "predictive" vehicle externalism (e.g. Clark 2017a) counterbalanced by equally consistent wave of internalism (e.g. Hohwy 2016). Their clash rapidly centered around *Markov Blankets* (MBs), focusing on questions like: "is there a privileged MB surrounding the thinking machinery?" (e.g. Ramstead *et al*. 2019); and: "if yes, does it enshroud *only* the brain?" (e.g. Hohwy 2016).

Here, I wish to take a step back from these questions, to observe the role MBs play in the debate over "predictive" vehicle externalism. To anticipate, I will argue that MBs neither adjudicate, nor *help* to adjudicate, that debate. My plan is as follows. In the next section, I introduce the FEP, focusing on MBs. In section 3, I argue that, on their own, MBs do not provide a solution to the debate over vehicle externalism. In section 4, I argue that MBs do not even provide a good way to *frame* that debate, showing that the framing offered by MBs has thus far forced us to adopt an extensionally inadequate criterion to identify the constituents of our thinking machinery. In section 5, I argue that resorting to MBs leads us to sidestep, in an important sense, the debate over vehicle externalism, as they make vehicle internalism *vacuously* true[2]. A brief concluding paragraph follows.

2 - Free-energy: a selective sample

The FEP is standardly presented as an account of biological self-organization, stating that the persistence of biological systems is guided by *free-energy minimization* (Friston 2011; 2012; 2013; Friston and Stephan 2007; Hesp *et al.* 2019).[3]

[2] A "disclosure statement": I endorse vehicle externalism. But my aim here is *not* to defend it. My only aim is to argue that the debate over vehicle externalism should leave MBs behind. So the problem I raise in section 5 is *not* that MBs make vehicle internalism true, but that they do so *vacuously*.

[3] Readers familiar with the FEP might object that the FEP is "scale free", as it provides an account of existence *in general* (e.g. Hipolito 2019; Friston 2019). It is hard, however, not to notice that the FEP is *typically* offered as a principled solution to the problem of biological self-organization (e.g. Friston 2013; Hohwy 2020; Corcoran *et al*. 2020; Seth 2020). At this juncture, then, it might be useful to state that the FEP is, theoretically speaking, a moving target, whose formal apparatus is still under construction (consider, for instance, the recent introduction of *dual information geometries*, see Parr *et al.* 2020) and whose epistemic status and ontological commitment are far from clear (see Colombo and Wright 2018; van Es 2020).

Consider the variables implicated in the prolonged existence of an organism (e.g. level of glucose in the blood, level of activation of photoreceptors, *etc*.). Their values collectively determine the organism's state, and their range determines the set of possible organismal states, which can be represented as a state-space having one dimension for every variable. Points in such a space represent particular states (e.g. glucose at level *x*, photoreceptors active at level *y*, *etc*.). In principle, an organism might occupy any point of that space - but typically won't. My bodily temperature, for instance, is typically around 36.6°; hence I tend to "occupy points" clustered around that value. The same holds true for other dimensions of that space. Therefore, organisms occupy only a small volume of that huge state-space; and they *must* do so if they want to continue living (Friston 2019: 175-178). If I start wandering towards the "54° bodily temperature" region of my state-space, I would be toasted (literally).

According to the FEP, that small volume in the state space formally captures an organism's *phenotype*.[4] To continue living, organisms must constantly occupy points in that volume; therefore they must constantly re-visit the states belonging to their phenotype. For this reason, the FEP assumes that organisms are *ergodic*; meaning that: "one can interpret the average amount of time a state is occupied as the probability of the system being in that state when observed at random" (Friston 2013: 2). An example might help clarify this concept. Suppose a fair dice is cast eternally. Being fair, the dice will occupy a state (i.e. displaying number *n*) ⅙ of the time. Moreover, since the dice is fair, were I to observe it randomly, I would observe it displaying *n* with a probability of *roughly* 0.17; that is, p(*n*) = ⅙ ≈ 0.17. Hence, the time the dice spends in a state (displaying *n*) approximates the probability that I would observe that state if I observed the dice at random.

[4] Technically speaking, the FEP conceives phenotypes as sets of attractors constraining the organism's path through its state-space (e.g. Friston 2013), which are determined by each organism's embodiment and evolutionary history (e.g. Friston *et al*. 2012). In the most recent renditions of the FEP, they are also referred to as the *non-equilibrium steady state density* (Friston *et al* 2020).

A probability function can thus be defined over the states in the state-space. States belonging to the phenotype will be highly probable, whereas non-phenotypic states will be highly unlikely. The distribution will thus be sharply peaked around phenotypic states (e.g. “bodily temperature 36.6°”) and very flat on all other states (e.g. “bodily temperature 54°”); technically speaking, that distribution has low *entropy*.[5] This also means that, *on average and in the long run*, organisms occupy states with low *surprisal* (given their phenotypes). Here, surprisal is an information theoretic quantity (basically, the negative logarithm of a probability) quantifying how “far off” any given state is from the phenotypic states; that is, the states the organism should occupy to prolong its existence. Thus, to survive, organism must keep the entropy over their states low, or, which is equivalent, avoid “surprisaling” states *on average and in the long run*. Notice that such a “long run” perspective is *essential* to the FEP. In fact, it is only because this “long run” perspective is adopted that the ergodicity assumption holds, thereby allowing biological self-organization to be described as *surprisal avoidance* overtime.

Importantly, organisms cannot quantify surprisal directly. According to the FEP, however, they can keep track of an *upper bound* on surprisal, which is (variational) free-energy (Buckley *et al*. 2017). Organisms can track free energy because it is a function of two probability densities organisms can track; namely a *generative density*, which specifies the joint probability of worldly and sensory states given a model of how sensory states are produced; and a *recognition* (or variational) *density* encoding the system’s “beliefs”[6] about worldly states. The recognition density is said to be encoded by the system’s internal states; whereas the generative density is said to be “entailed” by the system’s dynamics, meaning

[5] Notice that here entropy is an information-theoretic quantity, not physical entropy. See (Linson *et al*. 2018) for their relation.

[6] In the FEP literature, beliefs are estimated probabilities of external states of affairs, rather than propositional attitudes with linguistic content.

that the system's dynamics realize the inversion of a generative model (i.e. maps the organism's sensory states on their most likely causes; see Ramstead *et al*. 2020a: 7-8).[7]

Free-energy is an upper bound of surprisal because it can be mathematically decomposed into the sum of surprisal and a the *Kullback-Leibler divergence* between the system's variational density and the true probability distribution over worldly states. Bluntly put, the Kullback-Leibler divergence can be thought of as a measure of how much the organisms' guesses about worldly states are wrong. The lower the divergence, the better the guesses. Since the Kullback-Leibler divergence is always positive, free-energy will always be greater than surprisal. Hence minimizing it will implicitly minimize surprisal, keeping the organism alive.

There are two ways to minimize free-energy. A system can minimize free energy just by minimizing the Kullback-Leibler divergence. This is *perceptual* inference, which does *not* minimize surprisal. It only minimizes the Kullback-Leibler divergence. However, perceptual inference is necessary to ensure that free-energy is a *tight* bound on surprisal. For, only when free-energy is a tight bound on surprisal *active inference* (i.e. a self-generated change of sensory states) can minimize surprisal effectively (Bruineberg *et al.* 2018a).

Perceptual and active inference can be taken as corresponding to a form of perception and action[8] (Corcoran *et al*. 2020). Importantly, in more complex systems[9] free-energy minimization affords an optimal way to balance explorative (or epistemic) actions and exploitative (or pragmatic) actions (Friston *et al*. 2016), while making the agent learn the most efficient and minimalistic routes to success (Tschantz *et al*. 2020). In this way, the FEP

[7] In recent times, these points have been spelled out using complex mathematical constructs (namely, non-equilibrium steady states and dual information geometries) that are not immediately relevant to the purposes of this essay (see Parr *et al*. 2020; Friston *et al.* 2020).

[8] Saying that active inference corresponds to action (i.e. bodily movements fulfilling an intention) is imprecise. In fact, *each and every* self-generated change of sensory state (e.g. sweating to lower one's bodily temperature) is an instance of active inference (see Seth and Friston 2016). Here I'm momentarily sacrificing precision to ease of exposition.

[9] Namely, systems able to quantify their *expected* free-energy; that is, the-free energy expected under various courses of action, see (Friston *et al*. 2013) and (Millidege *et al*. 2020) for discussion.

makes contact with one of the core insights of vehicle externalism; namely the claim that often fast and fluid environmental interactions are the grounds upon which our cognitive successes rest (Clark 2017b).

Crucially, perceptual and active inference are *enabled* by Markov Blankets. Strictly speaking, a MB is a formal property of nodes (i.e. a variable) in graphical models. Graphical models are sets of nodes (representing variables) and directed edges connecting nodes (representing causal or probabilistic relation among variables) used to simplify the computation of complex probability densities (see Koski and Noble 2009 for an introduction). MBs are one tool servicing this simplificatory purpose. Within a graph, a MB is the minimal set of nodes that allows us to accurately estimate the state of the variable of interest. It typically includes the *parents* of the target node (i.e. the nodes *directly leading* to it), its children (i.e. the nodes to which the target node directly leads) and its eventual co-parents (i.e. the nodes directly connected to the target node's children) (Koski and Noble 2009: 50). Technically speaking, the nodes constituting a MB make the target (Blanketed) node *conditionally independent* from all other nodes in the graph; roughly put, the Blanket shields off their influence on the target node. In this way, they allow one to estimate *optimally* the state of a target node by considering only a few (typically a small fraction) of the nodes in a graph, simplifying the computation.

According to the FEP, MB are also *ontologically real*, functional boundaries which separate biological systems from the environment in a non arbitrary way (Friston 2013; Kirchhoff *et al*. 2018).[10] A typical example is that of a cell's membrane. The cell's membrane is a *functionally* relevant boundary which allows the cell to differentiate itself from its

[10] Notice, importantly, that such a reading of MBs as ontologically real functional boundaries does *not* follow directly from (and it is not justified by) the relevant literature on graphical models (see Menary and Gillett 2020; Bruineberg *et al*. 2020 for extensive discussion). However, FEP theorists seem to think it is (e.g. Friston 2013: 2; Kirchhoff and Kiverstein 2019b: 2). Since here I'm reporting the standard presentation of the FEP, and throughout the essay I'm interested only in how MBs, *as the FEP conceives of them*, are used in the debate over vehicle externalism, I concede the point.

environmental surroundings. In the parlance of the FEP, the MB induces a separation between *internal* states (the innards of the cell) and the *external* states (the environment the cell is embedded in). Internal and external states are separated by MBs in the sense that MBs make internal states *conditionally independent* over external states. Notice, importantly, that the insulation is *exclusively* statistical, and never causal. In fact, insofar MBs enable perceptual and active inference, they enable internal and external states to be causally coupled (Friston 2013). This is due to the internal partitions of MBs, and how these partitions interact. According to the FEP, each MB is partitioned into two disjoint sets of states, termed *sensory* and *active* states. The partition is roughly as follows: a state of a MB is a sensory state *iƒ* it is influenced by external states and influences internal (and active) states. Conversely, the state is an active state *iƒ* it is influenced by internal states, and influences external (and sensory) states. Notice that active and sensory states also influence each other, in a way that closely resembles perception-action loops (Fabry 2017). In this way, MBs allow an agent to couple sensomotorically with the environment.

The interplay of internal and external states through active and sensory states allows to formalize the notions of perceptual and active inference in terms of MBs and their states (e.g. Ramstead *et al.* 2018, figure 1). Recall, in perceptual inference, a biological system minimizes the Kullback-Leibler divergence between its “best guesses” and the actual states of the world. Conversely, through *active inference*, a system minimizes surprisal directly, inducing a self-generated, and selective, change in the sensory states it samples.

In perceptual inference, a system attunes its “best guess” (i.e. the variational/recognition density) encoded in its *internal states* to the environmental contingencies (i.e. external states) given the sensory states of its MB. But, given the interplay of internal and active states, these changes will, at least sometimes, change active states, which, in turn, change external and sensory states. Hence, changes in internal states lead to the selective resampling of sensory states and, more broadly, a change of external states so as to make *them* fit the organism's internal states, which is active inference. Thus, perceptual and active inference are naturally captured, and formalized, by MBs. In general, one can think of the coupling between internal and external states (that is, to perceptual and active inference) as the synchronization of internal and external states (Friston 2013; Bruineberg *et al*. 2018: 2433-2440).

One last step. Self-organizing systems must avoid *surprisal*, which is the (negative logarithm of) the probability of sensory states. The closer a sensory state is in the state-space, to the states making up the system's phenotype, the lower its surprisal. But surprisal is also the complement of Bayesian model evidence (Friston 2019: 177). This means that organisms (i.e. internal states) can be seen as a statistical models[11] of the environment which interact with their surrounding so as to acquire *evidence* for their correctness, thereby prolonging their existence. This is why, on the view the FEP proposes, biological self-organization is cast as a process of *self-evidencing* (Friston 2013; Hohwy 2016).

More could be said about the FEP and its explanatory ambitions. But, since here my target is the role MBs play in the debate over vehicle externalism, I believe this simple sketch is sufficient for present purposes. So, how does all this bear on the truth of vehicle externalism?

3 - Markov Blankets do not adjudicate the vehicle internalism/externalism debate

According to the FEP, MBs are ontologically real boundaries enabling perception and action. Given that perception and action intuitively are the interfaces separating the thinking machinery from the environment, it is tempting to resort to MBs to determine whether the

[11] The interpretation of "models" in this context is far from straightforward. A prominent reading applies to a very thin, but still ontologically committed, notion of model, according to which organisms *literally* are models of the organism-niche coupled system just in the sense that they actively regulate their encounters with the environment (Conant and Ashby 1970). Others have suggested an instrumental reading of models instead (van Es 2020; Baltieri *et al*. 2020). Given the purpose of this essay, however, I can stay neutral on the matter (thankfully).

thinking machinery includes environmental or bodily constituents, thereby determining the truth of vehicle externalism.

Yet it seems that doing so immediately begs the question against vehicle externalism. This is because, according to the summary of the FEP presented above, MBs are the boundaries of *organisms*. Vehicle externalism, however:

> "[...] is a view according to which thinking and cognizing may (at times) depend directly and noninstrumentally upon the work of the body *and/or the extraorganismic environment*." (Clark 2008: XXVIII; emphasis added)

Vehicle externalism claims that constituents of the thinking machinery can be located on *either side* of the boundary separating the biological agent from the environment. But, according to the official presentation of the FEP given above, that boundary just is the MB. So it seems that assuming, without any further argument, that MBs demarcate the thinking machinery simply begs the question against vehicle externalism.

Perhaps such an assumption can be justified by some argument *A* showing that the boundary of the organism is the relevant boundary of the thinking machinery. This is surely possible. Yet, in such a case, it would be *A*, rather than any theoretical appeal to MBs, to adjudicate the debate over vehicle externalism (in favor of the internalist). This is because *A* would show that the thinking machinery is entirely contained within the organism, thereby proving that vehicle externalism is false, leaving no role to MBs in adjudicating its truth.

The FEP theorist is now likely to point out I misrepresented MBs, because MBs are *multiple and nested* (Kirchhoff *et al*. 2018; Hesp *et al.* 2019). Cells, each with its own MB, sometimes join forces, constituting multicellular systems that are free-energy minimizers *in their own right*, and thus possess a MB. And in fact, a FEP theorist would claim that we find MBs at every scale of organization, from cells, to tissues and organs (Friston *et al*. 2015; Palacios *et al*. 2020), organisms (Kirchhoff and Kiverstein 2019a), agent in their ecological niches (Bruineberg *et al*. 2018b), and eventually the entire biosphere (Rubin *et al*. 2020).

Therefore, MBs do not *necessarily* coincide with the boundaries of an organism, and thus provide a principled way to identify systems as they are in nature. They can thus be used to identify specific systems, such as our thinking machinery, and determine whether it "extends" (e.g. Hohwy 2016; Ramstead *et al.* 2019; Sims 2020).

The core idea is simple: first, one finds the relevant MB. Then, one looks at what makes up the internal states. If the internal states encompass only neural components, then vehicle externalism is false. Otherwise, vehicle externalism is true. This seems the approach Hohwy (2016) adopted:

> "[...] there is a quite specific account of what happens in active inference, *which puts part of the boundary at the dorsal horn of the spinal cord*. [...] *This tells us how neurocentric we should be*: the mind begins where sensory input is delivered through exteroceptive, proprioceptive and interoceptive receptors and it ends where proprioceptive predictions are delivered, mainly in the spinal cord." (Hohwy 2016: 277; emphasis added)

The idea of using MBs in this way is attractive for a number of reasons. As said above, MBs are taken to be *principled* boundaries. In fact, MBs are "achieved by a system through active inference" (Ramstead *et al*. 2019: 11) and "result from a system's dynamics" (Ramstead *et al*. 2018:3). For this reason, they seem to provide a *non-arbitrary* way to identify systems. Moreover, the identification of MBs is an *empirical* matter, which hinges upon the empirical identification of patterns of statistical independence. In Howhy's quote above, for instance, the relevant account of what happens in active inference is the empirical account provided by (Friston *et al*. 2010). Thus, MBs seem to promise a principled *and empirically sound* solution to the debate over vehicle externalism, providing what many philosophers engaged in that debate have strived to provide (e.g. Kaplan 2012). Moreover, MBs appear able to deliver the desired good *circumventing* the host of thorny philosophical issues that often have halted the debate over extended cognition, such as issues concerning non-derived content (Kirchhoff and Kiverstein 2019a: 70-71; see Piredda 2017 for a nice summary). Yet, it seems to me that this usage of MBs raises at least three distinct problems.

First, the “multiple and nested” view of MBs smuggles a different conception of MBs into the debate. For now MBs are not (or not only) the boundaries of organisms, but rather the boundaries of *biological systems* in general, if not of *systems* in general (e.g. Hipolito 2019; Friston 2019; Friston *et al*. 2020). Perhaps this is the correct conception of MBs, but if it is, it does not immediately follow from the official presentation of the FEP provided above. Indeed it seems hard to square the claim that the FEP is a principled account of biological organization with the claim that MBs demarcate systems *in general*. Surely not every system is a biological system.

Secondly, it is not clear whether allowing such a proliferation of MBs would enable them to play a role in determining the truth of vehicle externalism. This is because, if MBs are multiple and nested within each other, we would need a criterion *C* to determine *which* MB, in this fractal sea of MBs, bounds the (perhaps extended) thinking machinery. However, in such a case, the truth of vehicle externalism would be determined *by C*, rather than the theoretical construct of a MB.[12]

Lastly, and I believe most importantly, there are reasons to deny that MBs can be used to *identify* systems (thinking machinery included). This is because the identification of a system is *logically prior to* the identification of a system’s MB. Hence, in order to identify the MB of the thinking machinery, we must already have determined what the thinking machinery is. But if we already have determined what the thinking machinery is, then we *already know* whether vehicle externalism is true or not: if the thinking machinery coincides with the brain, then vehicle externalism is false, otherwise it is true. As a result, we cannot use MBs to determine whether vehicle externalism is true, on the pain of circularity.

To see why this is the case, consider how MBs are defined in the relevant literature on graphical models. Here’s a canonical[13] definition:

[12] Andy Clark (2017a: 8) made a similar remark, but I think it is important to state the point explicitly.

[13] I’m not reporting Pearl’s (1988: 97) original definition only for brevity. This is because Pearl defines MBs in terms of further technical concepts (namely, independency maps) which require explanation in their own right.

> "Definition 2.20 (Markov Blanket) The Markov blanket of a variable X is the set consisting of the parents of X, the children of X and the variables sharing a child with X. " (Koski and Noble 2009: 50)

Notice that MBs are *defined* in terms of the target (blanketed) variable. The definition might be "expanded" so as to cover more than a variable, thereby capturing all the variables implicated in a description of a given system of interest. But still, given this definition, one *first* identifies a variable (or set of variables) of interest, and *then* identifies the relevant MB of that variable (or set of variables). There is thus no *absolute* notion of MB, and I cannot just point to a graph and ask: "Ok, now tell me where is the relevant Markov Blanket". To ask so, I must already have identified a node, whose MB I'm interested in. The direction of identification runs from variables to MBs, and not the other way around.

The technical literature on the FEP confirms this. For instance, Ramstead *et al.* (2019: 25) argue that we can *choose* the relevant MB, depending on our explanatory interests.[14] But this clearly entails that, in order to choose a relevant MB, we must *already* have identified the system of interest; that is, the system whose behavior we wish to explain. And if the system of interest is the thinking machinery, it follows we must *already* have identified it. But if this is the case, then we *already* know whether it "extends" or not. Similarly, Allen and Friston (2018: 2466) and Clark (2017a) inform us that what counts as the relevant MB depends on our explanatory interests. Even Hohwy (2016)[15] is forced to admit that the choice of what counts as the relevant MB is at least partially pragmatic, and that it depends on our explanatory goals.[16]

[14] A claim that hardly squares with other claims made by Ramstead and colleagues; such as the claim that MBs are "ontological" (see Ramstead *et al.* 2019: 3) and that MBs are "produced" by a system's behavior.

[15] Albeit, in all fairness to Hohwy, he does hold that the MB around the brain is *in principle* privileged, and thus it is the one identifying the thinking machinery proper. I will discuss this point below (see section 4)

[16] Importantly, the mere fact that, given our explanatory interests, we can "place" MBs around putative "extended" thinking machineries provides no vindication of vehicle externalism. For vehicle internalists do acknowledge that explanations of cognitive outputs must take into account environmental factors (Rupert 2009; Sterenly 2010). So, if MBs are identified depending on our explanatory interests, *both* vehicle externalists and internalists can identify "extended" MBs. They would nevertheless disagree on whether these MBs identify all *and only* the constituent parts of the thinking machinery.

This is of course a small sample, but it could be enlarged. However, the point seems reasonably clear: first, one "picks up" a system of interest, *and then* one identifies the system's MB. Thus, if the system of interest is the thinking machinery, one must *already* have identified it to find its MB. But if that machinery has already been identified, then we *already* know whether vehicle externalism is correct. Hence, MBs *cannot* (logically) adjudicate the debate over vehicle externalism, on the pain of circularity.

4 - Markov Blankets are not good framing devices to discuss vehicle externalism

At this juncture, it is easy to imagine the FEP theorist insisting that MBs *do have* a relevance for the debate on vehicle externalism, as they provide a "formal ontology"[17]: a formal framework enabling us to provide a crisp answer to hard philosophical questions, like "is vehicle externalism true?" (cfr. Constant *et al*. 2019; Ramstead *et al*. 2019; 2020b). Here's a clear statement of the idea, applied to the debate over vehicle externalism:

> "The Markov Blanket *formalism* as applied to systems that approximate Bayesian inference serves as *an attractive statistical framework for demarcating the boundaries of the mind.* Unlike other rival candidates for "marks of the cognitive" the Markov Blanket formalism has the virtue of avoiding begging the question in the extended mind debate. [...] The Markov Blanket concept escapes these problems." (Kirchhoff and Kiverstein 2019a: 69-70; emphasis added)

Perhaps this is correct. Maybe asking "which MB, in the fractal sea of MBs, bounds the thinking machinery" yields more satisfactory results than trying to find a "mark of the cognitive" (e.g. Adams and Aizawa 2008).[18]

[17] Notice, importantly, that the relevant conception of MBs has changed again. Now MBs are no longer ontologically real functional boundaries of systems in general, but only posits of a *formal framework*.

[18] However, there seems to be a general problem with this idea. Consider the relevant variables implicated in a psychological explanation. Depression, for instance, is correlated with a range of variables such as "job loss" or "humiliation" (I'm taking this example from Campbell 2007). These variables can be represented in a causal graph, and thus *might* end up constituting the MB of one's thinking machinery. Suppose this happens: in what sense would "being unemployed" be part of the functional boundary separating one's thinking machinery from the environment? I must confess that such a claim strikes me as simply unintelligible. But it also seems a claim that might be licensed by the present framework. Thanks to ANONYMIZED FOR BLIND REVIEW for having pointed this out to me.

Here, I wish to suggest that this is not the case. In fact, I beleive that framing the debate around vehicle externalism in terms of MBs has thus far[19] yielded consequences unpalatable enough to make the whole MB-based approach to vehicle externalism worth reconsidering.

To see why, consider two prominent MB-based approaches to the debate over vehicle externalism. One is Hohwy's (2016; 2017) defense of vehicle internalism; the other is Kirchhoff and Kiverstein's (2019a; 2019b) defence of vehicle externalism.[20]

Importantly, both approaches use MBs to frame the debate over vehicle externalism in roughly the same way. Both approaches take MBs to be "multiple and nested" (Hohwy 2016: 264; Kirchhoff and Kiverstein 2019a: 73-76). As a consequence, both accounts resort to MBs to frame the vehicle internalism/externalism debate in terms of *which* MB should be chosen as the MB bounding the thinking machinery, and *why* that specific MB should be preferred over all the other MBs (e.g. Hohwy 2016: 265; Kirchhoff and Kiverstein 2019a: 79-80). Both accounts agree on the fact that the choice of the relevant MB must be justified using only theoretical resources internal to the FEP. In a sense, thus, both accounts agree upon the fact that, if properly interrogated, the FEP will tell us where the mind stops and the rest of the world begins (Hohwy 2016: 267-273; 2017: 2-4; Kirchhoff and Kiverstein 2019a: 79-81; 2019b: 17-18). Importantly, both accounts agree upon a clear MB-based criterion to identify the boundaries of the mind; namely, that the relevant MB is the MB that identifies the internal states that minimize surprisal *over time*, or *on average and in the long run*. Here's Hohwy spelling it out:

[19] The "thus far" part is important: I'm open to the possibility that there *might* be a perspicuous way to use MBs in the vehicle externalist debate. The FEP theorist is thus challenged to articulate one.

[20] The choice of Hohwy's account as a representative account of the internalist front is somewhat forced by the fact that other philosophers defending forms of internalism (broadly speaking) about predictive processing/the FEP do not defend *vehicle* internalism directly (e.g. Gładziejewski 2017; Wiese 2018). The choice of Kirchhoff and Kiverstein as representatives of the vehicle externalist front is less forced, but still pretty much obliged: Clark (2017a; 2017b) is more concerned with predictive processing rather than the FEP. And (Ramstead *et al.* 2019) seem to believe that the choice of considering "extended" systems depends purely on our explanatory interests; a position that can be squared with an embedded, but still vehicle internalist, view (Rupert 2009; Sterenly 2010).

> “Another, somewhat more principled response [...] is to rank *agents according to their overall, long term prediction error minimization* (or free-energy minimization): the agent worthy of explanatory focus is the system that *in the long run* is best at revisiting a limited (but not too small) set of states. It is most plausible that such a minimal entropy system is constituted by the nervous system of what we normally identify as a biological organism: [...] *extended agents do not maintain low entropy in the long run*” (Hohwy 2016: 265; emphasis added)

where an “agent” is just a system surrounded by a MB. Here’s Kirchhoff and Kiverstein making essentially the same point:

> “The self-evidencing nature of biological agents blocks the threat from cognitive bloat. External resources form part of an agent’s mind when they are *poised to play a part in the process of active inference that keeps surprisal at minimum overtime*. [...] More generally we suggest an external resource will count as a part of an individual’s mind if it is a part of a system whose existence is *produced and maintained* through a self-evidencing process” (Kirchhoff and Kiverstein 2019b: 17-18)

Recall that such an “in the long run” criterion is intrinsic in the structure of the FEP. The FEP is an account of how biological systems (if not systems in general) persists *overtime*. According to the FEP, biological systems persist *through time* by minimizing entropy, that is, surprisal *on average*. And since surprisal is the complement of model evidence, this means that organisms are self-evidencing systems; that is, systems that, overtime, seek the evidence confirming their existence, thereby prolonging it.

Lastly, both accounts thake the “on average and in the long run” criterion to be *extensionally adequate*; that is, apt to single out the MBs enshrouding *all and only* the cogs of the thinking machinery. This is because the criterion is used to solve two deeply related problems concerning the way in which the “boundaries of the mind” are drawn; namely the “cognitive bloat” objection to vehicle externalism (i.e. too much stuff gets counted as a cog in the thinking machinery) and the “shrinking brain” objection to vehicle internalism (i.e. too little stuff gets counted as a cog, see Anderson 2017) at once.

The number of premises shared by the accounts proposed by Kirchhoff, Kiverstein and Hohwy immediately invites the following question: if the premises are the same, then why do the conclusions differ? If they all espouse the same premises[21] and the same relevant criterion to identify the thinking machinery, their conclusions *should* be the same. So, apparently, either Hohwy or Kirchhoff and Kiverstein *mis-applied* the criterion. This, it seems to me, suggests that framing the debate over vehicle externalism in terms of MBs does not, in and by itself, lead us *straight* to a solution of the debate. Now, one might object that framing that debate in terms of MBs does not, in and of itself, make that debate easier to solve. Fair enough. But then, why bother with the *formal framework* they provide? What sort of theoretical boon are MBs providing here?

I am inclined to answer "none". In fact, I believe that the criterion Hohwy, Kirchhoff and Kiverstein derive from the FEP is grossly extensionally inadequate. Recall that, during active inference, an agent "brings about" the sensory states it expects to encounter. Importantly, these states encompasse *all* the variables that define the state-space in which the agent phenotype is embedded. For us humans (an, broadly speaking, animals) this includes extero-, intero- and viscero-ceptive states (e.g. Seth and Friston 2016). Hence, we (and animals in general) must minimize free-energy in respect to all these states.

Consider now the following, often used, example (e.g. Bruineberg 2018: 3; Bruineberg *et al.* 2018a: 2423; Ramstead *et al.* 2019: 9, Veissière *et al.* 2020): human beings expect their bodily temperature to be around 36.6°. For a human, having a bodily temperature around 36.6° is the least surprisaling state; and deviations from that state, whether they increase or decrease the bodily temperature, increase surprisal. So, when our bodily temperature deviates from the predicted 36.6°, we engage in active inference to avoid dangerous surprisaling states. We do so, for instance, by sweating, so as to lower our bodily temperature when it is too high.

[21] One might object that Kirchhoff, Kiverstein and Hohwy do not *really* espouse the same premises, as their theoretical commitments differ widely. For instance, whereas Hohwy holds that the FEP yields a representationalist and inferentialist account of cognition (e.g. Kiefer and Hohwy 2019), Kirchhoff and Kiverstein argue that the FEP leans towards a radical form of enactivism (Kirchhoff and Kiverstein 2019a). This is surely correct. However, it is not clear how these different theoretical commitments bear onto the question at hand. In fact, vehicle externalism is surely compatible with both representationalism (e.g. Clark 2008) and anti-representationalism (e.g. Chemero 2009).

Or by trembling, so as to raise it when it is too low. We also keep our bodily temperature around 36.6° *by wearing appropriate clothes*. And clothes appear to be part of the physical machinery by means of which we minimize free energy, and thus avoid surprisal overtime, on average and in the long run: we wear clothes more often than not, and we surely wear them with the purpose of keeping our bodily temperature around 36.6°. It thus seems correct to conclude that, according to Kirchhoff, Kiverstein and Hohwy's criterion, the relevant MB identifying the thinking machinery will include clothes. But this conclusion surely seems wrong.[22] So, it seems correct to conclude that the proposed criterion is *not* extensionally adequate: it counts too much stuff as a cog in the thinking machinery.

Secondly, and perhaps more decisively, it seems that when it comes to *internal* (i.e. neural) vehicles, we do not judge whether they qualify as parts in the thinking machinery based on their role in free-energy minimization *on average and in the long run*. Hence, that criterion violates the core insight that the "parity principle" is trying to express; namely, that we should judge whether candidate external vehicles are part of our thinking machinery *with the same metric* we deploy to judge internal vehicles (Clark 2008; 77-78; 2013: 195).[23]

Consider the following scenario: after a severe head injury, a child gets a part of her brain *x* explanted at time *t*. After the surgery, she recovers and goes on to live a long (and cognitive unimpaired) life. It seems intuitively correct to say that, after *t*, the neural region *x* does not count as a cog in her thinking machinery. But it seems equally intuitively correct to say that, *before t*, the neural region *x* actually *was* a cog in her thinking machinery.[24] That is, at time *t*-

[22] Minimally, because the constituents of our thinking machinery are supposed to be information-processors of some sort, but clothes do not appear to be information processors of any kind. I suspect, however, that even the most extreme proponents of vehicle externalism would concede me this point without resistance.

[23] Vehicle externalists that emphasize the *complementarity* of inner and outer resources (e.g. Menary 2007; 2018; Kirchhoff and Kiverstein 2019a) find the parity principle problematic, as it might suggest that internal and external resources must be functionally similar. Importantly, however, even vehicle externalists stressing complementarity agree on the fact that whether a putative vehicle counts as a cog in the mental machinery depends *exclusively* on the sort of task it performs in the relevant sort of processing in which it takes part, regardless of its spatial location. Thus, they agree with the parity principle as stated in the main text. This point is sometimes explicitly acknowledged (Menary 2007: 55-57; Gallagher 2018).

[24] There is, to be sure, a call to intuition here. But I think it is fine, as, at the end of the day, determining what really qualifies as a cog in the mental machinery *is* based on our intuitions about what counts as cognitive (see

1 it seems intuitively correct to judge *x* a cog in the thinking machinery. And, more importantly, it seems unlikely that, at *t-1*, we *would* revise such a verdict, were we to discover that, due to an historical accident, *x* will not partake in free-energy minimization on average and in the long run (by stipulation, since "the owner" of *x* is a child, she spends most of her life without *x*). In other words, it seems correct to say that, *when x is appropriately wired*, it just is a cog in the thinking machinery, regardless of what its future career as a piece of a free-energy minimizing engine will be. The fact that a putative piece (neural or non-neural) of the thinking machinery can be contingently decoupled from the rest of that machinery by some future event "does not rule out cognitive status", as Clark and Chalmers (1998: 11) wrote.

Notice further that, albeit in less extreme form, many purely neural "candidate cogs" of our thinking machinery do not end up performing free-energy minimization on average and in the long run. Consider, for instance, synaptic pruning. According to the FEP, such a process should be understood in terms of a reduction in model parameters, bolstering neuronal efficiency (Friston 2010: 131). But such a description of synaptic pruning makes sense only if we concede that the "pruned" synapses *were parameters of the model* seeking evidence for itself. Yet, synaptic pruning is a process that naturally happens during development (e.g. Changeaux 1985), when one is still a child. Hence it seems that we are committed to the claim that the relevant model (i.e. the internal states enshrouded by a MB) has genuine constituents which are not there *in the long run*, and thus cannot contribute to long-term error minimization. Moreover, a neuronal region might fail to perform its own free-energy minimization duties in the long run without having to "leave the brain", for instance as a result of a disconnection syndrome (see Parr and Friston 2020). Yet, it seems correct to say that such a neural region is still a cog in the thinking machinery - indeed, it is only *because* such a cog is damaged that we can account for the symptoms brought about by the disconnection syndrome. Lastly, under normal conditions, neural regions organize in

Clark 2010: 53-54; 2019: 277); at least, until a suitably uncontested "mark of the mental/cognitive" is provided.

"transient" task specific neuronal devices (see Anderson 2014; Clark 2017b for a "predictive" take on the issue). But it is far from clear whether any such transiently created device performs free-energy minimization in the long run. Yet it seems intuitively correct to count them as cogs in the thinking machinery nevertheless.

Now, if all of this is true for neural candidate vehicles *and the parity principle is correct*[25], then the same must hold for putative external vehicles. Hence, given if we would not apply the "overtime, on average and in the long run" criterion to pieces of the brain, we should not apply it to putative external vehicles. And since (at least intuitively) the antecedent is correct, the consequent follows.

Now, perhaps we could substitute the "overtime, on average and in the long run" criterion with something better. However, the "overtime, on average and in the long run" criterion is taken to directly "fall off" out of the FEP. And, if fact, both Kirchhoff and Kiverstein (2019a: 80-81; 2019b 17-18) and Hohwy (2016: 272) derive it directly by the self-evidencing nature of living systems, for self-evidencing *just is* minimizing free-energy overtime, in the long run (e.g. Friston 2013; Friston *et al*. 2020). Hence, if this is correct, there seems to be no easy way to displace the "overtime, on average and in the long run" criterion *without* thereby introducing substantial modifications in the theoretical architecture of the FEP itself.

Perhaps the "overtime, on average and in the long run" criterion could be complemented by some further criterion, ensuring that the relevant thinking machinery is identified in an extensionally adequate and non question-begging way. The important thing to notice, I believe, is that such a criterion would be in the task of *correcting* the verdicts yielded by the "overtime, on average and in the long run" criterion. This heavily suggests that MBs do not provide a *good* theoretical framework for articulating the vehicle externalism/internalism

[25] Of course, one could provide an argument against the parity principle and counter this argument. But such an argument would effectively be a refutation of vehicle externalism, and so it would solve the vehicle externalism debate (in favor of vehicle internalism), leaving MBs no role to play in it.

debate - in fact, we would need a criterion precisely to correct the shortcomings of such a framing.

5 - Is vehicle externalism (conditioned over Markov Blankets) possible?

Thus far, I have argued that MBs will not resolve the vehicle externalism/internalism debate, and that framing such a debate in terms of MBs has thus far yielded very unwelcome theoretical consequences. Here, I wish to claim that we cannot *meaningfully* frame the vehicle externalism/internalism debate in terms of MBs, for doing so makes vehicle internalism *vacuously true*, leading to a purely verbal solution of the debate. My argument hinges on two premises.

The first premise is that the *relevant* meaning of "external(-ism)" and "internal(-ism)" is defined in terms of MBs, as seen in section 2. Recall: according to the FEP, what counts as internal and external depends on the presence of some relevant MB. This premise is widely shared in the FEP literature (e.g. Friston 2013; Wiese 2018: 223-227; Kirchhoff *et al.* 2018). Here's the way in which Hohwy puts it:

> "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? [...] A better answer is provided by the notion of Markov Blankets and self-evidencing through approximation to Bayesian inference. *Here is a principled distinction between the internal, known causes as they are identified by the model, and the external, hidden causes on the other side of the Markov Blanket.*" (Hohwy 2017: 6-7, emphasis added)

It seems to me there isn't much more to say: the meaning of "internal(-ism)" and "external(-ism)" is fixated by the relevant MB (see also Ramstead *et al*. 2019).

The second premise is that we should identify the thinking machinery by means of MBs. Again, this is a premise widely shared in the literature over "predictive" vehicle externalism. I think the references given in the previous sections substantiate this claim enough.

But then, if the thinking machinery is enshrouded by an MB, and if what is enshrouded by an MB is by definition *internal* in the relevant sense, then all the vehicles of the thinking machinery are by definition internal, vehicle internalism is by definition true, and everyone engaged in the debate over predictive vehicle externalism is by definition a vehicle internalist. In the continuation of the passage cited above Jackob Hohwy *almost* noticed the issue:

> "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. [...] *Notice that this definition of internalism makes Clark an internalist*" (Hohwy 2017: 6-7, emphasis added)

But if this is the case, then we should reject the proposed definition of "internal(ism)" and "external(ism)". We wish that our relevant definitions capture *at least* paradigmatic instances of the thing being defined. Hence, our relevant definition of "(vehicle) externalism" should capture at least paradigmatic instances of vehicle externalism; and the works of Andy Clark surely are one such instance. Hence, it seems correct to conclude that if MBs provide us with a partition between internal and external, then that partition is not *the relevant partition at issue* in the debate over vehicle externalism.

My argument has two premises. A good way to resist it is to deny one of them. Can premise one be denied? Well, the first premise is just that "internal" and "external" should be defined in reference to MBs. We can surely deny this, but this invites the question: if MBs do not decide what counts as internal or external, then why are they relevant to the vehicle externalism debate? Moreover, denying that MBs define what counts as internal and external seems in stark contrast with the FEP. So, I do not think the FEP theorist is free to deny premise one.

Does the denial of premise two lead to a better outcome? Well, since premise two is the claim that the thinking machinery should be identified by means of MBs, denying it seems just to *give up* on MBs, at least when it comes to draw the boundaries of the thinking machinery.

Perhaps it could be argued that premise one and premise two are fine, and that vehicle internalists have won the debate *via* MBs. As far as I can see this is a technically viable move, but not an *attractive* one; not even for vehicle internalists. In fact, accepting both premises makes vehicle externalism impossible. But everyone engaged in the debate over vehicle externalism claims that it is possible, vehicle internalists included (e.g. Adams and Aizawa 2008: 25-28). So, it seems that if vehicle internalists accept that vehicle externalism is possible (as they do), then they cannot accept a MB-based definition of "internal(-ism)" and "external(-ism)", on the pain of contradiction.

Vehicle internalists might however be tempted to accept the MB-based definition of "internal(-ism)" and "external(-ism)" and deny that vehicle externalism is possible. But doing so looks like *changing the topic of the conversation*. For the relevant debate concerns whether or not vehicle externalism is true *in the actual world*. Of course, it is possible to claim that something is not true in the actual world by claiming it is not true in all possible worlds. But such a claim needs to be justified, and the worries raised in the previous section suggest calling upon MBs to determine the truth of vehicle externalism is not a well-justified move.

Moreover, I doubt such a redefinition of "internalism" would buy the internalist something more than a purely *verbal* victory. For, were "internalism" to be redefined in such a way, there *would still be* a clash among internalists who believe that internal states are purely neural and internalists who believe that, at least sometimes, the internal states are not purely neural. It thus seems that accepting both premises does make vehicle internalism *vacuously true*. For, it seems that, thus secured, the truth of vehicle internalism has no relevant consequence - apart from forcing us to refer to vehicle externalism as "vehicle internalism", in a confusing way.

I thus recommend to abandon *at least* one of the two premises above. Given that abandoning premise one runs counter to the FEP, I believe the FEP theorist is better off giving up premise two; that is, I believe the FEP theorist should acknowledge that MBs do not matter in the debate over vehicle externalism.

6 - Concluding remarks

I have argued that MBs are not relevant to the debate over vehicle externalism. If the arguments I've provided here are on the right track, MBs do not solve, nor help to solve, the debate surrounding "the extended mind".

Importantly, I do not take my arguments to be "knockdown" arguments. I'm willing to concede that there might be some yet-to-be-discovered way to fruitfully apply MBs in the debate over vehicle externalism. So perhaps what I'm really doing here is challenging FEP enthusiasts to show us that there is such an application.

I would like to conclude by noticing that the challenge here raised might be deeper than my arguments show. Up to this point, I've been neutral on the metaphysical status of MBs, accepting (for the sake of discussion) that MBs really are as the FEP theorist thinks of them. Importantly, however, it is far from clear that the FEP theorist conceives MBs in only one way, as sometimes noticed in the essay. More generally, the metaphysical status of MBs is far from clear in the literature over the FEP. As extensively argued in (Bruineberg *et al*. 2020; Menary and Gillet 2020), MBs are, strictly speaking, only formal properties of nodes in graphical models; and it is not *immediately* clear whether they also denote the functional boundaries of biological agents, or systems in general. It thus seems to me that the FEP theorist faces a double challenge. First, the FEP theorist must determine the metaphysical status of MBs. Secondly, the FEP theorist must show that MBs are a viable tool to determine where the thinking machinery stops and the rest of the world begins. Time will tell how (and whether) these challenges will be met.

References

Adams, F., & Aizawa, K. (2008). *The Bounds of Cognition*. Oxford: Blackwell.

Allen, M., & Friston, K. (2018). From cognitivism to autopoiesis: towards a computational framework for the embodied mind. *Synthese*, *195*(6), 2459-2482.

Anderson, M. L.(2014). *After Phrenology*. Cambridge, MA.: The MIT Press.

Anderson, M. L. (2017). Of Bayes and bullets. In T. Metzinger, W. Wiese (Eds.), *Philosophy and Predictive Processing*: 4. Frankfurt am Main, The MIND Group. https://doi.org/10.15502/9783958573055.

Baltieri, M., *et al*. (2020). Predictions in the eye of the beholder. An active inference account of Watt governors. *ALIFE 2020:The 2020 Conference on Artificial Life*. https://doi.org/10.1162/isal_a_00288

Bruineberg, J. (2018). Active inference and the primacy of the "I can". In T. Metzinger, W. Wiese (Eds.), *Philosophy and Predictive Processing*: 5. Frankfurt am Main, The MIND Group. https://doi.org/10.15502/9783958573062.

Bruineberg, J., *et al*. (2018a). The anticipating brain is not a scientist: the free-energy principle from an ecological-enactive perspective. *Synthese*, *195*(6), 2417-2444.

Bruineberg, J. *et al*. (2018b). Free-energy minimization in joint agent-environment systems: a niche construction perspective. *Journal of Theoretical Biology*, *455*, 161-178.

Bruineberg, J. *et al.* (2020). The emperor's new Markov Blankets. *Preprint*. Retrieved at: http://philsci-archive.pitt.edu/18467/ Last accessed: 30/12/2021

Buckley, C. *et al.* (2017). The free-energy principle for action and perception: a mathematical review. *Journal of Mathematical Psychology*, *81*, 55-79.

Campbell, J. (2007). An interventionist approach to causation in psychology. In A. Gopnik, L. Schulz (Eds.), *Causal Learning* (pp.58-67). New York: Oxford University Press.

Changeaux, J. P. (1985). *Neuronal Man*. New York: Pantheon Books.

Chemero, A. (2009). *Radical Embodied Cognitive Science*, Cambridge, MA.: The MIT Press.

Clark, A. (1998). Author's response. *Metascience*, *7*, 95- 103

Clark, A. (2008). *Supersizing the Mind*. New York: Oxford University Press.

Clark, A. (2010). Memento's revenge: the extended mind, extended. In R. Menary (Ed.), *The Extended Mind* (pp. 43-66). Cambridge, MA.: The MIT Press.

Clark, A. (2017a). 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: The MIND Group. https://doi.org/10.15502/9783958573031.

Clark, A. (2017b). Busting out: predictive brains, embodied minds, and the puzzle of the evidentiary veil. *Nous*, *51*(4), 727-753.

Clark, A. (2019). Replies to critics. In M. Colombo, E. Irvine, M. Stapleton (Eds.), *Andy Clark and His Critics*, (pp. 266-302). New York: Oxford University Press.

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

Colombo, M, & Wright, C. (2018). First principles in the life sciences: the free-energy principle, organicism and mechanism. *Synthese*, https://doi.org/10.1007/s11229-018-01932-w

Conant, R. C., & Ashby, R. W. (1970). Every good regulator of a system must be a model of that system. *International Journal of System Science*, *1*(2), 89-97.

Constant, A., *et al*. (2021). Representation wars: enacting armistice through active inference. *Frontiers in Psychology*, *11*: 598733.

Corcoran, A., Pezzulo, G., & Hohwy, J. (2020). From allostatic agents to counterfactual cognizers: active inference, biological regulation, and the origins of cognition. *Biology and Philosophy*, *35*(3), 1-45.

Di Paolo, E. (2009). Extended Life. *Topoi*, *28*(1): 9-21.

Fabry, R. E. (2017). Transcending the evidentiary boundary: prediction error minimization, embodied interaction, and explanatory pluralism. *Philosophical Psychology*, 30(4), 395-414.

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

Friston, K. (2011). Embodied inference, or: "I think therefore I am, if I am what I think". In W. Tschacher, C. Bergomi (Eds.), *The Implications of Embodiment (Cognition and Communication*) (pp. 89-125). Exeter: Imprint Academic.

Friston, K. (2012). A free-energy principle for biological systems. *Entropy*, *14*(11), 2100-2121.

Friston, K. (2013). Life as we know it. *Journal of the Royal Society Interface*, *10*(86): 20130475.

Friston, K. (2019). Beyond the desert landscape. In M. Colombo, E. Irvine, M. Stapleton (Eds.), *Andy Clark and His Critics*, (pp. 174-190). New York: Oxford University Press.

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

Friston, K., *et al*. (2010). Action and behavior: a free-energy formulation. *Biological Cybernetics*, *102*(3), 227-260.

Friston, K. *et al.* (2012). Free-energy minimization and the dark-room problem. *Frontiers in Psychology*, *3*: 120.

Friston, K. *et al*. (2013). The anatomy of choice: active inference and agency. *Frontiers in Human Neuroscience*, *7*:598.

Friston, K*., et al*. (2015). Knowing one's place: a free energy approach to pattern regulation. *Journal of the Royal Society Interface*, 12(105), 20141383.

Friston, K., *et al*. (2016). Active inference and learning. *Neuroscience and Biobehavioral Reviews*, *68*, 862-879.

Friston, K. *et al*. (2020). Sentience and the origin of consciousness: from cartesian duality to Markovian monism. *Entropy*, *22*(5): 516.

Gallagher, S. (2018). The extended mind: state of the question. *The Southern Journal of Philosophy*, *56*(4), 421-447.

Gładziejewski, P. (2017). Just how conservative is conservative predictive processing?. *Internetowy Magazyn Filozoficzny Hybris*, *38*, 98-122.

Hesp, C., *et al.* (2019). A multi-scale view of the emergent complexity of life: a free-energy proposal. In G. Georgiev, J. Smart, C. L. Flores Martinez, M. Price (Eds), *Evolution, Development, Complexity: multiscale models in complex adaptive systems* (pp. 195-127), New York: Springer.

Hipolito, I. (2019). A simple theory of every "thing". *Physics of Life Reviews*, *31*, 79-85.

Hohwy, J. (2016). The self-evidencing brain. *Nous*, *50*(2), 259-285.

Hohwy, J. (2017). How to entrain your evil demon. In T. Metzinger, W. Wiese (Eds.), *Philosophy and Predictive Processing*, 2, Frankfurt am Main: The MIND Group, https://doi.org/10.15502/9783958573048.

Hohwy, J. (2020). Self-supervision, normativity and the free-energy principle. *Synthese*, https://doi.org/10.1007/s11229-020-02622-2

Hurley, S. (2010). The varieties of externalism. In R. Menary (Ed.), *The Extended Mind* (pp. 101 - 154). Cambridge, MA.: The MIT Press.

Kaplan, D. M. (2012). How to demarcate the boundaries of cognition. *Biology and Philosophy*, *27*(4), 545-570.

Kiefer, A., & Hohwy, J. (2019). Representation in the prediction error minimization framework. In S. Robins, J. Symons, P. Calvo (Eds.), *The Routledge Companion to Philosophy of Psychology* (2nd Ed.) (pp. 384-410). New York: Routledge.

Kirchhoff, M. D., & Kiverstein, J. (2019a). *Extended Consciousness and Predictive Processing: a Third Wave View*. New York: Routledge.

Kirchhoff, M. D., & Kiverstein, J. (2019b). How to demarcate the boundaries of the mind: a Markov Blanket proposal. *Synthese*, https://doi.org/10.1007/s11229-019-02370-y

Kirchhoff M. D., *et al*. (2018). The Markov Blankets of life: autonomy, active inference and the free-energy principle. *Journal of the Royal Society Interface*, *15*(138): 20170792.

Kiverstein, J. (2018). Extended cognition. In A. Newen, L. De Bruin, S. Gallagher (Eds.), *The Oxford Handbook of 4E Cognition*, (pp. 19-41). New York: Oxford University Press.

Koski, T., & Noble J. M. (2009). *Bayesian Networks: an Introduction*. Chichester, Wiley and Sons.

Linson, A. *et al*. (2018). The active inference approach to ecological perception: general information dynamics for natural and artificial embodied cognition. *Frontiers in Robotics and AI*, *5*: 21

Menary, R. (2007). *Cognitive Integration: Mind and Cognition Unbound*. Basingstoke: MacMillan.

Menary, R. (2018). Cognitive integration: how culture transforms us and extends our cognitive capabilities. In A. Newen, L. De Bruin, S. Gallagher (Eds.), *The Oxford Handbook of 4E Cognition*, (pp.187-216). New York: Oxford University Press.

Menary, R., & Gillett, A. J. (2020). Are Markov Blankets real and does it matter? In Merdoça, D. Curado, S., Gouveia S. (Eds.). *The Philosophy and Science of Predictive Processing* (pp. 39-58). London: Blomsbury Academic.

Millidege, B. *et al.* (2020). Whence the expected free-energy?. *Preprint*, arXiv preprint arXiv:2004.08128.

Palacios, E. E., *et al*. (2020). On Markov Blanket and hierarchical self-organization. *Journal of Theoretical Biology*, 486:110089.

Palermos, S. O. (2014). Loops, constitution, and cognitive extension. *Cognitive System Research*, *27*, 25-41.

Parr, T., & Friston, K. (2020). Disconnection and diaschisis: active inference in neuropsychology. In Mendoça, D. Curado, S., Gouveia S. (Eds.). *The Philosophy and Science of Predictive Processing* (pp. 171-186). London: Blomsbury Academic.

Parr, T. *et al*. (2020). Markov Blankets, information geometry and statistical thermodynamics. *Philosophical Transactions of the Royal Society A: mathematical, Physical and Engineering Sciences*, *378*(2164), 20190159.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. San Francisco: Morgan Kauffman.

Piredda, G. (2017). The mark of the cognitive and the coupling-constitution fallacy: a defense of the extended mind hypothesis. *Frontiers in Psychology*, *8*: 2061.

Ramstead, M. J. D., *et al*. (2018). Answering Schrödinger's question: a free energy formulation. *Physics of Life Reviews*, *24*, 1-16.

Ramstead, M. J. D., *et al*. (2019). Multiscale integration: beyond internalism and externalism. *Synthese*, https://doi.org/10.1007/s11229-019-02115-x

Ramstead, M. J. D., *et al*. (2020a). A tale of two densities. Active inference is enactive inference. *Adaptive Behavior*, *28*, (4), 225-239.

Ramstead, M. J. D., *et al*. (2020b). Is the free energy principle a formal theory of semantics? From variational density dynamics to neural and phenotypic representation. *Entropy*, *22*(8): 889.

Rowlands, M., *et al*. (2020). Externalism about the mind. In E. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (winter 2020 edition) URL = <https://plato.stanford.edu/archives/win2020/entries/content-externalism/>

Rubin, S., *et al*. (2020). Future climates: Markov Blankets and active inference in the biosphere. *Journal of the Royal Society Interface*, *17*:20200503.

Rupert, R. (2009). *Cognitive Systems and the Extended Mind*. New York: Oxford University Press.

Seth, A. K. (2020). Preface: the brain as a prediction machine. In D. Mendoça, M. Curado, S. Gouveia (Eds.), *The Philosophy and Science of Predictive Processing*. London: Bloosmbury Academic.

Seth, A. K., & Friston, K. (2016). Active interoceptive inference and the emotional brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1708), 20160007.

Sims, M. (2020). How to count biological minds: symbiosis, the free-energy principle, and reciprocal multiscale integration. *Synthese*, https://doi.org/10.1007/s11229-020-02876-w

Sterenly, K. (2010). Minds: extended or scaffolded?. *Phenomenology and the Cognitive Sciences*, *9*(4), 465-481.

Tschantz, A. *et al*. (2020). Learning action-oriented models through active inference. *PLoS Computational Biology*, *16*(4), e1007805.

Van Es, T. (2020). Living models or life modelled? On the use of models in the free energy principle. *Adaptive Behavior*, 1059712320918678.

Veissière, S. *et al*. (2020). Thinking through other minds: A variational approach to cognition and culture. *Behavioral and Brain Sciences* (Accepted preprint) https://doi.org/10.1017/S0140525X19001213

Wiese, W. (2018). *Experienced Wholeness*, Cambridge, MA.: The MIT Press.

Wiese, W., & Metzinger, T. (2017). Vanilla PP for philosophers: a primer on predictive processing. In T. Metinger, W. Wiese (Eds.), *Philosophy and Predictive Processing*. Frankfurt am Main: The MIND Group. https://doi.org/10.15502/9783958573024.