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 misplaced, 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 will provide us with an answer based on circular reasoning. In section 4, I consider whether Markov Blankets help us track the boundaries of the mind, answering in the negative. This is because resorting to Markov Blankets to track the boundaries of the mind yields extensionally inadequate conclusions which violate the parity principle. In section 5, I further argue that Markov Blankets led us to sidestep the debate over the extended mind, as they make internalism about the mind vacuously true. 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 is only a claim concerning the *physical constituents* (vehicles) of the thinking machinery. Hence, it is compatible with different accounts of how the thinking machinery functions, including computationalism (Clark 2008), ecological psychology (Chemero 2009), enactivism (Di Paolo 2009), dynamicism (Palermos 2014) and more. As a consequence, how

[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" to indicate the system that is supposed to extend, be it a cognitive, conscious or mental system.

vehicle externalism should be articulated and whether or not it is true are intensely debated topics (Kiverstein 2018; Rowlands *et al*. 2020).

The recent popularity of "predictive" approaches to the mind, especially Friston's free-energy principle (henceforth FEP e.g. Friston 2010), generated a wave of "predictive" vehicle externalism (e.g. Clark 2017a) counterbalanced by equally consistent wave of "predictive" vehicle internalism (e.g. Hohwy 2016). Their clash rapidly centered around *Markov Blankets* (henceforth 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, arguing that MBs neither adjudicate, nor *help* to adjudicate, whether vehicle externalism is true. In other words, my aim here is to examine whether MBs play a valuable role in determining whether vehicle externalism is true, suggesting that MBs *do not* play such a valuable role.

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 simplify the tracking of the boundaries of the thinking machinery, showing that, at least thus far, the usage of MBs has delivered unpalatable verdicts and has been incompatible with the parity principle. In section 5, I argue further that 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] 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*.

Before I start, however, I need to explicitly place two *caveats*.

Caveat #1: my focus concerns *exclusively* the role MBs are supposed to play in the debate over vehicle externalism. So, I will characterize the FEP as it is characterized in that debate; namely, as an account of life and cognition "from first principles". I will thus introduce the FEP as a non-empty, conceptually/mathematically laden description of how living systems persist through time and display adaptive and intelligent behaviors (cfr. Bruineberg 2018; Bruineberg *et al*. 2018; Clark 2017a; Colombo and Wright 2018; Constant *et al*. 2019; Corcoran *et al*. 2020; Fabry 2017; 2021; Friston 2013; Hohwy 2016; 2017; Kirchhoff and Kiverstein 2019a, 2019b; Kirchhoff *et al*. 2018; Kiverstein and Sims 2021; Linson *et al*. 2018; Palacios *et al*. 2020; Ramstead *et al*. 2019; Sims 2020). This is not the only way in which the FEP can be understood[3]; but issues concerning the FEP status as a theoretical object lie beyond the scope of the present treatment.

Caveat #2: relatedly, I will *assume* that the FEP comes with genuine ontological commitments, among other things, to Markov Blankets as real and objective boundaries of living/biological systems (see references given above).[4]

These *caveats* seem to me justified by the principle of charity: the theoretical status and commitments of the FEP are surely debated, but, given my aim here, the principle of charity suggests to assume each party engaged in the dispute over "predictive" vehicle externalism correctly interprets the FEP and its commitments. Moreover, these caveats entail a reading of the FEP that is *charitable*, at least given the purpose of this paper. Vehicle externalism and vehicle internalism are *fact stating* claims concerning what *really and objectively* are the

[3] For example, it could be understood as a framework or toolbox for model-building (e.g. Andrews 2021; Raja *et al*. 2021) or as a conceptual/mathematical analysis of *systems in general* (e.g. Hipolito 2019; Friston 2019).

[4] As above, this is not universally accepted; for instance (Bruineberg *et al*. 2020) argue that these commitments are due to a projection of formal properties of models over systems modelled, (Menary and Gillett 2020) claim that such commitments are not intrinsic to the FEP, but descend from an implicit adoption of a pythagorean/platonic metaphysics, and (Baltieri *et al*. 2020, van Es 2020) suggest to take an instrumentalist stance towards the FEP more generally.

constituents of our thinking machinery. They are not *epistemic* claims concerning how the thinking machinery is best studied. Nor are they claims spelling out otherwise useful fictions. Thus, when looking for the boundaries of the thinking machinery, one looks for something objective and "out there".[5] So, if one takes MBs to be such boundaries, one *must* take them to be objective and "out there".

Notice that these two *caveats* make my claim conditional: what I'm going to argue is that, *given the assumptions spelled out via caveat #1* and *#2*, MBs neither adjudicate nor help to adjudicate whether vehicle externalism is true.[6]

Notice further that these two caveats entail that I will *not* systematically distinguish MBs as formal properties of variables (or "Pearl Blankets") from MBs as real and objective boundaries of free-energy minimizing system (or "Friston Blankets"; see Bruineberg *et al.* 2020; Menary and Gillett 2020). For such a distinction is either not acknowledged in the literature on the FEP I'm interested in, or, if it is acknowledged, it is downplayed, in a way that strongly suggests that "Friston Blankets" are an unproblematic development of "Pearl Blankets" (see, for instance, Wiese and Friston 2021). This (I believe) makes the present treatment *complementary to* the analysis offered in (Bruineberg *et al.* 2020; Menary and Gillett 2020). If I understood them correctly, these authors contend (among other things) that

[5] Notice, for the sake of clarity, that when vehicle internalist and vehicle externalist make claims about the "boundaries of the thinking machinery" they need not commit to the existence of what I will here call a *fence*; that is, a physical object having contiguous spatiotemporal parts which demarcate the perimeter of the spatiotemporal region within which all and only the constituents of the thinking machinery are located. The "boundaries" of the thinking machinery might, but *need not*, consist in such a fence. For example, Chalmers (2008; 2019) has suggested that it is intuitive to think at such boundaries as constituted by perception and action. Clearly, Chalmers is *not* suggesting that perception and action form a single physical object with contiguous spatiotemporal parts encasing all the cogs of the thinking machinery. Rather, he is suggesting that perception and action are (intuitively) the *functional interfaces* separating the thinking machinery from the environment.

[6] One might wonder what would happen if one were to let these assumptions go. I think that what would happen is roughly this: that one stops regarding the FEP as a non vacuous, mathematically/conceptually leaded description of how living systems persist through time and display intelligent/adaptive behaviors, and that one stops regarding MBs as real and objective boundaries of living and/or cognitive systems. And once one stops taking MBs in such a way, one has let go of the idea that MBs matter when it comes to adjudicating the boundaries of cognition.

the usage of MBs to demarcate the real and objective boundaries of free-energy minimizing systems needs to be justified further. Here, I will instead assume that such an usage is perfectly justified (this is conceded by the two *caveats* above) and argue that, even in this case, MBs are not able to play the desired role, at least when it comes to demarcating the boundaries of the free-energy minimizing thinking machinery.

With these *caveats* in place, it is now time to briefly introduce the FEP (readers familiar with the literature I'm considering here might wish to skip to section 3).

2 - The Free-energy principle: a selective sample of selective sampling

The FEP states that the persistence of living systems is guided by *free-energy minimization* (Friston 2011; 2012; 2013; Friston and Stephan 2007). Consider an organism's prolonged existence. In order to continue to exist through time, an organism must find, in the space of all its possible states, the *subset* of states compatible with its prolonged existence, which it must continuously visit and re-visit. For instance, a human that "wants to" continue existing must continue to visit states in which their bodily temperature is around 36.6°. Failures to occupy these states might cause harm or even death (e.g. if the bodily temperature goes to 154°).

The FEP formalizes this idea claiming that an organism's existence defines a probability distribution over the space of all its possible states, and that such a probability distribution has low *entropy*[7]; i.e. it is sharply peaked around the states that the organism must continuously re-visit to prolong its existence. Since entropy is the long term average of *surprisal* (i.e. the negative logarithm of a state's probability), minimizing surprisal over time will ensure that the organism constantly revisits the "right" states (Friston 2011: 92-93). So, an organism's prolonged existence can be ensured by a process of surprisal avoidance.

[7] Notice that *this is not physical entropy*, but rather information-theoretic entropy. See (Linson *et al.* 2018) for further discussion of this point.

Yet organisms cannot track surprisal. They can, however, track its upper bound, which is (variational) free-energy (Buckley 2017). Organisms can track it 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 estimate about worldly states. The recognition density is encoded by the system's internal states; whereas the generative density is "entailed" by the system's dynamics, meaning 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). Since free-energy is an upper bound on surprisal, continuously minimizing it will afford organisms a way to avoid surprisal-inducing states. Thus, an organism's prolonged existence can be understood as a continuous process of free-energy minimization.

Free-energy can be minimized in two ways: either by *perceptual inference*, which optimizes the recognition density so that free-energy becomes a tight bound on surprisal, or through *active inference* (i.e. self-generated changes of states), which avoids surprisal directly (see Bruineberg *et al*. 2018: 2413-2428 for further discussion of these points). Perceptual and active inference can be taken as corresponding to a form of perception and action[8] (e.g. 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, while making the agent learn the most efficient and minimalistic routes to success (e.g. Friston *et al*. 2016; Tschantz *et al*. 2020). In this way, the FEP makes contact

[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.

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).

Here is where Markov Blankets come into play. In statistics and machine learning, MBs are formal properties of variables 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). Within this literature, MBs are defined as follows:

> "**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)[10]

Here, the parents of a variable X are the variables whose directed connections lead immediately to X; whereas the children of a variable X are the other variables to which the X leads immediately through its directed connections; see **figure 1**.

[Insert figure 1 here]

Figure 1: The Markov Blanket of X. Nodes in the blanket are labelled to simplify the identification of the parents of X (X_P), the children of X (X_C) and the variables sharing a child with X (X_S), also known as the co-parents of X. All other nodes are unlabeled (Drawing by the Author)

The nodes constituting the Markov Blanket of X make it *conditionally independent* from any other node in the graph. This means that, in order to optimally estimate the value of X, one needs *only* to consider the values of the variables constituting its MB. Knowing (or ignoring) the value of any other variable will *not* modify the estimate. This is the reason as to why MBs can simplify the computation of the value of a variable: they allow us to "throw away" the rest of the graph X is embedded in when estimating its value. So, for instance, if X is

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

embedded in a graph with a hundred variables but its MB consists only of five variables, one can *safely ignore ninety-five variables* in the computation.

Now, the FEP takes MBs to be *also* real and objective boundaries of living systems (e.g. Friston 2013; Kirchhoff *et al*. 2018).[11] As Ramstead and colleagues (2019: 3) put it:

> "The Markov blankets are a result of the system's dynamics. In a sense, we are letting the biological systems carve out their own boundaries in applying this formalism. Hence, we are endorsing a dynamic and self-organising ontology of systemic boundaries"

The identification of MBs with the boundaries of living systems rests on the idea that although living systems need to interact with their environment because they are *open* systems, they must also distinguish themselves from their environments; that is, their states must form a set of states that is *distinct* from the set of environmental states (Palacios *et al*. 2020).

The FEP cashes in the relevant sense of organism/environment distinction in terms of *conditional independence* (Friston 2013; Palacios *et al*. 2020). The idea is that, once the state of the organism/environment boundary (i.e. the MB) is fixed, the goings-on on one side of the boundary will no longer influence the goings-on on the other side. When this happens: "all the necessary information for explaining the behavior of the internal states is given by the states of the blanket" (Hohwy 2019: 203). This form of conditional independence is precisely what MBs bring to the table: they "shield" the blanketed node (or, in the FEP rendition, organism) from the influence of any other node in the graph (or, in the FEP rendition, environment).

[11] This (as a reviewer noticed) might come as a bit of a shock for readers hostile to the "Pearl Blanket"/"Friston Blanket" conflation and for readers which are not familiar with the FEP. Both groups of readers are here reminded of *caveats #1* and *#2*. Sadly, space limitations prevent me from elaborating this point further. But see (Bruineberg *et al*. 2020: 16-20) for a clear, detailed and accessible discussion of this issue.

However, MBs also mediate the causal coupling between organism and environment.[12] This is because, according to the FEP, each MB is partitioned into two disjoint sets of states, termed *sensory* and *active* states (e.g. Friston 2013: 2). The partition is roughly as follows: a state of a MB is a sensory state if it is influenced by external states and influences internal (and active) states. Conversely, the state is an active state if 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; Kirchhoff and Kiverstein 2019a: 67). In this way, MBs allow an agent to couple sensomotorically with the environment, and allow to further formalize perceptual and active inference (e.g. Ramstead *et al*. 2018, fig. 1).

A prototypical example of a MB so conceived is that of a cell's membrane (Friston 2013; Da Costa *et al*. 2021; Millidge *et al*. 2021). The cell's membrane is a *functionally* relevant boundary which mediates the causal coupling between the cell's *internal states* (e.g. the states of the cytoplasm and organelles) and the *external states* (i.e. the environment the cell is embedded in) while still keeping the two separated *via* the conditional independence it induces (e.g. if the state of the membrane does not change, then internal states will remain fixed even if external states change).[13]

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.

---

[12] Notice that this point is not contested, and that it is granted even by vehicle internalists (e.g. Hohwy 2017).

[13] Albeit paradigmatic, the example of the cell's membrane needs some careful handling, for it might suggest that MBs *must*, in some sense, be *fences*; i.e. physical objects having contiguous spatiotemporal parts that demarcate the perimeter of a spatiotemporal region within which all the constituents of the free-energy minimizing system are located (see *fn.* 5). This is not the case: MBs can, *but need not*, be fences. MBs are *primarily* functional boundaries, described as a set of states making two other sets of states (termed "internal" and "external") conditionally independent. Whatever satisfies this description is a MB in the relevant sense, whether it is a fence or not (cfr. Kirchhoff *et al*. 2018: § 3.1).

So, how do MBs bear on the truth of vehicle externalism?

3 - Markov Blankets do not adjudicate whether vehicle externalism is true or not

According to the FEP, MBs are real and objective boundaries of free-energy minimizing systems, able to formalize perceptual and active inference. Given that perception and action intuitively are the interfaces separating the thinking machinery from the environment (cfr. Chalmers 2008; 2019), it is tempting to resort to MBs to determine whether the thinking machinery includes environmental and/or bodily constituents, thereby determining the truth of vehicle externalism.

But 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 living systems such as *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 the 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, assuming without argument that MBs demarcate the thinking machinery simply begs the question against vehicle externalism.

Perhaps this assumption could be justified by an argument *A* showing that the boundary of the organism is *also* the boundary of the thinking machinery. But then *A* would show that the thinking machinery is entirely contained within organisms, thereby proving that vehicle externalism is false, and leaving no role for MBs to play in adjudicating its truth.

It could be objected that I just misrepresented MBs, because MBs are *multiple and nested* (e.g. 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* (e.g. multicellular organisms), and thus possess their own MB. And in fact, FEP theorists sometimes 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), and eventually the entire biosphere (Rubin *et al*. 2020). Moreover, some of them claim that MBs are also *plastic*: their placement can vary over time, as new ways to sensomotorically engage with the environment are acquired (e.g. Clark 2017a). These shifts might lead Markov Blankets to move in a way such that their newfound placement includes organism-external components within the thinking machinery (Kirchhoff and Kiverstein 2019a, 2019b). If these points are correct, then MBs are *in no way forced to coincide* with the organism/environment boundary, and can therefore legitimately be used to determine whether the thinking machinery, or some other system, "extends" (e.g. Hohwy 2016; Ramstead *et al.* 2019).

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, it 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 of free-energy minimizing systems. They are said

to be "achieved by a system through active inference" (Ramstead *et al*. 2019: 11) and to "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 seems to be (at least partially) an *empirical* matter: 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). Further, MBs appear able to deliver the desired goods while *circumventing* the host of thorny philosophical issues that often have halted the debate over extended cognition, such as issues concerning non-derived content (see Piredda 2017 for a nice summary).

Yet, it seems to me that this usage of MBs raises at least two distinct problems.

First, the "multiple and nested" view of MBs smuggles a slightly different conception of MBs into the debate. For, in this conception, MBs are not (or not only) the boundaries of organisms or living things, but rather the boundaries of *biological systems* in general.[14] Perhaps extending the FEP in this way is the correct thing to do. Yet, once the FEP is extended in this way, it is no longer clear that perceptual and active inference correspond to perception and action (or anything thinking machinery-related). The entire biosphere may be a free-energy minimizing system (Rubin *et al.* 2020), but it is far from clear whether the biosphere as a whole *perceives* and *acts*.

[14] Or even the boundaries of *systems in general* (e.g. Hipolito 2019; Friston *et al*. 2020). Notice, however, that such a reading would transform the FEP from an account of biological self-organization to an account of *things in general*, in a way that it is likely to change the status of the FEP as a theoretical object (plausibly, an account of biological self-organization is a part of a special science, namely biology, whereas an account of things in general is not). I will not discuss this issue here, as stated by *caveat #1*.

Secondly, it is not clear whether letting MBs proliferate in this way would allow them to play the desired role in determining the truth of vehicle externalism. If MBs really are multiple and nested within each other, then 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, whether vehicle externalism is true would be determined *by C*, rather than the theoretical appeal to MBs (see also Clark 2017a: 8).

Importantly, the need for such a criterion seems to be acknowledged in the FEP literature. For instance, Ramstead *et al.* (2019: 25) argue that we can choose the relevant MB partially *depending on our explanatory interests*. 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. So, it seems that in the FEP literature I'm considering here, the need of a criterion to "pick up" the relevant MB is acknowledged, and that such a criterion is provided by our explanatory aims and interests.

However, I think that using such a criterion is problematic for two reasons.

First, if what counts as the relevant MB depends on our explanatory interest, then it becomes a bit unclear in what sense MBs are ontologically real and *objective* boundaries that are the result of a system's dynamics (e.g. Ramstead *et al*. 2019). On a fairly intuitive and innocent reading of "objective", something is objective if it is not mind-dependent. But surely explanatory interests are mind-dependent: for there to be explanatory interests, there needs to be minds around. So, if the MB of a system depends on explanatory interests, then it seems that MBs are not objective.[16]

---

[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] Notice that putting things this way does *not* entail that there is no fact of the matter on which is the relevant MB: there might still be a fact of the matter about what are the relevant (i.e. MB-determining) explanatory

Perhaps a way to respond to this challenge is to say that all the various (multiple and nested) MBs are really and objectively present in a mind-independent way. The idea would be that of claiming that the ontological structure of biological systems is fractal, and made up of MBs within MBs (cfr. Kirchhoff *et al*. 2018). Our explanatory interests would only *select* one of these objectively real MBs, *singling that one out* as the MB bounding the system we are interested in. If I understand them correctly, Ramstead and colleagues (2019) articulate and defend precisely such a position.

However, this position makes the *second* problem emerge perspicuously: the truth of vehicle externalism *does not* depend on our explanatory interests and/or our explanatory practices.

As illustrated in §1, vehicle externalism is a *metaphysical* thesis concerning the vehicles or constituents of our thinking machinery, which is independent from *epistemic* claims concerning how we should explain its functioning. This is well recognized in the literature over vehicle externalism (e.g. Sprevak 2010). On the one hand, the fact that vehicle externalism is a metaphysical claim distinguishes it from embedded/scaffolded views (e.g. Rupert 2009; Sterelny 2010), according to which satisfactory explanations of how the thinking machinery functions will make reference to extra-cerebral and/or extra-organismal factors *which are not constituents of the thinking machinery itself*. On the other hand, as noted in the first section of this paper, vehicle externalism makes no claim regarding how the thinking machinery functions. Vehicle externalism itself is compatible with different explanatory tools belonging to very different explanatory projects. Explanatory concerns are thus *orthogonal* to the truth of vehicle externalism.

---

interests. Yet, MBs would still not be objective in the sense of being mind-independent.

The very same point might perhaps most strikingly emerge considering what would happen given very *internalistic* explanatory interests. Surely the fact that one's explanatory interests concern (for instance) just the hippocampus does not entail that the thinking machinery is the hippocampus *and only the hippocampus* (cfr. Clark 2008: 109-110). Hence, Externalist (or internalist) explanations and/or explanatory interests favoring "wider" (or "smaller") MBs do not entitle one to the conclusion that vehicle externalism (or internalism) is true.[17]

Now, I wish to point out that there is a sense in which, when it comes to determining whether vehicle externalism holds true, it is *irrelevant* whether MBs are boundaries of organisms rather than multiple and nested. This is because we should be skeptical of the very idea that MBs can be used to *identify* systems (thinking machinery included). The reason is simple: the identification of a system (i.e. of a variable or set of variables of interest) is *logically prior* to the identification of its MB. If this is correct, then we are simply not allowed to use MBs to identify systems, on the pain of circularity.

Notice that this very issue has repeatedly surfaced throughout this section. When it comes to adjudicating the truth of vehicle externalism *via* MBs, assuming that MBs "enshroud" organisms is problematic precisely because it *presupposes* that the thinking machinery coincides with the insides of organisms, thereby begging the question against vehicle externalism. And when it came to "choosing" the right MB in a sea of multiple and nested MBs, the same problem reappeared: our explanatory interests, presumably oriented towards a previously identified system (or behavior/phenomenon exhibited by a system), were in fact

[17] Reflecting on mental content yields similar results. Most contemporary theories of mental content endorse semantic externalism, claiming that contents are partially determined by environmental factors (e.g. Shea 2018). So they accept that the extra-organismal environment plays a role in explaining how the thinking machinery works; namely, the role of (partially) determining mental contents. But these theories also typically take vehicles to be internal to the system in which they are tokened. Indeed, that internalism/externalism about content and vehicle are orthogonal is a fairly uncontested fact (see Clark and Chalmers 1998; Hurley 2010).

needed to single out the relevant MB. In both cases, we started with a system *and then* "discovered" the MB *of that specific system*.

To see why the identification of a system logically precedes the identification of its MB, recall how MBs are defined in the relevant literature on graphical models:

> "**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.[18] 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: one cannot just point to a graph and ask: "Ok, now tell me where is the relevant Markov Blanket". To ask so, one *must* have already indicated which is the node whose MB one is interested in. The identification of the "blanketed" system *is logically prior* to the identification of its MB. The direction of identification runs from target variables to MBs, and not the other way around.

Notice that the same order of individuation is preserved in empirical (or semi-empirical) settings. Consider, for instance, the simulation presented in (Friston 2013). Without entering too much in the detail, the simulation aims to show that a "protocell" equipped with a MB will spontaneously emerge from a "primordial soup" of particles interacting through short-range physical forces. To do so, the "primordial soup" is simulated and the particles are left to interact for some time. Then, the eight most densely coupled particles are *identified as the internal states* (Friston 2013:6), and their MB is recovered and splitted into active and sensory states (depending on whether the states constituting it influenced or were influenced

[18] Alternatively, one could "collapse" all the variables describing a system in the macro-variable "state of the system".

by the internal states). So, it seems that even in the empirical (or semi-empirical) setting of this simulation, the direction of identification runs from free-energy minimizing systems to MB.[19]

Time to take stocks. If MBs are the boundaries of organisms, then using MBs to determine whether vehicle externalism is true simply begs the question against vehicle externalism. If MBs are *not* boundaries of organisms because they are multiple, nested, malleable and plastic, then using MBs to adjudicate the truth of vehicle externalism does not beg the question against it - but invites other problems. The first is that it provides a slightly different conception of MBs, in which perceptual and active inference cannot be *obviously* equated to perception and action. The second is that if MBs are multiple and nested, then we need a criterion to identify which is the MB of the thinking machinery; and it would be that criterion, rather than the presence of a MB, what adjudicates the truth of vehicle externalism. Moreover, the criterion currently in use in the FEP literature is problematic, as it casts more than a shadow of doubt on the objectivity of MBs and it is ultimately unsuited to adjudicate the truth of vehicle externalism. Lastly, there are reasons to be skeptical of the whole idea of identifying systems *through* or *by means of* their MBs. This is because, logically, the identification of a MB *presupposes* the previous identification of a relevant system (i.e. a variable or set of variables). Using MBs to identify systems would thus be obviously circular.

4 - Markov Blankets do not track the boundaries of the mind

[19] In more recent versions of the FEP, however, this is not necessarily true. Thus, for instance, although (Hipolito *et al*. 2021) use MBs to partition the nervous system in previously known sub-systems (such as neurons and canonical microcircuits), (Friston *et al*. 2021) try to "read" MBs directly out of the couplings of various neuronal components. Yet, their procedure seems very removed from the graph-theoretic apparatus from which MBs originated. Moreover, MBs identified through this procedure still seem to be multiple and nested in a way that invites all the problems discussed above in regard to multiple and nested MBs. At any rate, these versions of the FEP do not share the assumptions here made via *caveats* #1 and #2, and so I will not discuss them further.

In the paragraph above, I've put forth some reasons to think that resorting to MBs will not determine whether vehicle externalism is true or not. But perhaps it could be objected that I have misunderstood the whole endeavor, and misinterpreted what MBs are supposed to do in that debate. Maybe MBs are not intended to *directly determine* the truth-value of vehicle externalism. Maybe they are just *framing tools*: conceptual devices that help us, in some determinate way, to adjudicate whether vehicle externalism is true. Here's a clear statement of the idea:

> "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)

Notice how, in this quote, Kirchhoff and Kiverstein are presenting MBs as a *formal tool* with significant epistemic virtues: it avoids begging the question in the debate over vehicle externalism and escapes some thorny issue that have plagued that debate. Perhaps this is the correct way to think about the role MBs should play in the debate over vehicle externalism. Maybe asking "where can we draw a MB around the thinking machinery?" yields more satisfactory results than trying to find a "mark of the cognitive" (e.g. Adams and Aizawa 2008) or another way to tell apart external propst that *causally interact* with the thinking machinery from the ones *constituting* it (e.g. Kaplan 2012). Since the "classic" debate over vehicle externalism ended up in a stalemate (cfr. Adams 2019 and the reply by Clark 2019), new ways to tackle the debate are surely welcome.

Yet, as far as I can see, the idea of using MBs as formal tools to settle the debate over vehicle externalism is far from unproblematic. In my assessment, it suffers from two distinct problems.

The first concerns the ontological status of MBs. In the literature on the FEP I'm considering, a cell's membrane is often offered as the prototypical example of a MB (Friston 2013; Kirchhoff *et al*. 2018; De Costa *et al*. 2021). But, *prima facie*, cell membranes are not framing devices or formal tools: they are concrete objects.[20] Moreover, MBs are supposed to be the result of a system free-energy minimizing dynamics (Ramstead *et al*. 2019). It is hard to see how a system's free-energy minimizing activity could result in a *formal tool* or "an attractive statistical framework".[21]

Now, perhaps the concern above could be allayed just by saying that *talking* about MBs (i.e. framing the issue of vehicle externalism in terms of MBs) is a good way of *tracking* MBs (i.e. objective boundaries of free-energy minimizing systems, among which the thinking machinery). The idea would thus be that the MBs *talk* tracks the objective boundaries of systems, or that it is at least the best way currently at our disposal to track and identify the objective boundaries of the thinking machinery (which also happen to be called "Markov Blankets", cfr. Palacios *et al*. 2020: 6). This strikes me as a reasonable and charitable interpretation of the passage by Kirchhoff and Kiverstein cited above.

Yet, and this is the second concern, there seems no *prior* guarantee that MBs will track real and objective boundaries of free-energy minimizing systems.[22] Consider the variables implicated in some psychological explanations. The occurrence of depression, for instance, is correlated with a range of variables such as *being divorced*, *being jobless*, *having being*

[20] Of course, this worry is closely linked to worries about the FEP's status as a theoretical object and its ontological commitments, as well as the distinction between "Pearl Blankets" and "Friston Blankets" (Bruineberg *et al*. 2020; see also Menary and Gillett 2020). But, as amply clarified when making *caveats #1* and *#2*, I'm here assuming that the version of the FEP I'm considering gets both of them right. And, to restate, the version of the FEP I'm considering takes MBs as formal tools and MBs as real boundaries of systems to be identical.

[21] Notice that the fact that a cell's membrane is a *fence* (see *fn.* 5 and 13) is playing no role in the argument I just gave. What is playing a role in my argument is that MBs are supposed to be boundaries objectively "out there" in the real world, rather than formal tools pertaining to a statistical framework. And, as clarified above, MBs need not be *fences* to be objective boundaries "out there".

[22] I owe this observation (and the example) to ***Anonymized for blind review***.

humiliated (I'm taking this example from Campbell 2007). These variables might figure in a graph depicting the state of a subject. It is thus possible that they might end up constituting the MB surrounding the subject's thinking machinery. Suppose it happens. Then, if the formal tool provided by MB tracks the real and objective boundaries of the thinking machinery, it would follow that *being divorced* or *being jobless* are part of the objective boundary that functionally separates the subject's thinking machinery for the environment, which is established by the free-energy minimizing activity of the thinking machinery itself. I must confess that I find this claim simply unintelligible. And yet it is a claim that *could* be licensed by the assumption that MBs track the objective boundaries of systems. Generalizing from this example, the problem seems to be this: given a target variable (or set of variables) in a graph, the MB that the target variable (or set of variables) identifies may be composed of nodes that track things or states of affairs that might not constitute an *ontologically real and objective boundary* in any straightforward sense of the term.[23]

It could be objected that although such "weird" boundaries *could* be identified, nothing entails that they *will*. The fact that we have no prior guarantees that Markov Blankets will track the real and objective boundaries of the thinking machinery clearly does not entail that they *won't* track it. Perhaps, as a matter of contingent fact, they will. The proof is in the pudding.

However, observing how MBs have been used strongly suggests that they do not *in fact* track the objective and real boundaries of the thinking machinery. To be fair, I must state here that it has not forced us to say that *being jobless* is part of the boundary separating the thinking machinery from the rest of the world (at least, not yet). Nevertheless, MBs seem to

[23] Notice, for the sake of clarity, that the problem I'm raising here is *not* that such a blanket would not be a *fence* (see fn. 5 and 13). Nor the problem that I'm here raising is that variables such as "having been humiliated" do not map onto spatiotemporal parts of *fences*. The problem I'm raising is that such variables do not seem to map onto *any* functional boundary constituting a thinking machinery/world interface.

misplace such a boundary in a way significant 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) defense of vehicle externalism.[24]

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 to track the bounds of 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 thinking machinery 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*.[25] Here's Hohwy spelling it out:

---

[24] 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).

[25] Notice that this criterion identifies the relevant MB by what it bounds; namely, the physical machinery that performs free-energy minimization on average and in the long run. Hence it is fully consistent with the arguments provided in the end of the preceding section of this paper.

> "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 over time*. [...] 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 "over time, on average and in the long run" criterion is intrinsic in the structure of the FEP. The FEP is an account of how organisms/biological systems persist *over time*. 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, over time, seek the evidence confirming their existence, thereby prolonging it (cfr. Hohwy 2016).

Lastly, and most crucially for present purposes, both accounts take the "over time, on average and in the long run" criterion to be *extensionally adequate*; that is, apt to single out the MBs tracking the boundary 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 thinking machinery 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 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 seems to suggest that framing the debate over vehicle externalism in terms of MBs does not, in and by itself, simplify our tracking the boundaries of the thinking machinery. Now, one might object that framing that debate in terms of MBs is not supposed, in and of itself, to simplify our tracking.

Fair enough; but then, *why bother with MBs*? 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.

To see why, recall that, during active inference, an agent "brings about" the sensory states it expects to encounter. Importantly, these states encompass the variables that define the state-space of all of an organism's possible states. For us humans (and, broadly speaking, animals) this includes extero-, intero- and viscero-ceptive states (e.g. Seth and Friston 2016). Hence, us humans (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.* 2018: 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 minimize free-energy and avoid surprisal. We do so, for instance, by sweating, so as to lower our bodily temperature when it is

too high; 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 over time, 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 tracking the bounds of the thinking machinery will include clothes in the internal states. And this conclusion surely seems wrong: pretty much everyone agrees that the constituents of our thinking machinery are supposed to *do something with information*, either by processing and/or storing it (as cognitivists contend), by "resonating" with it (as gibsonians contend, see Raja 2018) or by responding to it and/or enabling an agent's response to it (as enactivists contend, see Hutto and Myin 2013). But clothes do not appear to do *anything* with information. 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.

Notice that the argument I've just given does not depend on a very *demanding* "benchmark" to adjudicate whether something counts as a constituent of the thinking machinery (cfr. Wheeler 2011).[26] As Wheeler notices, when determining whether a candidate constituent *really* qualifies as a constituent of the thinking machinery, we *need* to have some benchmark to determine whether the constituent contributes to *thinking* (in the broadest possible sense) as opposed to anything else - otherwise, every candidate constituent would be counted in by default! Traditionally, this benchmark is provided by the *mark of the cognitive* one endorses; that is, by what one (implicitly or implicitly) takes to be necessary and/or sufficient[27] to make something a *genuine* contributor to thinking (in the broadest possible

[26] I wish to thank a reviewer for having pressed me to make this point more explicit.
[27] In the literature, sets of either necessary (e.g. Adams and Aizawa 2008) or sufficient (Rowlands 2009)

sense).[28] Here, the "mark of the cognitive" I'm endorsing is not very demanding and does not rest on contentious assumptions on the nature of cognition (indeed, as highlighted above, ecological psychologists, enactivists and cognitivists can all easily endorse it).[29] In the present context, this is a virtue: it makes my "benchmarking" fairly uncontroversial, thereby making this "mark of the cognitive" very hard to reject, in a way that makes it hard to reject my conclusion *by rejecting* the "mark of the cognitive" on which it rests.[30]

A reviewer (whom I thank) noticed that the example provided above can be countered in this way: clothes keeps our free-energy low on average and in the long run *only considered as a type*, but no *token* piece of clothing is involved in free-energy minimization on average and in the long run - we change clothes far to often for that to be the case. Hence no token piece of clothing should be included in the thinking machinery. This remark is surely correct. And yet, the example can be easily modified so as to force the inclusion of *token* pieces of clothing in the thinking machinery. We can easily imagine a futuristic society in which clothes are

---

conditions have been proposed - but no set of necessary and sufficient conditions. I think this disparity depends on one's argumentative goals: defenders of vehicle externalism have only to show that some external component *really* qualifies as a constituent of the thinking machinery, and to do so they only need *sufficient* conditions. Conversely, vehicle internalists need to argue that no candidate constituent *really* qualifies as a constituent: hence, they typically need to show us that all plausible candidate constituents violate some *necessary* condition.

[28] For the sake of completeness, notice that, strictly speaking, a mark of the cognitive may not be *necessary* to determine what counts as a cog in the thinking machinery. For example, Kaplan (2012) has proposed a *mutual manipulability* criterion to do that, and that criterion does not qualify as a mark of the cognitive. Notice, however, that if one were to endorse Kaplan's criterion to identify the various bits and pieces of the cognitive machinery, one would not have any use for MBs in determining the boundaries of the mind.

[29] Notice the scare quotes: I do not mean to suggest that "doing something with information" is the real mark of the cognitive. I'm only *using* it as a mark of the cognitive for argumentative purposes. And this to me seems fairly justified given that most philosophers and cognitive scientists would take "doing something with information" to be at least *necessary* in order for something to qualify as a cog in a cognitive machinery. Notice further that the "mark of the cognitive" I'm here adopting makes no distinction between *mere sensing* and *real thought*. So one cannot object to my analysis that it begs the question against alleged instances of "extended sensing", such as the usage of sensory substitution devices ranging from tactile-visual substitution devices (cfr. Noe 2004) to the proverbial stick of a blind person (Merleau-Ponty 2013).

[30] And even if someone were to take issue with this "mark of the cognitive", I could still do without it by appealing to our *folksy intuitions* to substantiate my conclusion: I'm fairly sure no one has the intuition that clothes are part of our thinking machinery. Notice that such an appeal to intuition would not be something groundbreaking in the debate over vehicle externalism: indeed, it is what (Clark 2008) recommends. Notice further that for such an appeal to intuition to work I neither need to presuppose that our intuitions are *always* crisp and clear, nor I need to presuppose that such intuitions are universally shared or indefeasible. There surely are cases in which our intuitions on cognition are murky, defeasible and not universally shared (e.g. do bacteria cognize?, see Lyons 2015). But the case at hand does not seem one of such cases.

made out using a super resistant, self-cleaning fiber, and so everyone wears a single outfit for the entirety of one's life. And even letting aside sci-fi scenarios, our *livers* do contribute to our thermoregulation. And most of us have only a single token liver through their lifetime. Should we conclude our livers are cogs in our thinking machinery? I'd answer in the negative, adducing the same reasons I adduced to claim that clothes are not cogs in our thinking machinery. Notice further that counterexamples of this sort proliferate easily. We typically have a single token pair of lungs and kidneys, a single stomach, a single intestine, a single set of blood vessels, a single hearth, and so forth. All these organs and body parts perform a number of functions that keep our free-energy low. And they perform these functions on average and in the long run. But, plausibly, none of these organs counts as a cog in our thinking machinery.

A (different) reviewer objected that the original clothes example rests on a philosophical sleight of hand. In their view, I build up the scenario using clothes, whereas I should use "knowing to take action to put on/take off/change clothes" (*verbatim* quote). Clothes, the reviewer seems to suggest, only contribute to a subject's sensory states. But sensory states are fleeting. This, the reviewer suggests, makes the case I proposed significantly different from paradigmatic cases of extended cognition, such as Otto's usage of a notebook to remember a relevant piece of information (Clark and Chalmers 1998). In this case, memory is *not* treated as something fleeting, and it is its persistence that makes Otto's perceptuomotor access to the notebook count as a *bona fide* instance of extended cognition. What could be said in response?

I'm not sure, mainly because I'm not sure I understand what the reviewer is after.[31] I have a

[31] For the record, this means that I could have grossly misinterpreted the reviewer's point, and thus that the paragraph above might be a gross misrepresentation of the reviewer's actual position. If that is the case, I apologize: it is not my intention to misrepresent the reviewer's view. But that is what I've understood, and so I can only respond to that.

hard time seeing why, in the example above, clothes should be substituted by one's knowledge about which clothes one should wear. I'm willing to concede that the vehicle storing that piece of knowledge is a *bona fide* cog in a subject's thinking machinery, and I'm willing to concede that it plays a role in keeping one's free-energy low on average and in the long run. Trivially, if one thinks that a good way to resist cold temperatures is by getting naked, one's free-energy will increase. But surely that piece of knowledge alone is not *sufficient* to keep one's free-energy low. I might know that, given the cold temperature, I would be better off wearing a sweater. But if I have no sweater to wear, I *will* get cold, thereby failing to efficiently minimize my free-energy.[32] So, there seems to be nothing problematic in taking clothes to be parts of the physical machinery by means of which free-energy is minimized; hence, *at least in this regard*, I've performed no sleight of hand. And this seems all that is needed in order for my original example to work.[33]

Now, back to the main argument. I want to make a further claim. I want to argue that even if we set aside (for the sake of discussion) matters of extensional adequacy, we have a further reason *not* to endorse that "on average, in the long run" criterion. For it seems that when it comes to *internal* (i.e. neural) vehicles, we do not judge whether they qualify as constituents of the thinking machinery based on their role in free-energy minimization *on average and in the long run*.[34] Hence, that criterion violates the core insight that the "parity principle" is

[32] Bruineberg *et al*. (2018a: 2430-2432) make a very similar point.

[33] Moreover, I must confess that I find it hard to see *why* the fleetiness of sensory states (as opposed to the persistence of memory) might cause troubles here. Although Otto's case is (perhaps regrettably) one of the paradigmatic cases of extended cognition, and although in that case surely what "extends" (if anything) is a perduring dispositional state, vehicle externalism is in no way a claim whose scope is limited to perduring states. Indeed, the first case of cognitive extension proposed in (Clark and Chalmers 1998) is a case of extended mental rotation, entirely built upon the usage of fleeting sensory states.

[34] A reviewer noticed that the original formulation formulation of the parity principle embeds a temporal dimension: "if, as we confront some task, part of the world functions as a process which, were it to go on in the head, we would have no hesitation in accepting as part of the cognitive process, then that part of the world is *(for that time)* part of the cognitive process (Clark and Chalmers 1998: 8, emphasis added). The reviewer's comment is surely welcome, for it *reinforces my point*: if we follow the parity principle, we judge candidate constituents of the thinking machinery by how they contribute to cognitive processing *when* they contribute, rather than by the *overall duration* of their contribution. Hence a criterion based on "average, in the long run" contribution to

trying to express; namely, that we should judge whether candidate external vehicles are constituents of our thinking machinery *with the same metric* we deploy to judge internal vehicles (Clark 2008; 77-78; 2013: 195).[35]

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.[36] That is, at time *t-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

---

free-energy minimization is surely at odds with the parity principle. And that is what I'm claiming.

[35] Vehicle externalists that emphasize the *complementarity* of inner and outer resources (e.g. Menary 2007; 2018) find the parity principle problematic, as it might suggest that internal and external resources must be functionally similar. Yet,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 (cfr. Menary 2007: 55-57; Gallagher 2018).

[36] 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 Clark 2010: 53-54; 2019: 277); at least, until a suitably uncontested "mark of the mental/cognitive" is provided. But notice that the minimal "mark of the cognitive" deployed above licenses this conclusion too.

the long run. Consider, for instance, synaptic pruning. According to the FEP, such a process should be understood in terms of a reduction of 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 physically "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 "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*[37], then the same must hold for putative external vehicles. Hence, given that we would not apply the "over time, on average and in the long run" criterion to pieces of the brain, we should not

[37] 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.

apply it to putative external vehicles. And since (at least intuitively) the antecedent is correct, the consequent follows.

A reviewer suggested that the line of reasoning proposed above might be tainted by a conceptual confusion; that is, a confusion between *supporting* the existence of a free-energy minimizing system in the long run and *being part of* or *constituting* a free-energy minimizing system in the long run. The example the reviewer provided is the following: a neurotransmitter token (say, a particular serotonin molecule) can *support* the continued existence of a free-energy minimizing system without *thereby* being part of the system's continued existence: given neurotransmitter decay, that particular molecule will not be part of the system's future states. Now, if what matters is *just* supporting the continued existence of a system, then my thought experiment on child neural explant (and the subsequent points on synaptic pruning, disconnection syndromes and "transient" task-specific neuronal devices) would not be warranted.

While it is true that these points would not be warranted if what matters is just *supporting* a free-energy minimizing system continued existence, it is doubtful that what matters is *just* supporting. Conceptually, were just supporting an organism's continued free-energy minimization *sufficient* for being part of the thinking machinery, then all sorts of things would count as constituents of that machinery. For example, if I'm on fire and jump in a pond of water to put off the flames, that water is transiently *supporting* my continued existence. But it seems wrong to say that ponds of water are constituents of my thinking machinery, for the same reason it seems wrong to say that clothes are constituents of my thinking machinery: they do not do anything with information. Moreover, as a matter of interpretation, both Hohwy and Kirchhoff and Kiverstein seem to agree that *just transiently supporting* a free-energy minimizing system's prolonged existence is not enough:

> "It is crucial that this minimization happens on average and in the long run because the surprise that is sought minimized is defined in terms of the states the creature tends to occupy in the long run [...]. *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: 271 emphasis added)

> "The action of using the notebook is a part of how Otto succeeds in minimising expected free-energy [...] *Crucially, his use of the notebook is not simply a one-off action. It is part of how Otto minimises expected free-energy, on average and over time.*" (Kirchhoff and Kiverstein 2019b: 17-18; emphasis added)

So, it seems correct to say that the point I just raised does not misinterpret Kirchhoff, Kiverstein or Hohwy's thoughts on the matter.

Now, it is natural to wonder whether the "over time, on average and in the long run" criterion to identify the MB around the thinking machinery could be substituted by a better criterion. However, the "over time, on average and in the long run" criterion is taken to directly "fall off" out of the FEP. And, in 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 over time, on average and 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 "over time, on average and in the long run" criterion *without* thereby introducing substantial modifications in the theoretical architecture of the FEP itself.

Perhaps the "over time, 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 way. But that seems like an admission of defeat: such a criterion would in fact be in the task of *correcting* the verdicts yielded by the "over time, on average and in the long run" criterion, which strongly suggest that the "over time, on average and in the long run", in spite of being intrinsic to the FEP, is not up to the task.

But perhaps I've thus far dramatically misunderstood what "on average, in the long run" means; or so, at least, a reviewer contends. They argue that when, in the infantile neurosurgery example, I wrote "[...]by stipulation, since 'the owner' of *x* [NA:the neurosurgically removed region] is a child, she spends most of her life without *x*" the phrase "most of her life" was exactly what it is *meant* by the expression "on average, in the long run". The reviewer further argues that getting the meaning of that expression right is crucial for my argument, given that my entire argument turns on a distinction between statistical and physical boundaries. To help make this distinction clear, the reviewer proposes the following example: if one colours inside the lines *on average and in the long run*, one might be actually coloring outside the lines at any point in time. But, as the appropriate frequencies are taken into account, even the act of coloring outside the lines is part of one's coloring inside the line on average and in the long run. How could I respond?

To start, I wish to note that my argument *does not* turn out to depend on a distinction between statistical and physical boundaries. I indicated this explicitly in §1. In that section, I've explicitly stated that, for the purposes of the present argument, I was not going to distinguish between "Pearl Blankets", that is, Markov Blankets intended as formal properties of variables, and "Friston Blankets", that is, Markov Blankets intended as ontologically real boundaries of a system. This is also why, in the same section, I've explicitly stated that my claim here is conditional: it is conditional because I'm willingly not distinguishing the two (as commonly done in the literature) *for the sake of argument*.[38]

[38] Perhaps there is a sense in which my argument presupposes a distinction between statistical and physical boundaries, *if* by "physical boundaries" one means what I have here been indicating with the term "*fences*"; that is, a physical object having contiguous spatiotemporal parts which demarcate the perimeter of the spatiotemporal region within which all and only the constituents of the free-energy minimizing system are located (see *fn.* 5). But surely distinguishing statistical boundaries such as MBs from *fences* is not problematic, given that MBs are not supposed to be fences (see *fn. 13*). Distinguishing the boundaries of the thinking machinery from fences is similarly unproblematic, given that no one takes such boundaries to be fences. Otherwise, it would be fairly easy to argue against vehicle externalism: it would be sufficient to notice that no "extended fence" exists!

Let me now focus on the example of coloring within the lines. If I interpret it correctly, the example suggests that a process (coloring within the lines) going on on average and in the long run *need not* be constituted (or otherwise made up) by spatiotemporal parts occurring (or otherwise present) on average and in the long run (if the coloring outside the lines were to occur on average and in the long run, then arguably one wouldn't be coloring inside the lines on average and in the long run).

If the example is meant to convey this, then the reviewer is raising a point similar to the point examined above when I contrasted *supporting* and *being part of* the continued existence of a free-energy minimizing system. Lots of things (like jumping into ponds of water to put off flames) can be transient parts of the process of minimizing one's free-energy on average and in the long run. It is even possible to conceive realistic scenarios in which one's deliberate departure from low-surprisal states is part of one's *in the long run* free-energy minimization (e.g. skipping breakfast to take a blood test). But surely ponds of water and skipped breakfasts are not cogs in the cognitive machinery - they do not do anything with information.

One could perhaps contend these uncomfortable conclusions seemingly follow only because I've not changed my interpretation of the "on average, in the long run" phrase. But how should it be interpreted?

The reviewer suggests that "on average, in the long run" means roughly "most of a system's lifetime". *But this is how the phrase has been interpreted above*. Indeed, the infantile neurosurgery case works precisely *because* some extremely plausible cogs of the child's thinking machinery *are not there for most of her life*, and so, given Kirchhoff, Kiverstein and Hohwy's usage of MBs to demarcate the boundaries of the thinking machinery, we are pushed towards the (seemingly unwarranted) conclusion that these very plausible cogs (recall, in the

examples they are pieces of neural tissue!) are not cogs at all. Hence, it seems that all my points/counterexamples are left in good order by such a reading.

Perhaps it could be argued that the expression "on average, in the long run" names the system's *phenotype*; that is, the set of low-surprisal states that according to the FEP an organism must visit *on average and in the long run* in order to prolong its existence. If that were the case, the claim made by Hohwy, Kirchhoff and Kiverstein would be that something counts as a constituent in a subject's thinking machinery just in case it contributes to the organism's occupying the phenotypic states.

But this does not seem what they want to claim (see their citations above). They manifestly do *not* wish to call a constituent of the thinking machinery everything that contributes to a system's persistent occupation of its phenotypic states. Otherwise, why shouldn't Hohwy allow *guns* used to shoot tigers (his example) to count? And why would Kirchhoff and Kiverstein stress the fact that Otto's usage of his notebook is not a one-off action? Surely the one-off action of shooting a tiger to avoid being mauled to death does contribute towards one occupying one's phenotypic states.

One might perhaps contend further that Kirchhoff, Kiverstein and Hohwy are simply misguided, and that the reading above is the one they *should* have endorsed. I really do not see how such a view could be defended. After all, Kirhhoff, Kiverstein and Hohwy do *not* endorse that reading precisely because they realize endorsing it would force one to count an inordinate amount of stuff as a cog in someone's thinking machinery: jumping into ponds of water while on fire or shooting at a tiger to avoid becoming the tiger's dinner both contribute to one's prolonged occupation of one's phenotypic states, but neither ponds of water nor guns and bullets can be properly counted as cogs in the thinking machinery (they do not appear to do *anything* with information). Moreover, there can be very plausible cogs in one's thinking

machinery that do *not* contribute to one's continuous occupation of one's phenotypic states. Think about the patterns of neural activity that instantiate a person's suicidal (or otherwise self-harming) tendencies.

In summary, it seems to me entirely correct to conclude that such an alternative reading of the phrase "on average, in the long run" is not supported by textual evidence, and it is not able to solve the relevant issue at hand. Hence, it should be rejected. Notice, importantly, that nothing of what I've just said entails or suggests that the reading of "on average, in the long run" deployed in my main argument is *the correct* reading. Nor am I entailing or suggesting that it is the *only possible or coherent reading*. Other readings might be both possible and more apt. But, at present, I really am unable to see any such alternative reading. So, I'm happy to throw the ball in the other camp, challenging philosophers convinced that MBs do a good job at tracking the boundaries of the mind to spell out, in a clear manner, such an alternative reading.

Time to take stocks. In this section, I have argued that considering MBs as formal tools to identify the boundaries of the thinking machinery raises a puzzle on the metaphysical status of MBs. Even ignoring that puzzle, considering MBs as formal tools to track the boundaries of the thinking machinery does not guarantee us that MBs will identify boundaries in any relevant sense, and indeed the concrete application of such a tool has thus far yielded very unpalatable results.

This strongly suggests that MBs are not good formal tools to track the boundaries of the thinking machinery. In the next section I will further expand on this issue, suggesting that resorting to MBs forces us to *sidestep* the dispute over vehicle externalism in a very important sense.

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

In this section, I want to argue that resorting to MBs to settle the debate over vehicle externalism leads us to sidestep the whole debate in a very real sense, making vehicle internalism *vacuously* true. 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). Hohwy spells it the most clearly:

> "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 determined 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 Jakob 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 what is 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 denying 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 drawing 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 false by definition. But the point of vehicle internalists has *never*

been that vehicle externalism is false by definition - rather, their point is that vehicle externalism is false *as a matter of contingent empirical fact* (cfr. Adams and Aizawa 2008). The truth of vehicle externalism should thus *at least in part* depend on how the world factually is, and souldn't be entirely settled by the meaning of words. Accepting that the dispute over vehicle externalism is solved by a re-definition of "internalism" and "externalism" seems a significant change of topic.

Moreover, I doubt such a redefinition of "internalism" would buy the internalist something more than a purely *verbal* victory. For 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 abandoning *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.

Will FEP theorists be able to meet this challenge? Of course, only time will tell. But, on my assessment, the prospects for the FEP theorists are not rosy. For, as signaled when placing *caveats #1* and *#2*, here I have adopted the most charitable reading of the FEP and of its ontological commitments (at least when it comes to adjudicating the truth of vehicle externalism *via* MBs). So it seems to me correct to conclude that FEP theorists eager to meet my challenge will have to fight an uphill battle: they will both have to rebuke my arguments, *and* to persuade others (e.g. Bruineberg *et al*. 2020; Menary and Gillett 2020) that the conception of MBs they deploy is indeed the right one.

References

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

Adams, F. (2019). The elusive extended mind: extended information processing doesn't equal extended mind. In M. Colombo, E. Irvine, M. Stapleton (Eds.), *Andy Clark and His Critics* (pp. 21-32). New York: Oxford University Press.

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.

Andrews, M. (2021). The math is not the territory. *Biology and Philosophy*, *36*(3), 1-19.

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*. (2018). 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.* (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.

Chalmers, D. (2008). Foreword. In A. Clark. (Auth.), *Supersizing the Mind* (pp. IX-XVI). New York: Oxford University Press

Chalmers, D. (2019). Extended cognition and extended consciousness. In M. Colombo, E. Irvine, M. Stapleton (Eds.), *Andy Clark and His Critics* (pp. 9-20). 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). Extended active inference: constructing predictive cognition beyond skulls. *Mind and Language*, https://doi.org/10.1111/mila.12330

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.

Da Costa, L. *et al.* (2021). Bayesian mechanics for stationary processes. *Preprint*??, https://arxiv.org/abs/2106.13830 last accessed 26/07/2021

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.

Fabry, R. E. (2021). Limiting the explanatory scope of extended active inference: the implications of a causal pattern analysis of selective niche construction, developmental niche construction, and organism-niche coordination dynamics. *Biology and Philosophy*, *36*(1), 1-26.

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.

Friston, K., *et al*. (2021). Parcels and particles: Markov blankets in the brain. *Network Neuroscience*, *5*(1), 211-251.

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.

Hipolito, I., *et al*. (2021). Markov Blankets in the brain. *Neuroscience and Biobehavioral Reviews*, *125*, 88-97, https://doi.org/10.1016/j.neubiorev.2021.02.003

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. (2019). Quick'n'Lean or Slow and Rich? Andy Clark on predictive processing and embodied cognition. In M. Colombo, E. Irvine, M. Stapleton (Eds.) *Andy Clark and His Critics* (pp. 191-205), New York: Oxford University Press.

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.

Hutto, D., & Myin, E. (2013). *Radicalizing Enactivism*. 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.

Kiverstein, J., & Sims, M. (2021). Is free-energy minimization the mark of the cognitive?.*Biology and Philosophy*, *36*(2), 1-27.

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.

Lyon, P. (2015). The cognitive cell: bacterial behavior reconsidered. *Frontiers in Microbiology*, *6*: 264.

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.

Merleau-Ponty, M. (2013). *Phenomenology of Perception*. New York: Routledge.

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

Millidge, B., *et al*. (2021). A mathematical walkthrough and discussion of the free-energy principle. *Preprint*, ArXiv preprint arXiv:2108.13343

Noe, A. (2004). *Perception in Action*. Cambridge, MA.: The MIT Press.

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.

Raja, V. (2018). A theory of resonance: towards an ecological cognitive architecture. *Minds and Machines*, *28*(1), 29-51.

Raja, V., *et al.* (2021). The Markov Blanket trick: On the scope of the Free-energy principle and Active Inference. *Physics of Life Review*. https://doi.org/10.1016/j.plrev.2021.09.001

Rowlands, M. (2009). Extended cognition and the mark of the cognitive. *Philosophical Psychology*, *22*(1), 1-19.

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

Shea, N. (2018). *Representation in Cognitive Science*. New York: Oxford University Press.

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

Sprevak, M. (2010). Inference to the hypothesis of extended cognition. *Studies in History and Philosophy of Science Part A*, *41*(4), 353-362.

Sterelny, 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

Wheeler, M. (2011). In search of clarity about parity. *Philosophical Studies*, *152*(3), 417-425.

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

Wiese, W., & Friston, K. (2021). Examining the continuity between life and mind: is there a continuity between autopoietic intentionality and representationality?, *Philosophies*, *6*(1): 18.

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.