The Problem of Mental Action

Predictive Control without Sensory Sheets

Thomas Metzinger

In mental action there is no motor output to be controlled and no sensory input vector that could be manipulated by bodily movement. It is therefore unclear whether this specific target phenomenon can be accommodated under the predictive processing framework at all, or if the concept of "active inference" can be adapted to this highly relevant explanatory domain. This contribution puts the phenomenon of mental action into explicit focus by introducing a set of novel conceptual instruments and developing a first positive model, concentrating on *epistemic* mental actions and epistemic self-control. Action initiation is a functionally adequate form of self-deception; mental actions are a specific form of predictive control of effective connectivity, accompanied and possibly even functionally mediated by a conscious "epistemic agent model". The overall process is aimed at increasing the epistemic value of pre-existing states in the conscious self-model, without causally looping through sensory sheets or using the non-neural body as an instrument for active inference.

Keywords

Attentional agency | Cognitive affordance hypothesis | Cognitive agency | Epistemic agency | Epistemic agent model | Epistemic goal states | Epistemic self-control | Epistemic value | Interactive inference | Interoceptive inference | M-autonomous | M-autonomy | Mind wandering | Phenomenal self-model | Predictive control | Veto control

1 Introduction: The Problem of Mental Action

There is no obvious way to accommodate mental action within the framework of predictive processing (PP). Examples of mental action are the volitional control of endogenous attention (as, for example, in deliberately focusing one's attention on a perceptual object or attaching it to an abstract goal-representation), trying to retrieve a series of images from episodic memory, using semantic memory to bind an object as a token to its type, active categorization or the construction of part-whole relationships, as well as engaging in mental calculation, the "building" of an argument from premises or high-level reasoning. Mental actions are a large and relevant subset of the domain of mental events, but it is unclear if they can be made amenable to scientific explanation using the conceptual instruments offered by PP. The latter approach holds out the promise of uniting perception, attention, and bodily action in a single formal framework. Yet if we take as our starting point a reasonably well-established notion such as "embodied active inference" (Friston et al. 2014; Fabry 2015; Fabry 2017), then while perhaps *bodily* actions can be explained in terms of "self-fulfilling motor fantasies" cancelling out proprioceptive prediction errors, *mental* actions cannot thus be appropriated.

In mental actions there is no motor plant to be controlled, no sensory manifold that could be manipulated by bodily movement. It is not easy to assimilate mental actions to the idea of active inference (Pezzulo 2012), simply because they do not necessarily involve any relevant changes in the non-neural body or the prediction of sensory events. From a metatheoretical perspective, mental action poses the interesting challenge of describing the deeper principles of goal-state selection and action initiation while subtracting the non-neural body and abstracting from issues of motor implementation.

My main claim in this chapter is that mental action is the predictive control of effective connectivity, where what is predicted is the epistemic value of states integrated into the phenomenal self-model under counterfactual outcomes. I will also claim that the circular causality constituting genuine men-

Metzinger, T. (2017). The Problem of Mental Action - Predictive Control without Sensory Sheets.
In T. Metzinger & W. Wiese (Eds.). *Philosophy and Predictive Processing:* 19. Frankfurt am Main: MIND Group. doi: 10.15502/9783958573208

tal action does not embrace events on any sensory sheet,[1] that mental action is a rare event, and that the specific phenomenal signature of mental action can be explained by a new content layer in the conscious self-model, the "epistemic agent model" (EAM), which may sometimes transiently emerge in the brains of human beings. In addition, I present the cognitive affordance hypothesis, which proposes that a central function of autonomous activity in the mind wandering network is to create a constant stream of *affordances for cognitive agency*, a continuing internal competition among possible cognitive actions.

Section 2 lays some conceptual foundations, Section 3 presents four building blocks of a future theory, and Section 4 draws some interim conclusions towards a first positive model for the new target phenomenon of mental action. I end by sketching a model of action initiation as a functionally adequate form of self-deception and draw attention to a metaphysical dilemma constituted by what I see as the three most important issues for future research.

2 What *Is* Mental Action?

I will briefly introduce some conceptual distinctions and tools in this section. Later I will situate these tools in the logical context of current theorizing on predictive processing, to prepare for a brief application at the end.

2.1 Mental Action versus Mental Behavior

Philosophers have thought long and hard about what distinguishes "action" from other kinds of event in the physical world (Davidson 2001; Dretske 1988; Wilson and Shpall 2016). As a matter of fact, "action theory" can be considered a small subfield within the discipline of academic philosophy. However, one of the major deficits of "action-oriented" views in cognitive science (e.g. Engel et al. 2013; Engel et al. 2016) is that there is no shared or even clearly defined concept of "action" in the background, no framework which could unite the new field from a metatheoretical perspective.

For the purposes of this paper, let us distinguish between "actions" and "behaviors" as follows. Actions and behaviors are a subset of the overt output of information-processing systems which are conceptually distinguished from other outputs by their conditions of satisfaction, being directed at goal states. Actions and behaviors can be successful, or they can fail. For actions, however, *conscious* goal-representation plays a central causal role. Actions can be terminated, suspended, or intentionally inhibited, and they exhibit a distinct phenomenological profile involving subjective qualities such as agency, a sense of effort, goal-directedness, global self-control, and ownership. Arguably, there is also a phenomenal quality of "ultimate origination", the more-or-less implicit appearance of a robust ability to do otherwise (see Section 2.3). Behaviors, on the other hand, while purposeful, do not entail explicit, conscious goal-representation. They are functionally characterized by automaticity, decreased context-sensitivity, and low self-control. Although they can be more rapid than actions, we may not even notice their initiation. While their phenomenological profile can at times be completely absent, behaviors typically involve the subjective experience of ownership without agency, where the introspective availability of goal-directedness varies and meta-awareness is frequently absent.

We can add a second conceptual distinction: there are not only bodily actions but also *mental* actions. Mental actions belong to the internal, *covert* output of some information-processing systems. Deliberately focusing one's attention on a perceptual object and consciously drawing a logical conclusion are examples. As with physical actions, mental actions possess satisfaction conditions (i.e. they

1 A sensory sheet is a collection or population of receptors. Sensory receptors absorb physical energy from a stimulus; in this way they can also function as *transductors* by transforming physical energy into electrical energy in the form of neural firing. Examples are the photoreceptors on the retina, which hyperpolarize in response to electromagnetic energy, or the olfactory receptor cells enabling odor perception by forming a spatially discontinuous olfactory sheet in different nasal cavities and airflows. The example of mechanoreceptors across the body surface shows that there can be great variations in the ability for tactile discrimination on the skin, because the minimal interstimulus distance required to perceive two simultaneously applied stimuli as distinct can vary between 1 and 45 mm.

are directed at a goal state) and, although they mostly lack overt behavioral correlates, they can also be intentionally inhibited, suspended, or terminated. In addition, however, they are characterized by their temporally extended phenomenology of ownership, goal-directedness, a subjective sense of effort, and the concomitant conscious experience of agency and *mental* self-control. For the purposes of this paper I will also assume that mental actions are typically directed at epistemic goal-states. Examples of such states are "seeing this visual object more clearly and in greater detail", "knowing the sum of 2 + 3", and "having arrived at a valid conclusion".

Mental action is a specific form of flexible, adaptive task control with proximate goals of an *epistemic* kind: in consciously drawing conclusions or in guiding attention there is always something the system wants to *know*, for example the possibility of a consistent propositional representation of some fact, or the optimal level of perceptual precision. There may also be relevant classes of non-epistemic or purely "conative" mental actions, for example those that are directed towards reward events. Here, I would propose that for the very large majority of mental actions "reward expectation" can be conceptually reduced to "epistemic value", for example as a relevant fitness-enhancing information gain under counterfactual outcomes. But let me keep things simple by limiting the investigation to epistemic mental actions. If irreducibly non-epistemic mental actions exist, they are not covered by the main arguments in this chapter.

Not only are there mental actions, however; there are mental *behaviors* too. These also belong to the *covert* output of some information-processing systems. "Mind wandering", or spontaneous, task-unrelated thought, is a paradigm case of unintentional mental behavior (Metzinger 2013a; Metzinger 2015; Metzinger 2017; Smallwood and Schooler 2015). Some mental activities are not autonomously controllable, because one centrally important defining characteristic does not hold: they cannot be inhibited, suspended, or terminated. Mental behavior may often be purposeful, but it exhibits no conscious goal-representation nor overt behavioral correlate. It is characterized by an unnoticed loss of mental self-control and a high degree of automaticity, plus a lack of sensitivity to the external situational context. The phenomenological profile is marked by ownership without agency, a variable or null capacity for introspective availability of goal-directedness, and — frequently — lack of any meta-awareness (Schooler et al. 2011). To be sure, involuntary mental behavior may serve many important epistemic functions, for example creative incubation, the consolidation of long-term memory (Mooneyham and Schooler 2013), or the continuous updating and maintenance of an autobiographical self-model (Metzinger 2013a). It may also assist the refinement and consolidation of goal-representations, by gradually descending from the abstract level of their general satisfaction conditions towards concrete, embodied motor intentions (Medea et al. 2016), and in this way it may enable further epistemic benefits in the future (Bortolotti 2015). In addition, low levels of cognitive control can boost epistemic processes in open-ended tasks relying on the use of diverse sources of information and involving temporal delays (Amer et al. 2016, p. 911), and a considerable number of mind wandering episodes are even intentionally controlled in their onset (Seli et al. 2016). There is a difference between "zoning out" and deliberately "tuning out", and there are interesting and fine-grained phenomenological nuances connecting both phenomena: some episodes are intentionally initiated, and sometimes even sustained by wilful "rebooting", but as they unfold over time they typically become unintentional mental events — inner behaviors, not inner actions. As I will propose, unintentional mental behavior is interesting, because it helps to *constitute* mental action.

2.2 Mental Autonomy: Variable Degrees of Epistemic Self-Control

We have just seen that mental actions are typically directed at epistemic goal states. Interestingly, these states, if successfully brought into existence, are often "self states", because it is the organism itself which has acquired new (epistemic) properties. It is the system *itself* (and not the world) which now knows something new, has optimized the depth of perceptual object representation, has arrived at a

novel conclusion, and so on. There are obvious exceptions. For example, whenever human beings actively contribute, via more complex forms of social interaction, to the knowledge their *group* possesses, they change epistemic group properties as well. In the very large majority of cases, however, mental action is a process by which an individual changes their own epistemic properties. If we assume the conceptual distinctions above, general intelligence can be seen as the capacity for adaptive epistemic self-control and *mental* self-control plays a central role. Self-control comes with different and variable degrees of autonomy. This raises the question as to the minimal degree of functional autonomy which enables an information-processing system to become an agent, in the sense that it crosses the threshold from mental behavior to mental action.

A simple and empirically tractable concept is "M-autonomy" (Metzinger 2015). In general, autonomy is often framed as the capacity for rational self-control of overt behavior, whereas the term M-autonomy refers to the specific ability to control one's own mental functions, like attention, episodic memory, planning, concept formation, rational deliberation, decision-making, and so on. One route to a richer conceptual analysis is to describe it as the capacity for *second-order mental action*, i.e., vertical, intra-mental, top-down control. This can be decomposed into the following attributes:

- the imposing of rules on one's own mental behavior;
- explicit goal-selection and commitment, establishing goal-permanence;
- satisfying the constraints of rationality or rational guidance; and
- the ability intentionally to inhibit, suspend, or terminate a process.

This last condition, "veto control", is the central semantic element in defining M-autonomy: if one cannot terminate one's own activity, one cannot be said to be autonomous in any interesting sense. This element can be empirically grounded and gradually refined, and it may prove heuristically fruitful in guiding research. Veto control is a manifestation of the ability to suspend or inhibit an action voluntarily, and from a logical point of view it is a functional property which we do not ascribe to the brain but to the person as a whole. Let us call the capacity "intentional inhibition".[2] During a mind wandering episode, we do not have this capacity because we cannot actively suspend or inhibit our own mental activity (Metzinger 2013a; Metzinger 2017). Therefore, our degree of epistemic self-control is low.

How would one go about isolating the neural basis of autonomous epistemic self-control? From a conceptual point of view, every representation of agency must have three logical components: a model of an entity exerting control (the "self"), a model of the satisfaction conditions of the specific action (the "goal state"), and an asymmetric relationship dynamically connecting and transiently *integrating* the first two components (the "arrow of intentionality"). For the special case of mental action this has three implications. First, we must avoid any homunculus fallacy with regard to the first component. Second, we must do justice to the fact that the number of possible goal states is extremely large, because at any given point the number of possible targets for introspective attention — as well as the number of potential contents for the control of abstract, symbolic thought — is much larger than that for bodily action. For human beings, the set of *cognitive* target objects and states is much larger than what they could reach for, grasp, or run to — simply because our inner environment has become much richer and more complex than the concrete space of causal interaction in which our physical bodies are situated. Third, the dynamic, relational component connecting the first two elements has to be extremely fast and flexible, and it must be able to adapt to a complex task domain in a fluid and highly context-sensitive fashion. This already puts valuable constraints on possible architectures and means of realization.

For example, given the first conceptual constraint above, for every individual mental action, one would expect task-independent and task-dependent components to become integrated. Functional

2 In adopting this terminological convention, I follow Marcel Brass (Brass and Haggard 2007); an excellent and helpful recent review is (Filevich et al. 2012).

connectivity analyses point to a combination of intrinsic and task-evoked connectivity patterns, reflecting a global network architecture composed of an intrinsic part which is also present during rest, and task-general as well as task-specific patterns of connectivity evoked by specific demands (Cole et al. 2014). It is important to avoid implicit homunculus fallacies and to dissolve the "cognitive agent" component into a statistical analysis which assigns flexible sets of interacting cognitive subsets to every specific instance of task execution, but in a way that still reveals functional clusters or "network roles" (Mattar et al. 2015, p. 3). The problem is not to find some mysterious little "man in the head", but to develop a formal understanding of how the brain manages to continuously approximate an optimal balance between global integration and local flexibility, to develop a framework explaining how a stable state can be maintained while transient coalitions of network units generate specific cognitive behaviors. Such an understanding is gradually emerging. "Network roles" will be very complex functional states — context-sensitive, non-encapsulated, determinate — which can conceptually be defined as very large clusters of causal relations. If human beings belong to the class of probabilistic automata (Putnam 1967; Putnam 1975), however, then this engenders the challenge of how even to begin reliably mapping such complex probabilistic roles onto brain regions.

Recent empirical work reveals the dorsal fronto-median cortex (dFMC) as a candidate region for the physical realization of this very special form of purely mental second-order action.[3] It does not overlap with known networks for external inhibition, and its computational function may lie in predicting the social and more long-term individual consequences of an unfolding action, that is, in representing the action's socially and temporally more distant implications for the organism.[4] There is a considerable amount of valuable neurobiological data on the physical substrates of intentional inhibition in human beings, and a number of studies have already led to more abstract computational models of volitional control, action selection, and intentional inhibition itself (Filevich et al. 2012; Filevich et al. 2013; Campbell-Meiklejohn et al. 2008; Kühn et al. 2009; Brass and Haggard 2007). These data are valuable not only for understanding the "back end" of many mind wandering episodes — the transition from mental behavior to mental action — but also for a more comprehensive theory of mental autonomy (for more, see Metzinger 2013a, Section 3.3). From a philosophical perspective, the functional property of M-autonomy is interesting for a wide range of reasons, including its relevance to our traditional notions of a "first-person perspective" (1PP) and "personhood" (Metzinger 2015). If one cannot control the focus of one's attention, then one cannot sustain a stable perceptual first-person perspective, and for as long as one cannot control one's own thoughts, one cannot count as a rational individual.

What could be a first, empirically plausible candidate for a neural realization of the complex functional demands posed by selective cognitive self-control? The fronto-parietal network (FPN) may be a good candidate for this cluster of functional properties (Cole and Schneider 2007; Niendam et al. 2012; Cole et al. 2013). Kalina Christoff and colleagues (Christoff et al. 2016, p. 721) hypothesize that it supports the deliberate constraining of the contents of thought. The FPN plausibly plays a central role in mental health: impaired cognitive self-control and disrupted processes of goal-representation are markers of disease across a large spectrum of neuropsychiatric conditions (Cole et al. 2014), and domain-general measures of fluid, culturally invariant and knowledge-independent intelligence have been found to be specifically correlated with the lateral prefrontal cortex, a circumscribed region within the FPN (Cole et al. 2012). Recent models of cognitive control support the idea of a "flexible hub" which can at the same time monitor and causally influence a large variety of task-relevant information sources (Cole et al. 2012, p. 8997). Relative to other known networks, the FPN is especially active during phases of highly adaptive task control, exploiting global variable connectivity by flexibly

3 See (Kühn et al. 2009), (Brass and Haggard 2007), and (Campbell-Meiklejohn et al. 2008). A helpful recent review of negative motor effects following direct cortical stimulation, listing the main sites of arrest responses and offering an interesting discussion, is (Filevich et al. 2012).

4 This passage draws on (Metzinger 2013a). See also (Filevich et al. 2012, Filevich et al. 2013).

shifting the connectivity pattern across many different brain regions and a wide variety of tasks. It has also been proposed that the FPN employs principles of "compositional coding", which would allow for certain connectivity patterns and representational contents to be reused and recombined (Anderson 2015) in order to transfer existing knowledge across tasks, thereby enabling the rapid learning of novel tasks in new functional contexts (see Cole et al. 2013, fig. 1). While the FPN's variable connectivity is truly global, it has also been identified as one of the ten major functional networks which partition the brain into intrinsic functional clusters, independent of particular task- or goal states. In the current context, it may be relevant to investigate the FPN's causal interaction with the default-mode network (DMN). This is a distinct network of interacting brain regions, whose activity is highly correlated, which automatically activates when a person is not involved in any task — such as at wakeful rest, daydreaming, when the individual is simulating social situations, or during autobiographical rumination. Interestingly, the variable connectivity of the FPN has been found to be significantly greater than that of the DMN with the entire brain (cf. Cole et al. 2013, fig. 5) — a point to which I shall return in Section 3.2.

2.3 Mental Action Type 1: Volitional Attention

Many authors have recently begun to ask the question of whether predictive processing (PP) can be extended to a genuine model of high-level cognition (Barsalou 2016; Butz 2016; Pezzulo et al. 2016; Spratling 2016). Let us briefly distinguish the two main types of mental action, which have to be accommodated under the PP approach. I will return to them in Sections 3.1 and 3.2.

Type 1, volitionally controlled attention, brings about a specific set of phenomenal properties, as is the case for pain or the subjective quality of "blueness" in a visual color experience (Metzinger 1995). Attentional agency (AA) is the conscious experience of actually initiating a shift of attention, of controlling and fixing attentional focus on a certain aspect of reality. AA involves a sense of effort, and it is the phenomenal signature of our functional ability actively to influence what we will come to know, and what, for now, we will ignore. As with all other forms of agency, it also involves the subjective quality of "ultimate origination" mentioned above — from the first person perspective, it seems that one could have done otherwise; in the way one experiences the overall process from the first-person perspective, any unconscious causal precursors are necessarily unknown such that the first subjective event carrying the phenomenal character of control (say, determining the focus of attention) necessarily appears as spontaneous and uncaused, emerging "out of the blue" as it were. On this level of the hierarchy, it is an unpredicted internal event. What we call "agency" refers to an interpretation of this fact as "initiation" or "origination": it is the activation of an internal self-model trying to explain away, by creating an explicit, supra-modal representation of an entity capable of ultimate origination and spontaneous self-causation, the surprise involved in suddenly achieving autonomous self-control. I return to this point in Section 4.3.

Consciously experienced AA is theoretically important because it is probably the earliest and simplest form of experiencing oneself as a *knowing* self, as an epistemic agent. To consciously enjoy AA means that one (the cognitive system as a whole) currently identifies with the content of a particular self-representation, that one operates under an "epistemic agent model" (EAM; see Section 2.5 below and Metzinger 2013a; Metzinger 2013b) active in one's brain. Being a phenomenological entity, an EAM can always be a hallucination, but typically it will be a window of self-knowledge, telling the system that M-autonomy has been achieved — we must always be careful to keep functionalist, mech-

anistic, and epistemological readings of this new term apart (see Section 3.1). AA is fully transparent:[5] the content of one's conscious experience is not a self-representation or a process of self-modeling, of depicting oneself as a causal agent in certain shifts of "zoom factor", "resolving power", "resource allocation", and so on. Rather, one directly experiences *oneself* as, for example, actively selecting a new object for attention. During nocturnal dreams and mind wandering episodes we do not have AA, although these episodes can of course be *about* having been an attentional agent in the past, or *about* planning to control one's attention in the future. Other examples of situations in which this property is selectively missing are non-lucid dreams and non-REM-sleep (rapid eye movement) mentation (Metzinger 2013b; Windt 2015), and also infancy, dementia, and severe intoxication.

2.4 Mental Action Type 2: High-Level, Symbolic Cognition

What is reasoning, logical thinking, or mathematical cognition from a PP perspective? Can prediction-based mechanisms be fully detached from overt sensorimotor loops (Pezzulo 2016, p. 33)? We can conceptualize cognitive agency (CA) as an abstract mental simulation of embodied actions, first executed using the physical, non-neural body. Such actions could have been the manipulation of discrete symbolic tokens in the external world, the use of gestures or bodily sign language, or even full-blown speech acts. But we can also frame them as abstract, inclusively internal versions of adaptive action control, involving predictive loops (see Section 3.2).

Again, there is a distinct phenomenology of currently being a cognitive *agent*, which can lead to experiential self-reports such as "I am a thinking self in the act of grasping a concept", "I have just actively arrived at a specific conclusion", "I am attempting to build an argument", and so on. We should be careful not to confuse the level of functional analysis ("autonomous cognitive self-control") with phenomenological readings based on verbal self-reports. What AA and CA have in common is that, in both cases, we consciously represent ourselves as epistemic agents: according to subjective experience, we are entities that actively construct and search for new epistemic relationships to the world and ourselves. We are information-hungry; there is something we want to *know*.

2.5 The Phenomenology of Epistemic Self-Control: The EAM

If the EAM is what explains the unifying phenomenal signature of mental action, then it is the common denominator on the phenomenological level of analysis. Phenomenologically, for a conscious cognitive system to operate under an epistemic agent model means for it to *know that it knows*. According to subjective experience, some of its own states have an epistemic value. They seem to present information about the world, and they often do so in a stable and counterfactually rich way — we can imagine many situations or possible worlds in which they would stay the same. As yet, we are only operating on a phenomenological level of description, but from a computational perspective this additional feature could serve as an introspective indicator of the fact that there exist many counterfactual manipulations under which the information presented by these states would remain largely invariant because there is a specific, abstract form of robustness or stability characterizing their informational content. Knowledge in this sense is the possession of *reliable* information.[6]

5 "Transparency" is a property of conscious representations, namely that they are *not experienced* as representations. Therefore, the subject of experience has the feeling of being in direct and immediate contact with their content. Transparent conscious representations create the phenomenology of naïve realism. An opaque phenomenal representation is one that is experienced *as* a representation, for example in pseudo-hallucinations or lucid dreams. Importantly, a transparent self-model creates the phenomenology of identification (Section 3; Metzinger 2003; Metzinger 2008). There exists a graded spectrum between transparency and opacity, determining the variable phenomenology of "mind-independence" or "realness". Unconscious representations are neither transparent nor opaque. See (Metzinger 2003) for a concise introduction.

6 The classical philosophical issue here is at what point rich and reliable information turns into semantic information, something that could be true or false (Dretske 1988). It is important to distinguish this metatheoretical question from the empirical question of how and at what level of reliability and counterfactual invariance a given type of cognitive system *internally models* information it possesses as semantic information.

Having an EAM is an instance of self-consciousness. For a human being it means that, subjectively, it now possesses a specific kind of self-knowledge — *knowing* that one knows and that one is able actively to control certain epistemic states. An EAM is a model of a single entity capable of autonomous epistemic self-control and ultimate origination. As this happens on the level of conscious processing, it also creates the phenomenology of ownership for certain states of perceptual or cognitive knowledge. Whatever can be autonomously controlled generates the subjective experience of ownership — a general principle which holds not only for the sensorimotor control of one's own body but also for a wide range of phenomena, ranging from simple tool-use to virtual re-embodiment in avatars and robots (Cohen et al. 2014; Cohen et al. 2014; Lenggenhager et al. 2007; Blanke and Metzinger 2009; Metzinger 2008). The conscious processing does not only enhance adaptivity, flexibility and context-sensitivity in active acquisition of knowledge. It also adds the property of *globality*: because states with high epistemic value are subjectively "owned" and integrated into a phenomenal self-model, possessing them becomes a property of the system as a whole — something it ascribes to itself, a unified space of hypothesis generation.

An EAM is also a new unit of identification (Metzinger 2017; Metzinger 2013a, p. 10). It enables self-reports of the type "I *am* this knowing self!" Phenomenologically, identifying with the content of an EAM typically also refers to a high probability that this system will come to know further novel aspects of the world (bringing about "epistemic optimism", i.e. a high expected rate of error reduction). This also means that the system will be *motivated* to bring about a lot of those changes in its own global state of knowledge (it has a "capacity for epistemic self-control", and the related phenomenal sense of control is expressed as positive affect). Perhaps it also has an "autobiographical" model of past fluctuations in prediction error minimization and therefore expects to repeatedly initiate new epistemic actions after a certain time has passed (see Van de Cruys 2017, for an interesting discussion with regard to curiosity and affective value). This is what having an internal model — not only of some passively knowing self but of an epistemic *agent* — means. There are goal states plus a possibility of failure, there is a corresponding high-level capacity, and often this capacity is not just an abstract feature but something that is exerted — a concrete process consciously experienced.

In this way a running EAM is a phenomenal self-model containing an ongoing prediction of future epistemic states, plus an explicit representation of the capacity to bring these states about. That is, there is a specific, high-level capacity for self-control and there are variable rates of prediction error minimization and degrees of autonomy related to this capacity. These degrees of autonomy are always linked to a certain probability which can itself become the target of a predictive model. For example, one could in principle measure autonomy from the outside, by counting, relative to a specific situation, the number of times an agent is able to suppress a spontaneously occurring impulse for motor action. This relative frequency would yield a probability, which we could use to express the degree of autonomy. The same also could be done from the inside, for epistemic actions, with the help of a specific layer in the phenomenal self-model.

A unified self-model does not necessarily entail that a distinct entity such as "a" self exists as well: the phenomenology of "knowing that one knows" could be constituted by something other than the possession of reliable information, as in self-deception. That second-order state ("knowing that one knows") might not be epistemic at all, or, as for example in certain states of belief and subjective certainty caused by epileptic seizures or direct electrical stimulation of the insula (Picard 2013; Picard et al. 2013), the first-order state might be a physical artefact or based on highly unreliable information. The same goes for the phenomenal qualities of "autonomy" and "ultimate origination": the agency component characterizing the subjective experience of being an active, knowledge-seeking self could be a first-person phenomenon only. I will therefore return to non-phenomenological levels of analysis in Section 3.3, to discuss potential constitutive, causal, or explanatory relationships between predictive control and the conscious EAM.

3 Building Blocks for a Positive Model

We now have a set of conceptual distinctions and tools that permit us to develop a positive model of mental action. My aim, however, is more modest than this may sound: I merely want to arrive at a conceptual nucleus for the notion of "mental action", a first — and hopefully heuristically fecund — working concept which can be refined and empirically enriched as we go along. The claim in this section will be that mental actions are a specific form of EAM-mediated predictive control of effective connectivity, aimed at increasing the epistemic value of pre-existing states in the self-model without using the non-neural body as an instrument for active inference.

3.1 Building Block 1: Attentional Agency under PP

In his seminal and ground-breaking book, *The Predictive Mind*, Jakob Hohwy draws our attention to the central issue:

> Endogenous attention often seems to come with a volitional element that is not quite captured by […] relatively mindless cue directing. We decide to attend to some feature, and then act on that decision. In prediction error parlance, this aligns volitional attention with active inference. The question is, can we conceive of active inference such that it accommodates expectations for precisions? (Hohwy 2013, p. 197)

This is the right question. To find an answer, we need a better understanding of the epistemology of attention, a plausible computational story, and a representationalist analysis of its specific phenomenal character, i.e. the introspectively available conscious content activated along with attentional agency.

Epistemologically, every form of volitional attention is a form of introspection. External scaffolding by the non-neural body is not a necessary condition. We can visually attend to hallucinations with closed eyes or to the hypnagogic imagery preceding sleep onset, even after the body has entered sleep paralysis. During a lucid dream, we can wilfully attend to the dream environment as well as to sensations in the dream body (Metzinger 2008; Windt 2015). A clearer, and perhaps even more provocative, way to bring out this distinction is by saying that, functionally, all high-level, volitional attention is active introspection: it always operates on aspects of an internal world model. If I attend to a visual object in front of me, then I am actively optimizing the second-order statistics of this (inclusively internal) object model. I claim that this is not only true during a lucid dream, but also during waking life. If I attend closer to some visceral sensation in my stomach or to the sensations going along with my breath, then I actively optimize the second-order statistics of my interoceptive self-model. Functionally, both forms of mental action are exclusively internal, because their satisfaction conditions describe a goal state which is a state of an internal model. Phenomenologically, however, these situations are *very* different: in the first case I experience myself as attending to the environment; in the second case I experience myself as actively introspecting certain aspects of bodily self-consciousness (see below).

Jakob Hohwy writes:

> All this seems no different in principle from the active inference we have come across before: a counterfactual hypothesis induces a prediction error causing us to change our relation to the world. It is a slightly unusual instance of action because the way we change our relation to the world is to increase our sensory gain in one region of space. The difference is then that the active inference is driven not by which prediction error we expect, but by the precision of the expected precision error — just as we can sample prediction error selectively, we can sample their precisions selectively. (Hohwy 2013, p. 198)

I think the general idea is correct, but two details are not quite right. Firstly, in being a volitional attentional agent, the system changes its relationship not to "the" world, but to an internal world *model*. Of course, this neurally realized model is a part of the physical world too, but for the system this means that it actively changes its relationship not to the environment, but to a part of itself. Its epistemic perspective is one of active introspection. Secondly, the gain that is increased is not a gain relative to an extra-organismic "region of space", because gain is really only increased for one region of *internal* computational space.[7] The "signal" already is an exclusively internal event, because it is a non-sensory "signal".

Computationally, attentional agency (AA; Metzinger 2003, 6.4.3; Metzinger 2006, Section 4; Metzinger 2013a; Metzinger 2015; Metzinger 2017) is the ability to control precision expectations. However, it is important to make the following distinction: note that precision is a property of the *signal*. If attentional agency is conceptualized as a form of prediction only, then what is predicted is a property of the raw sensory manifold, which is not a representational entity in itself. Its statistics just reflect the causal structure of the world. If AA is conceptualized as a form of *control*, however, then what is controlled is a property of a representational entity, namely the precision expectation embodied by an internal model. This is what makes attentional agency in this richer sense an inherently *mental* kind of action: it optimizes second-order statistics by changing properties of an internal model. The standard version of active inference at most optimizes the precision of a non-representational entity; AA is mental because it optimizes a representational entity.

What about the phenomenology? Clearly, attentional agency is not just any form of controlling precision expectations: it is the subset accompanied by (and possibly functionally mediated through) a conscious model of the self as an epistemic agent — what I have called an EAM in Section 2.5 above. It is the best currently available global hypothesis about certain future epistemic states, in this case states of *perceptual knowledge*. Because this hypothesis originates in what I will call "EAM-space", it automatically predicts a specific form of *self-knowledge* as well: probably, there will not only be a richer and more detailed experience of some specific perceptual object in the near future, but this will also be modelled as the perception of a "knowing self", a self which will be phenomenally aware of this very fact. Recall how an EAM is a model of a single entity, capable of autonomous epistemic self-control, and how, on the level of conscious processing, this creates the phenomenology of ownership for certain states of perceptual or cognitive knowledge.

Being an attentional agent therefore is always a special form of *self*-consciousness: one becomes aware of one's *own* capability for — and the actual process of — controlling the quality of one's own perception, its depth and resolution. At the same time, while quashing prediction error about one's expected precision, one experiences a specific sense of effort, a sort of resistance — and, in combination with the subjective quality of ultimate origination described above, it is this which engenders the relevant, exclusively mental sense of agency. Perhaps we can describe this experiential tension[8] as the way in which active, top-down precision control "collides" with an automatically arising precision prediction error. If so, one might speculate that the temporal curve in which the phenomenal sense of agency unfolds over time exactly reflects this "collision" and its eventual resolution.

The possibility of relating the phenomenology to a fully internalist computational analysis of volitional attention, as the EAM-mediated control of second-order statistics for perceptual states, reveals what is perhaps the philosophically most important point: the phenomenology and epistemology of environment-directed versus introspective attentional agency can diverge dramatically. Functionally, both forms of mental action are exclusively internal. Phenomenologically, however, in the former I

7 Jakob Hohwy discusses volitional spatial attention here, but the point would also hold for volitional feature attention. The relevant processing stream is exclusively internal.

8 A much more detailed analysis of subjectively experienced tension and its relationship to self-deception can be found in (Pliushch in press). See also (Pliushch 2017).

will experience myself as attending to the world out there, while in the latter I will experience myself as introspecting aspects of my own mind, my emotional state, or my bodily self.

3.2 Building Block 2: Cognitive Agency under PP

In *Surfing Uncertainty*, Andy Clark reintroduces and develops the notion of an "affordance competition" (Cisek 2007; Cisek and Kalaska 2005; Cisek and Kalaska 2010; Clark 2016, p. 179). This approach could be extremely helpful in understanding what high-level symbolic cognition really is, and in situating it in a PP-framework. I will extend this idea further, into the domain of unintentional mental behavior and mental action. The interim goal will be to create a second building block to arrive at a positive model for the second main category of mental actions. Clark writes:

> One powerful strategy […] involves rethinking the classical sense-think-act-cycle as a kind of mosaic: a mosaic in each which shard combines elements of (what might classically be thought of as) sensing and thinking with associated prescriptions for action. At the heart of this mosaic vision […] lies the simultaneous computation of multiple probabilistically infected 'affordances': multiple possibilities for organism-salient action and intervention.
> The idea here is that the brain is constantly computing — partially and in parallel — a large set of possible actions and that such partial, parallel, ongoing computations involve neural encodings that fail to respect familiar distinctions between perceiving, cognizing, and acting. (Clark 2016, p. 177; 180)

Could there be something like an exclusively *internal* affordance competition? Let me propose that mind wandering, the almost continuous appearance of task-unrelated thoughts, may be exactly this — the creation of a constant flow of possible *mental* actions, a dynamic *inner* environment constituted by non-sensory events which need to be predicted and controlled. Call this the "cognitive affordance hypothesis".

We have already seen that apparently spontaneous, apparently task-unrelated thought can conceptually be described as a form of unintentional behavior. The cognitive affordance hypothesis states that a central function of autonomous activity in the mind wandering network is to create a constant stream of *affordances for cognitive agency*, a competition among possible cognitive actions. It is empirically plausible to assume that large parts of this pattern overlap with activity in the default mode network (DMN; Buckner et al. 2008; Christoff 2012; Christoff et al. 2009; Weissman et al. 2006; Stawarczyk et al. 2013; Andrews-Hanna et al. 2010; Mantini and Vanduffel 2012; Buckner and Carroll 2007; Mason et al. 2007; Spreng et al. 2009), but that it also extends to other functional structures such as the rostrolateral prefrontal cortex, dorsal anterior cingulate cortex, insula, temporopolar cortex, secondary somatosensory cortex, and lingual gyrus (for a recent meta-analysis, see Fox et al. 2015).

For an intuitive grasp, let us imagine a situation in which we had as little control over our bodily behavior as we have over our own minds. Imagine we were beings which in the awake state were plagued by recurrent involuntary motor twitches and behaviors for roughly 30-50 per cent of the time, but that we could sometimes "seize" one of these bodily behaviors by bringing them under conscious control, thereby turning them into proper, goal-directed actions. Or imagine a sleep-walker, unconsciously navigating the world on autopilot, reactively driven by low-level affordances. Sometimes, however, the sleep-walker briefly "comes to", wakes up and becomes capable of autonomous bodily self-control. After each such episode he loses conscious sensorimotor control again, automatically returning to autopilot mode. I propose that what we call "thinking", "reasoning", or even volitionally attending to some object is exactly like this: first we are mental sleep-walkers; then we episodically turn into genuine mental agents by "seizing control", perhaps with the help of a conscious EAM, namely, by re-instantiating the functional property of M-autonomy. We navigate a self-created, internal affordance

landscape, but only rarely do we achieve autonomous self-control (roughly for only one third of our conscious lifetime, cf. Metzinger 2015, fig. 2).

Mind wandering is mental sleep-walking. But unlike embodied somnambulism it actually *sets up* the inner world which later allows us to autonomously perform mental actions during a state of "full consciousness". One central function of the mind wandering network is to provide us with an internal environment populated by competing possible mental actions — unintentionally occurring mental events which have the potential to become part of a stable control loop, thereby turning into genuine inner actions. A rich inner action repertoire enables an equally rich landscape of possible interactions with our inner world; it continuously opens a whole range of new functional windows. There is no inner world containing static mental objects, seen through a single, rigid window frame. What mind wandering does, rather, is to create a fluid and highly dynamic task domain. Every spontaneously occurring "task-unrelated" mental event is a potential task in itself, a cognitive affordance, a dynamic state which has the potential to be selected and transformed from unintentional mental behavior into mental action. The mind wandering network sets up this task domain, and sometimes frontal regions of the brain latch onto the best candidate, and create a transient functional integration with the EAM, "pursuing" a possible thought further, now in a consciously controlled fashion, "trying to stay one step ahead" on the cognitive level. Perhaps we can imagine this process as the frontoparietal control network being flexibly engaged with either the default or dorsal attention network in support of goal-directed cognition (Spreng et al. 2010). Perhaps, in this specific case, we could even speak of mental "patterns of action readiness" and then describe the episodic emergence of cognitive control as the appearance of an "optimal grip", this time on an *internal* situation (Bruineberg et al. 2016; Bruineberg 2017).

I will not enter into a discussion of candidates for neural implementation in this paper, but an empirically plausible model could involve the posterior cingulate cortex (PCC) as a driver of this specific type of activity, with the medial prefrontal cortex (MPFC) modulating and directing the flow of high-level cognition, for example by acting as a gateway in selecting semantic self-representations (cf. Davey et al. 2016, p. 935). Please note that such a "semantic" self-model would already be an exact conscious representation of the system itself as a *carrier of meaning*, with not only physical but also epistemic properties. And this is what mental action is all about — the control and creation of new epistemic properties. Put differently, if conscious perception can be described as a form of controlled hallucination, conscious thinking is a process of controlled mind wandering.

This is very much in harmony with the recently developed idea of adaptive action control as the "navigation of an affordance landscape" (Pezzulo et al. 2016). What it adds are three major aspects. First is the mind wandering network as an automatic, subpersonal mechanism constantly creating a dynamic internal affordance landscape. Second is the phenomenal self-model as characterized by a hierarchy of predictive horizons (or nested timescales) generating different types of representational content. And third is the concept of an "epistemic agent model" as one specific content layer in the phenomenal self-model — the transient conscious correlate of the event of achieving autonomous epistemic self-control.

On the mental level, the mind wandering network is the system that enables the rapid switching of actions (as Pezzulo et al. 2016, p. 415 demand), by continuously processing alternative potential mental actions and creating *expected* cognitive affordances even during phases of goal-directed activity. As the human self-model integrates different predictive horizons, different timescales generate different types of nested self-representational content: the body-model predicts somatosensory events within very small time-windows of only a few hundred milliseconds, whereas the autobiographical self-model predicts events which may lie years ahead, including the organism's eventual death. The cognitive layer not only biases affordance competition at lower hierarchical levels, it crucially allows an agent to adaptively destroy and create new cognitive affordances in order to realize long-term goal-representations. M-autonomy involves veto control and refers to the functional property enabling the system not

only to "reactively pick up currently available affordances" (Pezzulo et al. 2016, p. 416), but to actively sculpt its own internal affordance landscape in accordance with the process sketched above. The EAM is its transient conscious correlate, making the current fact of successful predictive control in this task domain introspectively available.

So the core idea of the cognitive affordance hypothesis is that mental action is the control of spontaneous mental behaviour, which can be described as an internal affordance competition mediated by a large neural network overlapping with the MPFC. Recent connectivity analyses reveal the DMN as consisting of regions at the top of a representational hierarchy which describe the current representational landscape in the most abstract terms (Margulies et al. 2016). Cognitive agency, critically involving the MPFC, then becomes the active sampling of an internal, highly abstract, and non-interoceptive environment. There are no sensory signals involved; hence their causes cannot be revealed. Physical events on sensory sheets play no proximal causal role. This leads to the conclusion that what is revealed can only be non-sensory, abstract properties of ongoing neural dynamics. But what exactly is it that could be predicted by this form of mental action?

3.3 Building Block 3: The Convergence of Instrumental and Epistemic Action for the Special Case of Mental Action

In his reply to Wanja Wiese's commentary on his target article in the Open MIND collection, Anil Seth writes:

> I suggest thinking instead of a continuum between epistemic and instrumental active inference. This is simply the idea that active inference — a continuous process involving both perception and action — can be employed with an emphasis on predictive control (instrumental), or on revealing the causes of sensory signals (epistemic). (Seth 2015b, p. 7)

Let us extend this idea to the domain of mental action. The target paper itself (Seth 2015b, cf. Section 2.3) had drawn attention to a deep connection between PP and mid-20th century cybernetic approaches which focused on the prediction and control of behavior. Clearly, for any system that has begun to manifest unintentional *internal* behavior (such as spontaneous, task-unrelated thought) it becomes necessary to predict and control exactly this internal behavior as well, because it is a constant source of uncertainty. If this behavior is an *epistemic* activity (as mind wandering seems to be), the new task would have to be described as "predictive epistemic self-control". Seth writes:

> [R]ather than seeing PP as originating solely in the Helmholtzian notion of "perception as inference", it is fruitful to see it also as a process of model-based predictive control entailed by a fundamental imperative towards internal homeostasis. (Seth 2015b, p. 9)

What is predictively controlled in cognitive agency is the spontaneous flow of subpersonal proto-thought, and we now begin to see how the explicit model on which this control is based could possibly be what was introduced on the phenomenological level as the EAM in Section 2.5. At this point we should be extremely careful, because the questions we confront are mainly empirical, not philosophical. For example, currently available data on mind wandering and cognitive control underdetermine possible metaphysical interpretations: on one end of the spectrum we might treat the EAM as a mere epiphenomenon devoid of any functionally relevant properties; at the other end it could be exactly the missing computational element that *constitutes* epistemic, model-based predictive control on the mental level. An intermediate position could be that most mental actions begin without an EAM (for example, by subpersonal processes of mental affordance competition), but *end* by transiently culminating in a conscious model of agentive second-order knowledge. Here, the EAM would not constitute, but *represent*. It might causally enable relevant second-order functional properties like veto

control (see Section 2.2), but it would not constitute the scientific target property of predictive control. Currently available empirical data do not help to decide between these three metaphysical options.

The same point holds for potential explanatory and causal relationships between the computational notion of "predictive control" as described by Anil Seth and the new concept of an EAM. There clearly is something like an EAM; it is a theoretically relevant part of our phenomenology which has been largely ignored in the past. But when I introduced the cognitive affordance hypothesis in the preceding section by saying that we are "seizing control", perhaps with the help of the FPN and a conscious EAM, I left at least two possibilities open: the EAM could be what causally *explains* predictive control, or it could be a mere phenomenal artefact *generated* by predictive epistemic self-control (I am very grateful to Jona Vance for critical discussion here). And when in Section 3.1. I introduced attentional agency (AA) as a type of epistemic self-control not involving sensory manifolds but exclusively aimed at precision expectations embodied in an exclusively internal model, I did not provide any concrete idea of the *form* of control that was presupposed. If it can be applied to processes not involving the non-neural body, then Anil Seth's work on predictive control is relevant, because we cannot solve the problem of mental action by appealing to an ill-understood and unexplained kind of control. Currently, the explanatory priorities are unclear. If I am on the right track, then novel computational models for mental action are needed, firmly grounded in empirical data and entailing testable predictions. To make a start, here are three.

From a third-person perspective, it remains true that the very large majority of all processes and hierarchical levels having to do with constructing a viable model of reality remain completely unconscious and are not part of the organism's phenomenological life-world. But it is conceivable that, in standard situations, an EAM could causally enable a new feature — *global* epistemic self-control plus *knowing* that one possesses this feature. Epistemic states are now represented as states of a single entity that functions as the system's current unit of identification, in this case as the currently active model of a "knowing self". This model makes the content of those states available for introspective attention, and for selective action control on the whole-system level. But it also possibly predicts and controls the acquisition of epistemic properties like the formation of a concept or the having of a belief. As with any other model, we can treat the EAM as an evolving space of possible hypotheses. EAM-space is a novel space, a new subregion or partition of the organism's global self-model, constituted by all those hypotheses exclusively predicting the organism's *epistemic* properties. Integration into the phenomenal self-model in turn creates new functional properties — the content of this space is now globally available to many other processing levels at the same time: for introspective attention and the flexible control of behaviour, be it motor or mental. The level of conscious processing is the level on which a unified ontology first emerges, and epistemic properties such as the possession of knowledge become elements of this ontology, parts of the system's reality model.

These are admittedly speculative observations, but a number of testable hypotheses can already be derived. For example, if an EAM is in place, then selectively blocking its introspective availability should suffice to trigger a mind wandering episode — a bout of spontaneous, task-unrelated thought (H1; see Axelrod et al. 2015; Broadway et al. 2015 for early examples of relevant studies). Conversely, if we find an experimental procedure to reliably re-establish predictive self-control in the relevant level of the hierarchy, then we should be able to turn a non-lucid dream into a lucid dream in a sleep laboratory (H2; Voss et al. 2014; Metzinger 2008), or experimentally to end a mind wandering episode in waking subjects, letting them unexpectedly "come to" again (H3).

Let us take another example: visually searching for a perceptual object by directing and sustaining the focus of visual attention, say, by "looking for" a lemon on the table in front of me. There are three very general expectations involved: that a lemon exists, that this lemon will soon be seen more clearly, and that it will be seen more clearly by *myself*. There is a hyperprior for "perceptual objecthood" (or, perhaps, only a mutual specification of counterfactually rich and hierarchically deep predictions), we have a high-level expectation of successful precision control in the visual domain, and there is a hy-

perprior for "epistemic agency". What is predicted is the holding of a relation between a subject and an object, in the near future, with the relation transiently uniting these elements being of a specific type: the category of "perceptual knowledge" or sensory-driven object representation.

Elsewhere, I have called the conscious correlate of this specific type of transient, dynamic, subject-object-representation the "phenomenal model of the intentionality relation" (or PMIR, see Metzinger 2003; Metzinger 2004; Metzinger 2006), because I claim that it is the conscious brain's subsymbolic and naturally evolved answer to the problem that Brentano and Husserl wanted to solve, much later, on a theoretical level. This model, in a fluid and highly context-sensitive manner, describes the temporal evolution of the "arrow of intentionality", that is, the asymmetric relationship of a system being *directed* at a goal state. For epistemic action, this goal state is one of possessing knowledge — for example, of being related to the lemon through the active, embodied process of "seeing" the lemon, of first establishing and then successfully sustaining a stable causal loop passing through one specific sensory sheet, namely, the receptor system constituting the outer statistical boundary of the visual system. The PMIR is the conscious model of a currently standing loop. Here, my empirical prediction would be that selectively blocking this and only this loop, say, by transcranial magnetic stimulation, would selectively destroy the phenomenology of epistemic goal-directedness (H4). It would leave the epistemic subject self-conscious and situated in a consciously experienced visual world, but without the phenomenal element of "epistemic seeing", because the quality of being a "visually-perceptually knowing self" would be selectively eliminated (e.g. Lamme 2003; Vandenbroucke et al. 2014).

So there are many ways in which Building Block 3 for the first working concept of "mental action" I am trying to construct could be empirically investigated. For something to be a representation means that misrepresentation is always possible (Dretske 1986). An EAM, if more than a mere phenomenal artifact, could always be suboptimal or false, a *bad* model, for example in cases of self-deception or delusion where we find many examples of human beings exhibiting a robust phenomenology of "knowing that they know" while from a third-person perspective they clearly don't (Section 2.2; Picard 2013). This point brings out another important aspect: if the relevant layer in the human self-model exists, then we can now think of it as something that is *itself* continuously optimized via hierarchically structured, context-sensitive forms of prediction error minimization. What it predicts are not so much events on a sensory sheet, but expected outcomes of epistemic self-control — its own epistemic properties as represented by the epistemic value of future states of the organism's self-model (see the next section).

Subpersonal events, if integrated with the EAM, can afterwards be attributed to the person as a whole. In this way we could perhaps say that mind wandering is model-free epistemic self-control, whereas mental agency is model-based epistemic self-control. If future research shows that the EAM is a real and distinct part of a cortical-control hierarchy characterized by causal powers and a functional role of its own, then the conscious model of the "knowing self" could be interestingly described as an internal representation of a good epistemic self-regulator. Mental action is exactly the special case of cybernetic self-regulation for which there is no non-neural body or "world in between", where what counts is the epistemic content of cognitive affordances, and the selective exploitation of pre-existing internal models created in an ongoing process of spontaneous activation. For this reason, "instrumental" and "epistemic" inference almost converge for EAM-optimization: the causes revealed in this process are only causes of hierarchically deep perturbations; the causal structure to be found comprises only pre-existing internal models.

On a purely functional level of analysis, we can then conclude that mental action is predictive control not of events on some sensory sheet, but of effective connectivity in deeper hierarchical levels. It is an intelligent way of sculpting patterns of effective connectivity. As mental actions are M-autonomous conscious events, they are marked out by a capacity for veto control plus explicit goal representation. What is still lacking is their "intentional object" — the satisfaction conditions to which they are direct-

ed. From a more analytical, third-person perspective we should now ask: what could be the abstract property that is represented as the "goal" of mental actions?

3.4 Building Block 4: Epistemic Value

Here is what Karl Friston and colleagues write about the exploitation/exploration distinction:

> In this setting, action reduces the difference between current and (unsurprising) goal states that are defined much like cybernetic formulations (Miller, Galanter, & Pribram, 1960). This difference can be reduced in two ways. First, by executing a pragmatic action, that fulfils goals directly (i.e., exploitation); for example, by visiting a known reward site in the context of foraging. Second, by performing an epistemic action (i.e., exploration) to disclose information that enables pragmatic action in the long run; for example, exploring a maze to discover unknown reward sites (Kirsh & Maglio, 1994). Clearly, most behavior has both pragmatic and epistemic aspects. (Friston et al. 2015, p. 2)

It is very tempting to think of high-level symbolic cognition, as well as introspective attention, as a process of "epistemic foraging in one's own inner world" (pragmatic *mental* action), cyclically re-visiting the well-known stream of internal affordances. And perhaps some types of mind wandering can be conceptualized as a form of epistemic but unintentional explorative behavior, creating ever new trajectories through the maze of one's own mind — for example, in an attempt to discover "unknown reward sites" by mental data-mining. Obviously, for any system which, as with human beings, already *has* a very rich and hierarchically deep internal model of reality, this model itself becomes an important resource, something that can be continuously exploited and explored — at least whenever there are no pressing problems presented by the organism's physical environment. In such situations our internal task domain becomes causally dominant, as it were, immediately driving us into epistemic foraging or into an automatic exploration of potentially unknown aspects of our inner reward landscape, of possible positive hedonic experiences. Of course, CA, AA and mind wandering can sometimes be more reward-seeking and sometimes more knowledge-seeking. But for mental action, "epistemic" and "pragmatic" aspects converge, because the "exploitation-exploration dilemma" is very small:[9] here, we find an exclusively brain-based form of active, epistemic self-control.

Recall that according to PP, the brain has no direct access to the causal structure of its own body or the external environment — it has to infer them via interoception (i.e. visceroception and proprioception) and exteroceptive sensation. This raises the following question: what parts of the world can be accessed by *neither* exteroceptive *nor* interoceptive predictive processing? In principle, what relevant epistemic target lies outside transductor space and the space of causal interaction via active inference? One general answer is: the brain itself; the neural body. The brain is blind to itself because it has no self-directed receptor system. The brain is not part of any receptive field and, in inferring properties of itself, it cannot use representational hierarchies that bottom out into sensory sheets. Therefore, should it ever be necessary or adaptive for the brain to access more abstract properties of its *own* causal structure, then active inference using the non-neural body cannot help. To be more precise, above a certain level of complexity there will be "inferentially encapsulated"[10] but epistemically relevant causal

9 In more traditional philosophical terminology, if we take the concept of "theoretical intentionality" as referring to all mental states whose content-specifier consists of truth conditions (i.e. directedness towards an epistemic goal, as in mental states of thinking, believing, etc.) and "practical intentionality" as constituted by all states where the content-specifier consists of satisfaction conditions (i.e. directedness towards an action goal, as in mental states of willing, desiring, etc.), then mental action is special in that it falls under both concepts at the same time.

10 I owe this term to Anil Seth. I propose that this type of inferential encapsulation from the non-neural body is very close to the classical Cartesian idea of "the mental". What Descartes really discovered was that there are two fundamentally different layers in the phenomenal self-model of *Homo sapiens*, one in which events are coded as having spatial *and* temporal properties (the internal model of the non-neural body) and one in which events are coded as temporal, but *not* related to a body-centered spatial frame of reference any more (i.e. the internal model of the inferentially encapsulated parts of the neural body).

properties of the neural body which are not characterized by direct interaction with sensory or motor surfaces. On the control or "instrumental" side, this fact becomes a problem if the brain needs to control its own functional profile. Mental action is exactly what achieves this, because it is a form of self-control aimed mainly at effective connectivity, and not at properties of the non-neural body plant. If you will, it is an attempt to solve the problem of "neural disembodiment", an attempt to take autonomous self-control to a new level.

The good regulator theorem (Conant and Ashby 2007) says that the relevant structure must instantiate a model of the system to be controlled, where the system includes the body plus causal interaction space in the world. What exactly would be the system to be controlled in the special case of epistemic agency? It is the level of knowledge the system possesses and the epistemic quality of its world model, in the human case as physically realized by its neural body. Generally speaking, we could say that for epistemic self-control, the *goal state* of the regulator is the optimal level of knowledge the system could possess, the overall epistemic quality of its world model. For the special case of mental epistemic self-control, however, we could redefine the question, and ask: what is the "good epistemic regulator" or non-sensory "model of the system" that is instantiated? It is tempting to say that this would have to be the EAM, because what matters is the overall epistemic quality of the cognitive *self*-model. We should be careful, however, not to over-interpret the conclusion of Conant and Ashby's argument (I am grateful to Michael Anderson for pointing this out). All that follows is that the regulator models its plant, in the sense that its control structures bear systematic isomorphisms to its states. There is no implication that the control system actually uses an explicit model *of* the plant as a separate element of its control architecture. For the special case of epistemic self-regulation as stated by the cognitive affordance hypothesis, this would mean that mind wandering is self-regulation on the mental level *without* an explicit model of the "epistemic plant", whereas mental action is precisely the rare, special case where we actually operate under a conscious and explicit model of ourselves as epistemic agents.

For individual mental actions, we could then say that their *target* is the expected epistemic value of own cognitive states, under counterfactual outcomes. Karl Friston and colleagues write, "Epistemic value is the expected information gain under predicted outcomes" (Friston et al. 2015, p. 6). This, then, would have to be what is maximized. Geoffrey Hinton (2005, p. 1765 quoted from Clark 2016, p. 196) said that a perceptual state can be seen as nothing but "the state of a hypothetical world in which a high-level internal representation would constitute veridical perception". Analogously, we can now say that a *cognitive* state is the state of a hypothetical knowing *self* in which a high-level internal *self*-representation (namely, the EAM) would itself carry high epistemic value because it would constitute veridical self-knowledge. If so, this would explain how EAMs could serve a causal role in a control hierarchy: they could function as explicit representations of the current level of optimization, of more or less successfully achieved epistemic goal states. I propose that for all non-reward-oriented forms of mental action, epistemic value may be the abstract property that constitutes their goal state (what philosophers call the "intentional object"). This leaves us with a first working concept of "epistemic mental action" under the PP approach: it is the predictive control of effective connectivity aimed at optimizing the epistemic value of attentional and cognitive states embedded in the system's phenomenal self-model, accompanied by the emergence of a new content layer on introspectively available levels of the hierarchy, namely, the EAM.

4 Open Questions

Let me end by sketching what I take, on a conceptual level, to be the three most important unresolved issues, perhaps the most relevant targets for future research. They refer to the homunculus problem, the metaphysics of goal-state selection and action initiation, and the possibility of eliminating the concept of "mental action" altogether.

4.1 Problem 1: The Bubble in the Bubble

As long as we model mental action analogously to active inference, we will always introduce a second statistical boundary into the brain's model of reality, a new evidentiary boundary potentially leading to a vicious regress of inferential seclusion (see Hohwy 2016; Hohwy 2017 and Clark 2017 for an excellent discussion). If, as I have proposed, what is predicted and actively controlled via changes of effective connectivity is the epistemic value of attentional and cognitive states as embedded in the self-model, then the *causes* of such shifts in epistemic value will lie hidden behind a Markov blanket — they will constitute an inner environment from which the mental agent is secluded. For example, if the cognitive affordance hypothesis points in the right direction, then this internal environment is constituted by the dynamic affordance landscape set up by the mind wandering network. What we call "thinking" is the process of trying to predict and control the epistemic dynamics unfolding in this network, of "surfing inner uncertainty" (Clark 2016) — the constant mental entropy brought about by the stream of spontaneously occurring thought, unbidden memories, task-unrelated goal-state simulations, and so on.

This threatens to create an acerbated version of the homunculus problem, because it is now tempting to posit a little "epistemic surfer" in our head, a truly disembodied entity that always tries to stay ahead of the next cognitive wave and maximizes evidence for its own existence. Mental action would therefore create a "bubble in the bubble" in the brain's model of reality, where all we need to predict our little mental agent's behavior is information about events from *their* side of the statistical veil, because an observation of events within the smaller bubble of mental agency and prior expectations about the cognitive system will suffice to predict his or her epistemic behavior. From a scientific third-person perspective, all causes beyond the smaller Markov blanket would then be uninformative for the purpose of predicting mental actions and for a scientific understanding our own cognitive phenomenology.

Problem 1 consists in avoiding new versions of the homunculus problem by proliferating evidentiary boundaries, action types, and nested explanatory-evidentiary circles (Hohwy 2016, pp. 5 & 12). We do not want to build epistemological Matryoshka dolls dressed in Markov blankets, and we also want to avoid setting the stage for bizarre, new, and *internalist* versions of the "extended mind debate" (Menary 2010). Therefore, the number of agents should not be multiplied without necessity.

4.2 Problem 2: Dropping Naïve Realism About "Goals"

In the conceptual framework I am proposing, action initiation becomes a functionally adequate form of self-deception. One open question is if this point holds for bodily and mental actions in the same way. Let me explain.

Many neuroscientists use the term "goal" in a naïvely-realistic manner, as if "goals" were something we could in principle find out there in the world. But goals are not given. It is important to understand that, from a scientific third-person perspective, all goal-representations are misrepresentations. Viewed from the philosophical perspective of a thoroughgoing naturalism, there are no intrinsically normative facts, quite simply because there are no states in the physical world that could count as "goal states" in any more rigorous, metaphysical sense. There are no essences, no intrinsic values which remain invariant across all contexts and situations. If you will, the harshness of naturalist metaphysics exactly consists in the point that nothing has *intrinsic* value, because any possible or actual fact is only normative *relative* to a certain organism, biological population, self-sustaining robot, or other such entity. Of course, a naturally evolved information-processing system may *represent* its own procreative success or the sustaining of its own existence as an intrinsically valuable goal state. It may generate an internal model of the world and of itself in it, in which organismic integrity and homeostatic stability *are* ultimate normative facts, and as this model is physically grounded any such process will have direct causal consequences — for example, the system may turn into a "crook-

ed scientist" (Bruineberg et al. 2016, Section 3) and begin maximizing evidence for its own existence (Friston 2011, p. 117; Limanowski and Blankenburg 2013). Clearly, any self-model successfully integrating (mis)representations of (non-existing) goal states as possessing intrinsic normativity will lead to the instantiation of new functional properties, which can then be selected for in the process of natural evolution (Metzinger 2003).

Call this "the self-deception model of goal selection and action initiation". This may refer to an important step in the biological history of mindedness, because to manifest higher levels of intelligent behavior any naturally evolved agent has to solve two major problems: autonomous self-control and autonomous self-motivation. Biological organisms like ourselves solve the second problem by *hallucinating* goals. One aspect that makes the theory of predictive processing attractive from a philosophical perspective is that it offers a concrete mechanism for what has in the past been called "conscious volition" or "deliberate action initiation": we can dis-attend to the current state of the body, thereby enabling proprioceptive predictions — which are currently *false* representations of body and limb position in space — to "become real", initiating a new cycle of active inference. I think that attenuating sensory input by reallocating expectations for prediction-error precisions across different levels of the generative model (see also Limanowski 2017) is exactly what allows us to "hallucinate goals" — blocking out ascending sensory-prediction error while simultaneously setting a dynamic, top-down context in hierarchical motor control by maximizing the precision of (and hence confidence in) a more *abstract* type of belief. That belief may be normative, and it may be false.

Whenever we begin to hallucinate a goal, we are actively optimizing a high-level, multi- or amodal model describing a possible path into counterfactual states of the organism, where "optimizing" means conferring intrinsic value on it. The self-model theory (Metzinger 2003) would say that this crucial step happens when an internal representation of some state of the world as organism-salient — as a positive state of affairs *relative* to homeostasis, procreative success, long-term self-sustainment and so on — turns into an *intrinsic* (i.e. context-invariant) value for the system whenever it is functionally embedded in its intrinsic self-representation, when it becomes integrated with the deeper, phenomenally transparent, and more context-invariant levels of the phenomenal self-model (Blanke and Metzinger 2009). Before movement onset, sensorimotor estimates are misrepresentations. The moment of action initiation is the moment in which the counterfactuality disappears from the content. What begins as an imagined bodily movement gradually becomes a real bodily movement; phenomenal opacity is transformed into phenomenal transparency; an allocentric representation of a successfully terminated action in a larger predictive horizon is gradually transposed into a shorter timescale, until it becomes "embodied" in an egocentric frame of reference leading to overt action. The intrinsic normativity in the original model's content, however, remains a hallucination — it is not an objective property of the successfully terminated motor pattern depicted by the original allocentric representation. In addition, if we hold on to the notion of active inference, then even the low-level sensory precision estimates generated in the process are systematically misrepresentational *and* beneficial for the agent, simply because they enable action. I believe this general conceptual point may be of considerable philosophical relevance, for example, because, in the words of Wanja Wiese, it draws our attention to the fact that "systematically beneficial misrepresentations may lie at the heart of our neural architecture" (Wiese 2016, Section 8).

Problem 2 consists in developing a fine-grained phenomenological analysis of the process just sketched without assuming a realist metaphysics of "goals" and "intrinsic value" — or reintroducing nested Matryoshka dolls and mental homunculi (Problem 1). Obviously, arriving at a deeper understanding of goal-state selection and action initiation is of particular relevance for the special case of *mental* actions and *epistemic* goal states. It would therefore be of importance to the construction of computational models that can later explain the emergence of the specific phenomenological profile going along with transiently "hallucinating intrinsic value".

4.3 Problem 3: Phenomenology and the Metaphysics of Mental Action

It might turn out that the problem of mental action just cannot be solved in a satisfactory manner which simultaneously does justice to all its relevant aspects — the first horn of the dilemma. One classical and respectable option is to *eliminate* the concept of "mental action" from our scientific taxonomy (Churchland 1981; Churchland 1989). "Mental action" would then have to be dissolved into a set of empirically grounded successor concepts. But even if this could be done without shrinking the explanatory scope and predictive power of future theories developed under the PP-approach, it would still leave us with an extremely counterintuitive perspective on our own phenomenology — the second horn of dilemma. Let us take a look at this second horn first.

Here is the sketch of a three-phase model, by which an eliminativist could try to account for our robust phenomenology of mental agency:

- Phase 1 is termed "**Deep Social Embodiment**" (indirectly alluding to the terminology developed by Lisa Quadt, see also Quadt 2017), and it begins with the early prenatal development of the fetus' brain. We often ignore the fact that the human brain finds itself in a situation of "nested double embodiment" right from the origin. The very first social interactions it has to predict and control are shaped by this special, convoluted form of embodiment. Some of these interactions may be motor, but many take place on a molecular level. They are special variants of what Quadt calls interactive inference: the fetus' brain has to predictively control *two* interoceptive environments, namely, a) the stimulus landscape internal to the fetus body itself, and b) the chemical landscape surrounding it, which is in turn predictively controlled by the mother's brain in a process Anil Seth has termed "interoceptive inference". Viewed from the epistemic perspective of the mother's brain, the fetus' body is a robust and new interoceptive signal source. Computationally, there are two overlapping task domains, and we can now view pregnancy as an escalating conflict between two control systems leading to a fundamentally competitive dynamic, ultimately based on the antagonistic coevolution of fetal and maternal genes (e.g. Haig 2015). This leads to the emergence and consolidation of a second self-model in what is still a single, metabolically autonomous organism — but this time in the brain of the fetus — ultimately achieving and culminating in the instantiation of a new phenomenal property — an individual "sense of presence" (Blanke and Metzinger 2009; Seth et al. 2011).

- Phase 2 is "**Amnestic Enculturation**". First, it is important to note that CA as well as AA are enculturated processes (Fabry 2015; Fabry 2017). One relevant cultural practice is that human beings educate their children by ascribing the capacity for veto control ("You *could* have refrained from doing this!") and ultimate origination ("You *could* have done otherwise!") right from the very beginning, and at a stage where, from a scientific perspective, young children almost certainly do not possess these capacities. According to the PP approach, this cultural practice will automatically become internally modelled, and the overall process will influence the gradual development of the child's EAM. All socially embedded PP systems have to extract and exploit the causal structure of their sociocultural niche — they have to predict their caregivers' behavior — and the best resource for the generation and updating of hypotheses they have is their own internal self-model.
  There is empirical evidence demonstrating the social transmission of the experience of agency (Khalighinejad et al. 2016), and not only do children observe adults controlling external and internal events, their social context is constituted by pre-enculturated agents already operating *under the belief* of having the capacities of mental veto control and ultimate origination themselves (Hauf and Prinz 2005). In this way, socially constituted norms for autonomy and successful self-control become internalized via social interactive inference. There is, however, an important second aspect which the eliminativist adherent of the PP approach could highlight: It is only much later, between the ages of 2 and 4 years, that the *autobiographical* self-model comes

onstream. This leads to the well-studied effect of infantile amnesia: the inability of adult human beings to retrieve episodic memories from this period. For the grown-up human being, the process of amnestic enculturation itself will never be introspectively available via the phenomenal self-model.

- I call Phase 3 the period of the "**Agentive Narrative**", because it is dominated by an inner life narrative: the existence of an autobiographical self-model creating an illusion of transtemporal identity (Metzinger 2013a, p. 5). Of course, there really is no such thing as a "narrative self"; our inner life narrative has no author, no mysterious self-behind-the-self that could function as the narrator, it too emerges out of a continuous process of dynamic self-organization and social interaction, leading to an ever-changing autobiographical self-model. But the process just sketched has an important phenomenological effect: when autobiographical memory begins to consolidate, a strong hyperprior for autonomy and moral responsibility is already firmly established. We believe and automatically predict that we (and others) have the capacity for veto control and ultimate origination, because "it has always been like that, since we can remember". We are embodied models of our early social environment. Our phenomenal self-model reflects implicit expectations of our primordial sociocultural niche and says that we have always been autonomous, morally responsible agents. This robust, self-related hyperprior has been installed early through highly reliable social feedback and it is a major factor in determining the phenomenology of self-consciousness and M-autonomy.

Now let us look at our target phenomenon of mental action against this background. According to subjective experience, and as pointed out in Section 2.3, the sudden emergence of M-autonomous epistemic self-control is an unpredicted internal event. Strictly speaking, what we call "agency" refers to an *interpretation* of this fact as "initiation" or "origination": it is the activation of an internal self-model trying to explain away the surprise involved in suddenly achieving autonomous self-control, by creating an explicit, supramodal representation of an entity capable of ultimate origination, veto control, and spontaneous self-causation. It likely has internal and external aspects, because it will involve internal simulations of cultural practices, as explained above. It is therefore something we learn before our conscious inner life-history has even begun. In addition, as Patrick Haggard has pointed out, the phenomenology of intending to act is actually quite thin and evasive, and lacks the vivid subjective quality of, for example, visual experience (Haggard 2005, p. 291). One might certainly argue that if we introspect as closely as we can and on a very fine-grained timescale, there really only is a phenomenological element of "surprise" going along with action initiation. The human adult's conscious model of the "self" involving ultimate origination and self-causation could therefore be an enculturated *post-hoc* confabulation. It could be computationally described as a causal-inference illusion that has become part of a sociocultural niche-model. Ultimately, it is based on an internalization of social interactions and deeply engrained language games that lead to self-evidencing habits of a) shallow introspection, and b) unreflected but "theory-contaminated" phenomenological self-reports. Therefore, the second horn of the dilemma may not be so counterintuitive after all — if we dare to take a closer look.

However, if we really take the option of eliminating the concept of "mental action", then, under PP, we suddenly face the problem of *bodily* action initiation: conventionally, mental actions explain the initiation of overt bodily action via a reweighting of precision expectations from somatosensory sources to goal-representations by generating descending proprioceptive predictions, which I called, in the introduction, "self-fulfilling motor fantasies". If AA is a causally inert phenomenal artifact and only unintentional mental *behaviors* (like the automatic, subpersonal reweighting of precision expectations just mentioned) do exist, then we have to find a different solution. We might say that a) bodily actions do exist, b) mental actions never existed, and c) our picture of bodily action initiation has changed in a way that only leaves us with a very weak and impoverished notion of the term "action". If

— now returning to the first horn of the dilemma — we want to say that mental actions are respectable citizens of a scientifically grounded ontology, then we need a model of *mental action initiation* for the special case of attentional agency, and we need one that avoids a vicious regress. Introducing higher-order precision shifts, for example by positing third-order precision-control targeting processes of second-order precision optimization ("second-order AA"), would create exactly such a vicious regress (Problem 1). It would also give birth to a lineage of nested Matryoshka dolls.

I have therefore proposed a non-circular model for the special case of cognitive agency: the cognitive affordance hypothesis. Cognitive action initiation is explained by the navigation of an internal affordance landscape which is continuously set up by the mind wandering network (Section 3.2). But at present it is not clear if the mind wandering network could play the same role for the initiation of *attentional* action, an equivalent of providing an automatic stream of goal-representations which compete for motor control by inducing sensory attenuation (Problem 2). So we may have a model of action initiation for CA, but not for AA. I take this to be another highly relevant future target for interdisciplinary research.

Problem 3 lies in deciding which horn of the dilemma to take. The realist option continues to treat mental actions as part of our scientific ontology, but bears the explanatory burden of developing a multi-level theory of goal-state selection and action initiation spanning all levels of the hierarchy — ranging from attentional to overt bodily agency — by solving Problem 1 and Problem 2. The eliminativist option has to dissolve the concept of "mental action" into a series of successor concepts possessing at least equal unificatory, predictive, and explanatory power — while making intelligible why, for centuries and in our very own autophenomenological reports, we have described our inner life so falsely.

5 Conclusion

Are there mental actions *at all*? I will not answer this question here. My overarching goal in this article has been to put the problem of mental action into explicit focus. But as we have now seen, this problem has many philosophical and empirical facets. It must be clearly stated that to date, there is no satisfying theoretical account of mental action under the PP approach. I have offered a series of new conceptual instruments and connected them to four central ideas discussed in recent philosophical and scientific research on PP. My aim was to integrate core semantic elements in order to arrive at a first working concept for "mental action", a hypothetical construct which can now be further developed and refined. According to this first notion, mental actions are a specific form of predictive control of effective connectivity, accompanied and possibly even functionally mediated by what I have termed the EAM. They are aimed at increasing the epistemic value of pre-existing states in the conscious self-model, without causally looping through sensory sheets or using the non-neural body as an instrument for active inference. I have sketched four empirically testable hypotheses, assuming that what is most needed at this stage are more fine-grained computational models yielding testable predictions. [11]

11 I am deeply indebted to Jona Vance, Michael Anderson, Carsten Korth, Jakob Hohwy, Giovanni Pezzulo, Iuliia Pliushch, Anil Seth, and Wanja Wiese for critical discussion and a considerable number of extremely helpful comments and Lucy Mayne for equally helpful comments plus excellent editorial help with the English version of this text. I have learned a lot from all of them. They are not responsible for the shortcomings of the final version of this paper.

References

Amer, T., Campbell, K. L. & Hasher, L. (2016). Cognitive control as a double-edged sword. *Trends in Cognitive Sciences, 20* (12), 905–915. https://dx.doi.org/10.1016/j.tics.2016.10.002.

Anderson, M. L. (2015). Précis of After phrenology: Neural reuse and the interactive brain. *Behavioral and Brain Sciences, 16*, 1–22.

Andrews-Hanna, J. R., Reidler, J. S., Huang, C. & Buckner, R. L. (2010). Evidence for the default network's role in spontaneous cognition. *Journal of Neurophysiology, 104* (1), 322–335.

Axelrod, V., Rees, G., Lavidor, M. & Bar, M. (2015). Increasing propensity to mind-wander with transcranial direct current stimulation. *Proceedings of the National Academy of Sciences of the United States of America, 112* (11), 3314–3319. https://dx.doi.org/10.1073/pnas.1421435112.

Barsalou, L. (2016). Can cognition be reduced to action? In A. K. Engel, K. J. Friston & D. Kragic (Eds.) *The pragmatic turn: Toward action-oriented views in cognitive science*.

Blanke, O. & Metzinger, T. (2009). Full-body illusions and minimal phenomenal selfhood. *Trends in Cognitive Sciences, 13* (1), 7–13.

Bortolotti, L. (2015). The epistemic innocence of motivated delusions. *Consciousness and Cognition, 33*, 490–499.

Brass, M. & Haggard, P. (2007). To do or not to do: The neural signature of self-control. *The Journal of Neuroscience, 27* (34), 9141–9145.

Broadway, J. M., Zedelius, C. M., Mooneyham, B. W., Mrazek, M. D. & Schooler, J. W. (2015). Stimulating minds to wander. *Proceedings of the National Academy of Sciences of the United States of America, 112* (11), 3182–3183. https://dx.doi.org/10.1073/pnas.1503093112.

Bruineberg, J. (2017). Active inference and the primacy of the 'I can'. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*.

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

Buckner, R. L. & Carroll, D. C. (2007). Self-projection and the brain. *Trends in Cognitive Sciences, 11* (2), 49–57. https://dx.doi.org/10.1016/j.tics.2006.11.004.

Buckner, R. L., Andrews–Hanna, J. R. & Schacter, D. L. (2008). The brain's default network. *Annals of the New York Academy of Sciences, 1124* (1), 1–38.

Butz, M. V. (2016). Toward a unified sub-symbolic computational theory of cognition. *Frontiers in Psychology, 7*, 925. https://dx.doi.org/10.3389/fpsyg.2016.00925.

Campbell-Meiklejohn, D. K., Woolrich, M. W., Passingham, R. E. & Rogers, R. D. (2008). Knowing when to stop: The brain mechanisms of chasing losses. *Biological Psychiatry, 63* (3), 293–300.

Christoff, K. (2012). Undirected thought: Neural determinants and correlates. *Brain Research, 1428*, 51–59. https://dx.doi.org/10.1016/j.brainres.2011.09.060.

Christoff, K., Gordon, A. M., Smallwood, J., Smith, R. & Schooler, J. W. (2009). Experience sampling during fMRI reveals default network and executive system contributions to mind wandering. *Proceedings of the National Academy of Sciences, 106* (21), 8719–8724.

Christoff, K., Irving, Z. C., Fox, K. C. R., Spreng, R. N. & Andrews-Hanna, J. R. (2016). Mind-wandering as spontaneous thought: A dynamic framework. *Nature Reviews Neuroscience, 17*, 718–731. https://dx.doi.org/10.1038/nrn.2016.113.

Churchland, P. M. (1981). Eliminative materialism and the propositional attitudes. *The Journal of Philosophy, 78* (2), 67–90.

——(1989). *A neurocomputational perspective: The nature of mind and the structure of science.* MIT Press.

Cisek, P. (2007). Cortical mechanisms of action selection: The affordance competition hypothesis. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 362* (1485), 1585–1599.

Cisek, P. & Kalaska, J. F. (2005). Neural correlates of reaching decisions in dorsal premotor cortex: Specification of multiple direction choices and final selection of action. *Neuron, 45* (5), 801–814.

—— (2010). Neural mechanisms for interacting with a world full of action choices. *Annual Review of Neuroscience, 33*, 269–298.

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

——(2017). How to knit your own Markov blanket: Resisting the second law with metamorphic minds. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*.

Cohen, O., Koppel, M., Malach, R. & Friedman, D. (2014). Controlling an avatar by thought using real-time fMRI. *Journal of Neural Engineering, 11* (3), 035006.

Cohen, O., Druon, S., Lengagne, S., Mendelsohn, A., Malach, R., Kheddar, A. & Friedman, D. (2014). FMRI-based robotic embodiment: Controlling a humanoid robot by thought using real-time fMRI. *Presence: Teleoperators and Virtual Environments, 23* (3), 229–241. https://dx.doi.org/10.1162/PRES_a_00191.

Cole, M. W. & Schneider, W. (2007). The cognitive control network: Integrated cortical regions with dissociable functions. *NeuroImage, 37* (1), 343–360.

Cole, M.W., Yarkoni, T., Repovš, G., Anticevic, A. & Braver, T. S. (2012). Global connectivity of prefrontal cortex predicts cognitive control and intelligence. *The Journal of Neuroscience, 32* (26), 8988–8999.

Cole, M. W., Reynolds, J.R., Power, J.D., Repovs, G., Anticevic, A. & Braver, T.S. (2013). Multi-task connectivity reveals flexible hubs for adaptive task control. *Nature Neuroscience, 16* (9), 1348–1355.

Cole, M. W., Repovs, G. & Anticevic, A. (2014). The fronto-parietal control system: A central role in mental health. *The Neuroscientist*.

Conant, R. C. & Ashby, R. (2007). Every good regulator of a system must be a model of that system. *International Journal of Systems Science, 1* (2), 89–97. https://dx.doi.org/10.1080/00207727008920220.

Davey, C. G., Pujol, J. & Harrison, B. J. (2016). Mapping the self in the brain's default mode network. *NeuroImage, 132*, 390–397. https://dx.doi.org/10.1016/j.neuroimage.2016.02.022.

Davidson, D. (2001). *Essays on actions and events: Philosophical essays.* Oxford: Oxford University Press.

Dretske, F. (1986). Misrepresentation. In R. Bogdan (Ed.) *Belief: Form, content, and function* (pp. 17–36). Oxford, Oxford University Press.

——— (1988). *Explaining behavior: Reasons in a world of causes.* Cambridge: Cambridge University Press.

Engel, A. K., Maye, A., Kurthen, M. & König, P. (2013). Where's the action? The pragmatic turn in cognitive science. *Trends in Cognitive Sciences, 17* (5), 202–209.

Engel, A. K., Friston, K. J. & Kragic, D. (2016). *The pragmatic turn: Toward action-oriented views in cognitive science.*

Fabry, R. E. (2015). Enriching the notion of enculturation: Cognitive integration, predictive processing, and the case of reading acquisition. In T. K. Metzinger & J. M. Windt (Eds.) *Open MIND*. https://dx.doi.org/10.15502/9783958571143.

———(2017). Predictive processing and cognitive development. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing.*

Filevich, E., Kühn, S. & Haggard, P. (2012). Intentional inhibition in human action: The power of 'no'. *Neuroscience & Biobehavioral Reviews, 36* (4), 1107–1118.

———(2013). There is no free won't: Antecedent brain activity predicts decisions to inhibit. *PLoS ONE, 8* (2), e53053.

Fox, K. C. R., Spreng, R. N., Ellamil, M., Andrews-Hanna, J. R. & Christoff, K. (2015). The wandering brain: Meta-analysis of functional neuroimaging studies of mind-wandering and related spontaneous thought processes. *NeuroImage, 111*, 611–621.

Friston, K. (2011). Embodied inference: Or I think therefore I am, if I am what I think. *The implications of embodiment (Cognition and Communication)*, 89–125.

Friston, K., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T. & Dolan, R. J. (2014). The anatomy of choice: Dopamine and decision-making. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 369* (1655). https://dx.doi.org/10.1098/rstb.2013.0481.

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., FitzGerald, T. & Pezzulo, G. (2015). Active inference and epistemic value. *Cognitive Neuroscience, 6* (4), 187–214. https://dx.doi.org/10.1080/17588928.2015.1020053.

Haggard, P. (2005). Conscious intention and motor cognition. *Trends in Cognitive Sciences, 9* (6), 290–295. https://dx.doi.org/10.1016/j.tics.2005.04.012.

Haig, D. (2015). Maternal-fetal conflict, genomic imprinting and mammalian vulnerabilities to cancer. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 370* (1673). https://dx.doi.org/10.1098/rstb.2014.0178.

Hauf, P. & Prinz, W. (2005). The understanding of own and others' actions during infancy: "You-like-me" or "Me-like-you"? *Interaction Studies, 6* (3), 429–445.

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

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

——— (2017). How to entrain your evil demon. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing.*

Khalighinejad, N., Bahrami, B., Caspar, E. A. & Haggard, P. (2016). Social transmission of experience of agency: An experimental study. *Frontiers in Psychology, 7* (974), 313. https://dx.doi.org/10.3389/fpsyg.2016.01315.

Kühn, S., Haggard, P. & Brass, M. (2009). Intentional inhibition: How the "veto–area" exerts control. *Human Brain Mapping, 30* (9), 2834–2843.

Lamme, V. A. F. (2003). Why visual attention and awareness are different. *Trends in Cognitive Sciences, 7* (1), 12–18.

Lenggenhager, B., Tadi, T., Metzinger, T. & Blanke, O. (2007). Video ergo sum: Manipulating bodily self-consciousness. *Science, 317* (5841), 1096–1099. https://dx.doi.org/10.1126/science.1143439.

Limanowski, J. (2017). (Dis-)attending to the body. Action and self-experience in the active inference framework. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*.

Limanowski, J. & Blankenburg, F. (2013). Minimal self-models and the free energy principle.

Mantini, D. & Vanduffel, W. (2012). Emerging roles of the brain's default network. *The Neuroscientist, 19* (1), 76–87. https://dx.doi.org/10.1177/1073858412446202.

Margulies, D. S., Ghosh, S. S., Goulas, A., Falkiewicz, M., Huntenburg, J. M., Langs, G., Bezgin, G., Eickhoff, S. B., Castellanos, F. X. & Petrides, M. (2016). Situating the default-mode network along a principal gradient of macroscale cortical organization. *Proceedings of the National Academy of Sciences*, 12574–12579.

Mason, M. F., Norton, M. I., Van Horn, J. D., Wegner, D. M., Grafton, S. T. & Macrae, C. N. (2007). Wandering minds: The default network and stimulus-independent thought. *Science, 315* (5810), 393–395.

Mattar, M. G., Cole, M. W., Thompson-Schill, S. L. & Bassett, D. S. (2015). A functional cartography of cognitive systems. *PLoS Computational Biology, 11* (12), e1004533.

Medea, B., Karapanagiotidis, T., Konishi, M., Ottaviani, C., Margulies, D., Bernasconi, A., Bernasconi, N., Bernhardt, B. C., Jefferies, E. & Smallwood, J. (2016). How do we decide what to do? Resting-state connectivity patterns and components of self-generated thought linked to the development of more concrete personal goals. *Experimental Brain Research*. https://dx.doi.org/10.1007/s00221-016-4729-y.

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

Metzinger, T. (1995). *Conscious experience.* Exeter, UK: Imprint Academic.

——— (2003). *Being no one: The self-model theory of subjectivity.* Cambridge, MA: MIT Press.

——— (2004). Précis of "Being no one". *PSYCHE - An Interdisciplinary Journal of Research on Consciousness, 11* (5), 1–35. http://psyche.cs.monash.edu.au/symposia/metzinger/precis.pdf.

——— (2006). Conscious volition and mental representation: Towards a more fine-grained analysis. *Disorders of Volition*, 19–48.

——— (2008). Empirical perspectives from the self-model theory of subjectivity: A brief summary with examples. *Progress in Brain Research, 168*, 215–278.

——— (2013a). The myth of cognitive agency: Subpersonal thinking as a cyclically recurring loss of mental autonomy. *Frontiers in Psychology, 4*, 931. https://dx.doi.org/10.3389/fpsyg.2013.00931.

——— (2013b). Why are dreams interesting for philosophers? The example of minimal phenomenal selfhood, plus an agenda for future research. *Frontiers in Psychology, 4*. https://dx.doi.org/10.3389/fpsyg.2013.00746.

——— (2015). M-autonomy. *Journal of Consciousness Studies, 22* (11-12), 270–302.

——— (2017). Why is mind wandering interesting for philosophers? In K. C. Fox & K. Christoff (Eds.) *The Oxford handbook of spontaneous thought*.

Mooneyham, B. W. & Schooler, J. W. (2013). The costs and benefits of mind-wandering: A review. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale, 67* (1), 11–18. https://dx.doi.org/10.1037/a0031569.

Niendam, T. A., Laird, A. R., Ray, K. L., Dean, Y. M., Glahn, D. C. & Carter, C. S. (2012). Meta-analytic evidence for a superordinate cognitive control network subserving diverse executive functions. *Cognitive, Affective, & Behavioral Neuroscience, 12* (2), 241–268.

Pezzulo, G. (2012). An active inference view of cognitive control. *Frontiers in Psychology, 3*, 478. https://dx.doi.org/10.3389/fpsyg.2012.00478.

——— (2016). The contribution of pragmatic skills to cognition and its development: Common perspectives and disagreement: Common perspectives and disagreement. In A. K. Engel, K. J. Friston & D. Kragic (Eds.) *The pragmatic turn: Toward action-oriented views in cognitive science* (pp. 19–33).

Pezzulo, G., Cartoni, E., Rigoli, F., Pio-Lopez, L. & Friston, K. (2016). Active inference, epistemic value, and vicarious trial and error. *Learning & Memory, 23* (7), 322–338.

Picard, F. (2013). State of belief, subjective certainty and bliss as a product of cortical dysfunction. *Cortex; A Journal Devoted to the Study of the Nervous System and Behavior, 49* (9), 2494–2500. https://dx.doi.org/10.1016/j.cortex.2013.01.006.

Picard, F., Scavarda, D. & Bartolomei, F. (2013). Induction of a sense of bliss by electrical stimulation of the anterior insula. *Cortex; A Journal Devoted to the Study of the Nervous System and Behavior, 49* (10), 2935–2937. https://dx.doi.org/10.1016/j.cortex.2013.08.013.

Pliushch, I. (2017). The overtone model of self-deception. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*.

——— (in press). *Self-deception within the predictive coding framework*.

Putnam, H. (1967). Psychological predicates. In W. H. Capitan & D. D. Merrill (Eds.) *Art, mind, and religion* University of Pittsburgh Press.

——— (1975). *Mind, language and reality (vol. 2)*. Cambridge: Cambridge University Press.

Quadt, L. (2017). Action-oriented predictive processing and social cognition. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*.

Schooler, J. W., Smallwood, J., Christoff, K., Handy, T. C., Reichle, E. D. & Sayette, M. A. (2011). Meta-awareness, perceptual decoupling and the wandering mind. *Trends in Cognitive Sciences, 15* (7), 319–326.

Seli, P., Risko, E. F., Smilek, D. & Schacter, D. L. (2016). Mind-wandering with and without intention. *Trends in Cognitive Sciences, 20* (8), 605–617.

Seth, A. K. (2015a). Inference to the best prediction. In T. K. Metzinger & J. M. Windt (Eds.) *Open MIND*. https://dx.doi.org/10.15502/9783958570986. http://open-mind.net/papers/inference-to-the-best-prediction.

——— (2015b). The cybernetic Bayesian brain. In T. K. Metzinger & J. M. Windt (Eds.) *Open MIND*. https://dx.doi.org/10.15502/9783958570108.

Seth, A. K., Suzuki, K. & Critchley, H. D. (2011). An interoceptive predictive coding model of conscious presence. *Frontiers in Psychology, 2*.

Smallwood, J. & Schooler, J. W. (2015). The science of mind wandering: Empirically navigating the stream of consciousness. *Annual Review of Psychology, 66*, 487–518. https://dx.doi.org/10.1146/annurev-psych-010814-015331.

Spratling, M. W. (2016). Predictive coding as a model of cognition. *Cognitive Processing*, 1–27.

Spreng, R. N., Mar, R. A. & Kim, A. S. N. (2009). The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: A quantitative meta-analysis. *Journal of Cognitive Neuroscience, 21* (3), 489–510. https://dx.doi.org/10.1162/jocn.2008.21029.

Spreng, R. N., Stevens, W. D., Chamberlain, J. P., Gilmore, A. W. & Schacter, D. L. (2010). Default network activity, coupled with the frontoparietal control network, supports goal-directed cognition. *NeuroImage, 53* (1), 303–317. https://dx.doi.org/10.1016/j.neuroimage.2010.06.016.

Stawarczyk, D., Cassol, H. & D'Argembeau, A. (2013). Phenomenology of future-oriented mind-wandering episodes. *Frontiers in Psychology, 4*, 425.

Van de Cruys, S. (2017). Affective value in the predictive mind. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*.

Vandenbroucke, A. R. E., Fahrenfort, J. J., Sligte, I. G. & Lamme, V. A. F. (2014). Seeing without knowing: Neural signatures of perceptual inference in the absence of report. *Journal of Cognitive Neuroscience, 26* (5), 955–969. https://dx.doi.org/10.1162/jocn_a_00530.

Voss, U., Holzmann, R., Hobson, A., Paulus, W., Koppehele-Gossel, J., Klimke, A. & Nitsche, M. A. (2014). Induction of self awareness in dreams through frontal low current stimulation of gamma activity. *Nature Neuroscience, 17* (6), 810–812. https://dx.doi.org/10.1038/nn.3719.

Weissman, D. H., Roberts, K. C., Visscher, K. M. & Woldorff, M. G. (2006). The neural bases of momentary lapses in attention. *Nature Neuroscience, 9* (7), 971–978.

Wiese, W. (2016). Action is enabled by systematic misrepresentations. *Erkenntnis*. https://dx.doi.org/10.1007/s10670-016-9867-x.

Wilson, G. & Shpall, S. (2016). Action. In E. N. Zalta (Ed.) *The Stanford encyclopedia of philosophy* Metaphysics Research Lab, Stanford University.

Windt, J. M. (2015). *Dreaming: A conceptual framework for philosophy of mind and empirical research*. Cambridge, MA: MIT Press.