ψυχη
PSYCHE

The Sense of Self in the Phenomenology of Agency and Perception

Jakob Hohwy
Department of Philosophy
Monash University
Clayton, VIC 3800
Australia
Jakob.Hohwy@arts.monash.edu.au

PSYCHE 13/1, April 2007

KEYWORDS: agency, visual perception, minimal self, narrative self, generative models, predictive coding, default mode, delusions, psychopathology

Abstract: The phenomenology of agency and perception is probably underpinned by a common cognitive system based on generative models and predictive coding. I defend the hypothesis that this cognitive system explains core aspects of the sense of having a self in agency and perception. In particular, this cognitive model explains the phenomenological notion of a minimal self as well as a notion of the narrative self. The proposal is related to some influential studies of overall brain function, and to psychopathology. These elusive notions of the self are shown to be the natural upshots of general cognitive mechanisms whose fundamental purpose is to enable agents to represent the world and act in it.

1. Introduction

There is a growing awareness that the phenomenology of agency and of perception is associated with a sense of a *minimal self*: the pre-reflective feeling that a given movement is performed by *me*, or that a given perceptual experience is had by *me*. This reference to self is distinguished from the autobiographical sense of having a *narrative self* that persists across experiences. The minimal self is more like an instantaneous feeling of "mineness" with which experiences are labelled. Recognition of the minimal self is inspired by Husserlian phenomenology (Zahavi 2006) and, though somewhat elusive, it seems introspectively valid to say that our experiences come furnished with such a background feeling of mineness (for an influential review, see Gallagher 2000; see below for a fuller characterisation).

In the case of the phenomenology of agency, there is reason to believe that the sense of minimal self is reductively explained by a cognitive mechanism based on generative models. I develop and defend the hypothesis that the same kind of cognitive mechanism can reductively explain the sense of minimal self in visual perception. I then show how this proposal integrates well with findings on overall brain functioning relating to the sense of having a narrative self. Finally, I contrast this account with psychopathological experience.

To prepare the ground for a reductive account of core aspects of the self in agency and perception, I outline three core cognitive tasks that seem intimately related with the self:

> *Self in agency and bodily movement.* An individual needs to be able to generate and intimately track motor commands in accordance with her desires and beliefs about the world. There must be a distinction available between changes in her body and in the environment that are due to her own agency and those changes that are due to other factors in the environment or her sensorimotor system. This agency-based distinction between the doings of the mind and of the world relates to the sense of self in agency.
>
> *Self in perception.* The individual also needs to be able to represent the real environmental causes of noisy, context-dependent sensory signals. This is in part a reality testing competence ("is the cause of the sensory input really what it seems?") that grounds a distinction between representation of persistent external objects and the noisy contributions of the sensory system of a perceiving self.
>
> *Self in planning and attention.* The individual needs to be able to distinguish between those environmental causes that she should attend to and act on and those that are not relevant to her current tasks, plans and preferences. This attentional distinction between the world that would afford her plans and the world that wouldn't relates to the sense of a cohesive, extended self.

The importance of these tasks to the self can be seen by imagining how the self would be compromised if one had persistent difficulties with them. One would tend to lose track (i) of one's body as a locus of mental causation and volition; (ii) of how the world is and how it seems—of where the mind ends and the world begins; and (iii) of one self as a reasonably cohesive person able to discern and prioritise tasks and to attend to salient features of the world.

A number of people have proposed that a cognitive system based on generative models used for forward modelling of re-afferent signals contributes to the sense of self associated with the phenomenology of agency in bodily movement (for reviews, see de Vignemont and Fourneret 2004; Frith 2005; Gallagher 2005: Ch 8). This cognitive system may be a particular instance of the brain's overall cognitive system for representation and attention based more generally on generative models and predictive coding (for reviews, see Friston 2002; Kersten *et al*. 2004). This gives rise to the hypothesis, to be explored and defended here, that this overall cognitive system underpins the three cognitive tasks I mentioned and, further, that properties of these cognitive mechanisms explain some of the core properties of the conscious sense of self associated with perception and attention, as well as agency.[1]

The conclusion is that core properties of the sense of self are underpinned by properties of a unified fundamental cognitive brain system. This approach to the nature of the self reveals the sense of minimal self as a sense of already being familiar with new sensory input, which is sustained by predicting what happens and, for the narrative self, of a see-saw between pondering one's role in a given task and forgetting oneself in the task. The self in agency and perception transpires as a predicting and pondering self.

Section 2 reviews approaches that relate self and agency through forward modelling cognitive systems and explains how a sense of minimal self in agency may arise. Section 3 describes how this general framework is extended to perception (my example will be visual perception, other sensory modalities may be treated similarly) and demonstrates how it explains properties that from the perspective of a conscious subject is a pre-reflective, minimal self. Section 4 extends the generative model perspective to findings on whole brain function as it relates to attention and rest, and shows how this contributes to the experience of a more extended, reflective and pondering self. Section 5 considers some psychopathological aspects of this proposal.

2. The Predicting Self in Agency

In order to track one's causal doings in the world and to track what other people and external events do to one's body, individuals have a need to accurately distinguish self-initiated movement from externally initiated movement. The evidence simply that movement has occurred is not sufficient on its own to decide between the hypothesis that the movement was initiated by one-self or externally. If one has an intention to move that corresponds to the movement, then that is some evidence that the movement was self-initiated. However, though a movement may correspond to the subject's intentions there is still a need to test whether it was the individual herself that acted on the intentions or something else that coincided with the intentions (de Vignemont and Fourneret 2004). Intentions, in any case, can be crude and unreliable evidence for self-ascription of mental causation (Bargh and Chartrand 1999; Wegner 2002; for a philosophical review, see Hohwy 2004a). Moreover, the need to distinguish one's own observed movement from other's movement is complicated by the fact that own and other's movements enjoy partially shared representations primarily in the motor system (Blakemore and Frith 2003; Jeannerod and Pacherie 2004). It therefore seems that something further than movement representation and intention matching is needed.

One way for the cognitive system to deal with these needs is to operate with efference copies of its motor commands (Frith 1992; Blakemore, Wolpert *et al*. 2002; Frith 2005). Efference copies of motor commands can be used in forward modelling of the movement. This allows accurate prediction of the re-afferent sensory (e.g., visual and proprioceptive) consequences of action. These predictions can be compared (i) to the intended state, and (ii) to the actual sensory consequences of the movement. Good predictions cancel out or attenuate the other input and produces less error signal than bad predictions.

Comparisons with the intended state allow fast central error correction in case the forward model reveals that execution of the movements will not correspond to the desired state (Frith and Done 1989). Comparisons with the actual state allows fast movement correction in case the modelled movements do no match the actual re-afferent input of the ongoing movement, for instance in the case of unexpected obstacles in the trajectory of one's hand. In both cases, fast on-line control of the movement is made possible and enables the individual to control movements without having to consciously monitor the environment and make post hoc adjustments.

Self-initiation of intended movement can also be represented in virtue of how it, in contrast to externally initiated movement (e.g., when someone pushes you), is accompanied by forward models. In self-initiated movement, forward models are compared with the re-afferent signal, when the fit is good the incoming signal is attenuated such that only a relatively little error signal is transmitted forwards in the system. This enables the individual to distinguish self-initiated from externally initiated movement since the latter is not associated with control and attenuation of re-afferent feed-back. The individual is then able to represent and track herself and her mental states as a causal influence in the world and keep track of when bodily movements result from other people interfering with her body.

As noted in the introduction above, it seems that there is an immediate, pre-reflective experience of minimal self or "mineness" as one voluntarily performs bodily movements (at least in non-pathological cases). A central aspect of this experience of mineness is the sense of agency such that I experience the movement as intended, initiated and controlled *by me* (for discussion, see Stephens and Graham 2000; de Vignemont and Fourneret 2004; Gallagher 2005: Ch 8). Thus:

> [T]he sense of agency is conceptualized as a transient, in-the-moment type of self-awareness: when performing an action, one represents the action as one's own and under one's control. It seems natural to classify this as an aspect of the ecological or minimal self (Boyer *et al*. 2005).

I may sometimes have to reflect and use inference on the basis of, for example, visual evidence to evaluate the consequences and ultimate success of my intended actions. But the experience of mineness in agency seems an integrated, non-inferential and "on-line" aspect of the phenomenology of agency itself, even when the controlled movement doesn't ultimately have the desired consequences. This feeling is not associated with the experience of having a perduring self that occurs—more or less as the same—across different controlled bodily movements. Rather, each individual experience of agency comes with a feeling, in the specious present, that this is *my* movement. It is this sense of minimal self associated with mineness in agency that we now explain with in terms of properties of the forward modelling system.

The sense of agency is partly based on predicting the sensory consequences of movement (de Vignemont and Fourneret 2004; Hohwy and Frith 2004). When predicted signals are attenuated, only the discrepancy between the prediction and the input is sent forward as an error signal, so when things go well the incoming signal will not be informationally salient, relative to less well predicted signals. Counterintuitively, then, the immediate conscious sense of being in control during agency transpires as the feeling of *not* having to perform active on-line control of what we predict well. It is literally non-observational and immediate in as much as it is being free from attending to incoming signals. This is in distinction to the experience of other people's movement as well as passive movement of one's own limbs. Here we have a candidate for the cognitive mechanism that gives rise to mineness. In terms of conscious experience, mineness is the feeling of already being familiar with the movement's sensory consequences when they actually occur, we are so to speak already 'at home' in the movement because the incoming signals are predicted (I explain the notion of 'familiarity' further in the next section). In contrast, externally initiated movement cannot be predicted in the same way and we cannot be 'at home' in them in the same way. In this way it is possible to explain why the pre-reflective experience of the minimal self in agency has one kind of conscious feel rather than another (for elaboration, see Hohwy and Frith 2004).

The other side of the feeling of control is based on matching motor intentions for movement with the forward model. Here too a sense of mineness can arguably arise. When there are discrepancies between intentions and model, the model fails to match or "predict" the intended state. For relatively small discrepancies, adjustments must be made to the model. If the discrepancy is large enough it may be because no feasible model can be found. In that case, the subject may have to revise her intentions and try again. Assume now that, just as we process the sensory consequences of movement through comparisons with the model, we process the intended state through comparisons with the model. Then the subject will feel mineness or 'at home' in her intentions in so far as they are matched by the model. In contrast, the feeling of mineness is compromised when the model fails to match the intentions. In that case the intentions accrue informational salience to her and any consequent movement may be disowned as being on 'auto-pilot'. The anarchic hand sign, where patients report that their hand moves according to its own intentions (Marchetti and Della Salla 1998) may be an example where the model fails to be updated in the light of discrepancies with the intentions.

Notice that I construe processing of our intentions analogously to processing of the sensory consequences of action. Our intentions are input, or driving signal, to the model, just as the sensory signal is driving input to the model. By extension, we become aware of our intentions by comparing them with the model. This is controversial because it implies that we are not aware of our intentions for movement until just before we begin acting on them. It is however consistent with this that we are aware of our longer-term intentions and whether our actions are in accordance with them (Hohwy 2004a). The construal is consistent with Libet's famous experiments (Libet, Gleason et al. 1983; Libet 1985), and developments of them (Haggard and Eimer 1999), which shows that we become aware of the "urge" to move well after the brain has begun preparing to perform the movement, and we become aware of having initiated the movement before it has actually begun.

This account of mineness is related to the temporal aspects of experience that Husserlian phenomenological analysis reveals:

> any moment of experience contains a retentional reference to past moments of experience, a current openness (primal impression) to what is present, and a protentional anticipation of the moments of experience that are just about to happen (Gallagher and Zahavi 2005: Sec. 3; see also Zahavi 2006: Ch 3; Grush forthcoming).

The classic example of this temporal structure is listening to music, where one does not just hear one note at a time but somehow at any moment incorporates both past notes and anticipates future ones in consciousness. In these terms, my account links mineness in agency, and as we shall see below, in perception too, specifically to the protentional anticipation of what is about to happen. Gallagher (2005: Ch 8) explores this link too. My account differs, however, because the system needs not just to anticipate what is going to happen but also to compare the anticipation to what actually happens; mineness is then the feeling of familiarity with the incoming signal constituted by its attenuation. In addition, the account incorporates a sort of anticipation and comparison for own intentions.

This account of the pre-reflective minimal self pertains to the sense of self in the phenomenology of agency. I now explore the sense of minimal self in visual perception by showing how the forward modeling account of agency is an application of a general neurocomputational principle.

3. The Predicting Self in Perception

Underlying the focus on forward models in the motor control system is a computational problem. If an individual intends to grasp a cup, this goal can be attained in an indefinite number of ways combining the many (i.e., at least 2^{600}) ways our muscles can be activated with the many possible trajectories through space one can take to arrive at the cup. This is the inverse problem of trying to reason from (here, desired) effects to causes. The sensorimotor system overcomes it, as we have just seen, by employing forward or generative models of the sensory consequences of movement. Instead of arriving at the series of motor commands solely by an inference from a representation of the desired effects to the causes that will bring about the effect, it works the other way, by an inference from the model based on a series of motor commands to predictions of the sensory consequences, were that series to be executed. If the consequences fit the desired state (the effect), then that is probably a good series of commands to execute.

This reliance on generative models and prediction is very likely also found in our perceptual systems. From inside the cranium, the brain must figure out what the external causes of its sensory input are. It is however exceedingly difficult to compute causes from effects because the same effects may in principle be caused in indefinitely many different ways and the same causes may in principle have indefinitely many different effects. This is again the inverse problem. One way to overcome this problem is to begin with hypotheses about the cause, and then use those hypotheses to generate predictions that could explain the input. Hypotheses that best explain the input are relatively probable and the system can then infer what the most probable cause is. This static description of the use of generative models can, using a predictive coding scheme, be expanded in temporal and spatial terms such that the predictions concern what would happen next or what would be present in other parts of the visual field, given a certain hypothesis about the cause. Thus a system that works with generative models and predictive coding can overcome the inverse problem and come to represent its environment (for various approaches and overviews, see Friston 2002; Rao *et al.* 2002; Eliasmith 2003; Friston 2003; Grush 2004; Kersten *et al.* 2004; Friston 2005). Put simply, if your input looks a bit like a bicycle, then you can test whether it is really caused by a bicycle by working on the hypothesis that it is a bicycle, and then generate predictions about what the actual input, the sensory consequences, would be should one move to the left around the supposed object. If the predictions are confirmed, then the probability that it is a bicycle will increase. The bicycle hypothesis best explains the evidence in the form of a given series of sensory input. On the other hand, if it is not a bicycle after all (if it is a picture of one, say) then the predictions will not be confirmed and a lower probability will be assigned to this hypothesis. (This idea of perceptual inference goes back to at least Helmholtz (1860), see also (MacKay 1956; Neisser 1967; Gregory 1980; Mumford 1992)).

On a predictive coding scheme, the cognitive system implemented in the brain is hierarchical such that relatively higher levels represent hypotheses about probable causes and issues predictions about future sensory input backwards to lower levels in the system. If the predictions are good then the incoming signal is attenuated such that only the error signal is propagated forwards in the system. The error signal is then used, together with lateral connections, to update the hypothesis so as to generate even better predictions. There is therefore considerable computational similarity to the forward modeling account of the sensorimotor system described above. Notice also that this interconnectivity occurs between any two levels and that the hypotheses of one level can serve as input to the next level up; in this way predictive coding can enable percepts to be built up in stages beginning with very primitive components such as lines at different orientations and ending with very complex, category-specific percepts such as bicycles (see, e.g., Rao and Ballard 1999; Murray *et al*. 2002).

Some of the salient properties of a predictive coding scheme can be appreciated in terms of *perceptual reality testing* (see Hohwy and Rosenberg 2005). The scheme allows a distinction between how things *seem* on the basis of some sensory input and how things *are* in the light of testing hypotheses by predicting future input. A given sensory input is, viewed on its own, a representation of how the world seems to be. For example, in the case of the bicycle inference, the subject will be able to test her initial belief about how the world seems against reality. If in contrast there were no inferential link between the sensory inputs at different moments, the individual would have available only a disjoint series of seemings; the contribution of noise in the individual's own sensory system, and in the world, to the representation could not be assessed. Generative models and predictive coding therefore enable a distinction between a persistent external world and more transient, subjective seemings. (Notice that the distinction is not between imagery and perception, it is a distinction that concerns how we manage the transition from non-veridical perception to veridical perception.)

The ability to draw this kind of distinction between how things seem and how they are is, as I mentioned in the introduction, highly relevant to the sense of self. If this ability is impaired, one would tend to lose track of, so to speak, where the mind ends and the world begins (for the application of this to psychopathology, see Section 5). This allusion to the self suggests that this system is a candidate for the cognitive mechanism that explains the sense of a minimal self in perception. The basic idea is that, as one manages to make the transition to veridical perception, the sensory input becomes better and better predicted, and one therefore acquires a feel of familiarity with the offerings of the sensory systems. If what is familiar is in some sense mine, then this could be what explains the feeling of mineness. In a slogan: as you gain the world you gain a sense of self. I now spell out this proposal, and in particular the notion of familiarity, in more detail.

From a phenomenological perspective, many of the experiences that we saw are associated with the sense of minimal self in the phenomenology of agency recur in the phenomenology of visual perception. The minimal self in perception is thus also characterised in terms of a quality of mineness of conscious experience which is an immediate, pre-reflective and non-observational access to one self.

> All […] experiences are given (at least tacitly) as *my* experiences, as experiences *I* am undergoing or living through (Gallagher and Zahavi 2005; for an nice description, see Zahavi 2006: 124-132).

It is a non-objectified self that is somehow given in ordinary conscious experience. It constitutes the sense of mineness that is a recurrent constant through changes in the stream of consciousness and that immediately reveals one's experiences as one's own (Zahavi 2006: 124). It is that aspect of the self that remains when one abstracts away from the experience of temporally extended, 'narrative' personal identity, which is partially sustained by memory of past events (Gallagher 2000). Though this description grasps at a somewhat elusive experience that is difficult to describe and seems to belong to the subtle background of experience (Zahavi 2006; Grush forthcoming), it does seem that experience is in fact infused with more than just the bare, freely floating representations of the world: it seem reasonable to say that, as we have perceptual experience, it is given in a mode such that we are aware that this is our own experience (see also Shoemaker 1963: Ch 3).

As I indicated above, this experience of the minimal self in perception is explained by properties associated with generative models and predictive coding. A feeling of mineness requires some kind of cognitive frame of reference in which to place the experience. This frame of reference cannot however be an autobiographical *narrative*, which is meant to be distinct from the minimal self. It must somehow be able to play its role in the specious present. We must therefore go to a deeper cognitive level to discern a frame of reference required for mineness. In accordance with the predictive coding scheme set out above, and its interpretation in terms of reality testing, I propose that we interpret the experience of perceptual mineness as the experience of having predicted, and thus already being familiar with, what one perceives. It is mineness in the sense in which what is experienced as familiar is felt to be mine, even when we cannot place the familiar object in a concrete autobiographical narrative. The contrast is when the perceived causes in the environment are deeply unexpected. Then the feeling of mineness would be replaced by a feeling of bewilderment and alienation towards the offerings of the sensory system—the minimal self would begin to fragment (see also Section 5 on psychopathology). The proposal fits with the non-observational, pre-reflective nature of the minimal self because it concerns a contrast in the way perceptual content is given (is it predicted or not?), rather than a contrast in the representational content itself. It also fits with the idea that mineness is in the background of consciousness since the feeling arises when the incoming sensory signal is attenuated, and is therefore not something to which attention is directed.

Of course, stated in terms of hypotheses and predictions this explanation sounds like an implausible sort of inferentialism and intellectualism**.** But it is fundamental to the idea of perceptions as hypotheses that perceptual inference is non-conscious: the system is as mentioned hierarchical and predictions are generated automatically for low level, subpersonal computations. It might in turn be objected that there is a feeling of mineness even when the environmental causes are not predicted. As an observation this seems correct for category-specific contents accessed at relatively high hierarchical levels**,** but the objection overlooks that in such cases there is much successful low level modeling and prediction of line segments at different orientations, and of edges, shadows, contours, shapes, colours, trajectories of movement etc. For example, even if, at an intellectual level, we do not expect a bear to enter our office, we will nevertheless experience it as it comes in; and this experience will have mineness. But this is no objection to the proposal because the experience will be based on rapidly updating hypotheses that best explain the outcomes of predictive systems very low in the visual hierarchy.

We can then unify the pre-reflective sense of minimal self in agency as well as visual perception under the same independently motivated cognitive model. In both cases, it appears meaningful to say that a sense of mineness or familiarity arises that fits well with the character of the minimal self. We can then understand the elusive sense of minimal self in terms of having internal models that successfully predict or match the sensory consequences of our own movement, our intentions in action, and our sensory input.

This proposal fits well with the idea that perception and agency are closely related (see Wolpert *et al*. 2001, who holds that the brain is for movement). In order to test predictions, and thereby in order to possess a minimal self, the creature must somehow move around in the world. So from this perspective, the phenomenology we associate with perceiving the outer world, and with the minimal self, is fundamentally tied to agency.[2]

In the discussion of phenomenology of agency, I noticed that the minimal self seems associated with the Husserlian phenomenological notions of temporal structure, in particular with the notion of protentional anticipation of the future. This also holds for the phenomenology of visual perception. A cognitive system with generative models implemented with predictive coding seems well suited to capture such a notion of anticipation. In these terms, my proposal regarding the minimal, pre-reflective self thus incorporates a protentional anticipatory temporal aspect of experience. But anticipation alone does not explain mineness. It is what happens as the anticipations, or predictions, are tested (at low levels in the visual hierarchy) that explains mineness. The sense of minimal self arises when the anticipations are correct. To tell this story, the neurocomputational approach is needed.

4. The Pondering Self in Planning and Attention

We have seen that the notion of forward models in sensorimotor control connects with the more general notion of generative models and predictive coding in perception. Now I tie these notions to studies of overall brain function. This allows us, within a shared framework, to connect the above proposal concerning the pre-reflective self with a proposal about a more common idea of a temporally extended, reflective, ‘narrative’ self.

If the brain's basic cognitive system is characterised by generative models and predictive coding, then we should conceive of sensory information as something that merely *modulates* existing, on-line representations rather than wholly *determines* brain activity. As a consequence, the activity we see in, e.g., brain imaging studies such as fMRI of the brain would have to do with how the incoming sensory signal is met by and partially cancelled out by predictions, and how hypotheses or models are updated in the light of error signals.

Recent studies on human brain function seem consistent with this interpretation of cognition and human brain function, at least to the extent that there is considerable spontaneous, task-independent activity. At rest (resting quietly with the eyes open or closed), there is roughly the same ratio of oxygen consumption to oxygen delivery throughout the brain (Gusnard and Raichle 2001; Raichle *et al*. 2001). This seems to identify a reasonable baseline state in the human brain during rest in which there is nevertheless considerable metabolism. Raichle and his colleagues suggest that this indicates a default mode of brain function.

Though no one area uses relatively more oxygen than any other area, the metabolic rate during rest is higher in a group of areas including medial prefrontal and parietal cortices as well as lateral parietal cortices bilaterally. It has turned out in a great number of imaging studies that when subjects successfully perform goal-directed and attention-demanding task, these particular areas of the brain are *deactivated* relative to the baseline while task-relevant (visual, auditory etc) areas are activated (Shulman *et al*. 1997; Gusnard and Raichle 2001; Raichle 2001). The deactivation is task-independent and increases with the difficulty of the goal-directed task. There is also evidence that these areas are internally correlated with slow, spontaneous fluctuations in the BOLD signal (< 0.1 Hz), and that these in turn are anticorrelated during rest with the fluctuations internal to the frontal and parietal areas typically activated in attention-demanding tasks (Fox *et al.* 2005).

In contrast, these medial prefrontal and parietal areas are *activated* relatively to the baseline in tasks that can be said to involve the self such as asking subjects to make explicit evaluative judgments about themselves, taking a first- vs third-person perspective, reflecting on aspects of one's mental state, or in episodic memory retrieval (Gusnard *et al*. 2001). Activity in these areas is however not restricted to self-related tasks as they are also recruited during controlled or strategic coordination in multiple demanding tasks where one for example integrates the outcomes of two or more separate cognitive operations in the pursuit of a higher behavioural goal (e.g., Ramnani and Owen 2004).

It has therefore been suggested that these areas of the brain are especially important for enabling a sense of self-awareness. Further, it seems that the deactivation in these areas in attention-demanding tasks correspond to the experience of immersion in activities without much reflective self-awareness (Gusnard 2005), it is quite literally 'losing one self' in the current task. Since activity in these areas is also seen in tasks involving multiple demanding cognitive operations, Gusnard suggests further that the sense of self may be associated with the ability (of higher primates) to disengage from current stimuli and be able to plan behaviour and prioritise cognitive resources according to the tasks at hand.

This is an intriguing proposal. It concerns one of the core aspects of the sense of self that I mentioned at the beginning because it concerns the way we make sense of ourselves as cohesive, planning agents that can attend to the aspects of the world around us that would afford our preferences and plans. This captures some of what Dennett and others (for review, see Zahavi 2006: Ch 5) have called the 'narrative' self.

> [W]e are all virtuoso novelists, who find ourselves engaged in all sorts of behaviour, more or less unified, but sometimes disunified, and we always put the best "faces" on it we can. We try to make all of our material cohere into a single good story. And that story is our autobiography. The chief fictional character at the center of that autobiography is one's *self* (Dennett 1992).

Moreover, the proposal connects this sense of the narrative self, or at least a "proto-narrative self', to fundamental overall brain function ('proto' because it doesn't immediately concern the actual telling of the narrative). It is especially interesting that it reveals a very fitting seesaw of activity between self-reflection and attention, with the default state being at the balancing point. Without a sense of who one is in terms of one's plans and preferences there would be no sense of a reflective self as we know it, but it is also difficult to imagine our sense of a cohesive self without it being put to use on a set of tasks to selectively engage ourselves in, and lose ourselves in.

As mentioned, these findings on overall brain function integrate well with a cognitive framework employing generative models and predictive coding, at least to the extent that sensory input is presented as modulating rather than determining brain activity. One can therefore speculate about the cognitive operations performed by these cortical networks during self-reflection, rest and attentional immersion. If correct, these speculations throw light on this notion of the proto-narrative self.

Before one engages actively with a task, attends to its components and loses oneself in it, one must figure out what the task is all about, what one's role is in this situation. This is a kind of theorising where one needs to arrive at good hypotheses that make sense of the situation relative to one's own role in it. There are two aspects to this. Models, or hypotheses, must be able to explain the incoming input, the actual state, and this is done under uncertainty as to what is really going on. And, models must be formed about how the preferences of the agent can be satisfied, that is, become the actual state, while taking the overall situation into consideration. In other words, at this high cognitive level that involves reflective rather than pre-reflective self-awareness, the system needs to arrive at some relatively probable generative models for perception and action that explains the data and probabilifies the desired state. The nature of the reflective self as identified in these studies is as a *pondering self* in search of good high level hypotheses with a high prior probability, that can generate good predictions about what happens in complex situations, and that will be fruitful for bringing about the desired state. As we have seen for generative model frameworks in general, these hypotheses do not appear de novo but are based on the context and what has gone before (e.g., the experimenter's instructions) and in practice they will be updates of older hypotheses in the light of error signals. It seems plausible that this temporal aspect will contribute to the feeling of having an extended self rather than a 'staccato' self that on each new occasion finds itself applying some hypotheses with high priors or good fit to desires.

Some further supporting observations can be made. The interpretation of the default mode in terms of generative models gives a rationale for why there should be deactivations when one immerses oneself in attention-demanding tasks. If the function of the self-network is to identify the best hypotheses in a given complex behavioural situation, so that predictions can be generated on the basis of these, then once these are identified it would be very disruptive to the correct functioning of the whole system if further, less probable high-level hypotheses were allowed to generate hypotheses too. Only if the original hypotheses turn out to generate fairly large error signals should activity be seen in the areas related to self-reflection. Then those areas are recruited in order to help in updating or replacing the hypotheses in question. And indeed this seems to be the case for there is activity here in precisely those tasks where a clear and easy choice of good hypotheses is difficult, namely in tasks involving multiple demanding cognitive operations where one presumably quite often must re-appraise the global situation to engage successfully. These two types of tasks associated with activity in these areas are therefore closely related. Possibly, self-reflection tasks has more to do with modulating generative models for bringing about desired states whereas multiple demanding operations has more to do with modulating models that can explain the incoming complex data. Of course, in real life outside of the highly constrained tasks set in the scanner, these two aspects are intertwined.

Even more speculatively, the seesaw between self-reflection and attentional immersion may be an instance, at a very high level, of a commonly viewed pattern of activity at many levels in the brain called repetition reduction. It is normal to observe a great deal of activity as tasks are first engaged in. But as they are mastered, neural activity dwindles (Henson 2003; Maccotta and Buckner 2004). The explanation could be the same at all levels: first hypotheses must be identified and then the search for novel hypotheses must stop until something unpredicted or undesired comes up. This yields a cognitive interpretation of the default mode during rest identified by Raichle and his colleagues: the important thing is not that the subject is resting but that no novel sensory input needs to be explained and that the actual state corresponds close enough with the desired state (the subject has after all agreed to lie in the scanner). The prediction is that we should observe an oxygen extraction fraction after repeated tasks (such as prolonged finger-tapping) that is very similar to that observed during rest. The explanation of the high level of metabolism during the default mode would then be that the system is maintaining the current hypotheses and predictions while being open for whatever desires and sensory input that comes next; as long as the nature of these are not known the default state cannot favourise any particular hypothesis. From this cognitive perspective, the default mode therefore signifies a state of epistemic contentment (all is sufficiently explained and desired) and openness to a relatively uncertain future.

Notice that it is possible to treat self-reflection about desires similarly to how we treated intentions for movement in the discussion of movement (Section 2). That is, perhaps we are not aware of our desires, they are the unknown state of the ‘inner world’ that the generative self-model must try to match (just as the world model tries to match the unknown states of the world). We are then only aware of our desires through their processing in the generative model—and we acquire a reflective sense of self in so far as the model matches the desires that function as input to the model. The sense of narrative self is jeopardised when there are mismatches between model and desire (dissociative identity disorder may for example arise when the self-model fragments to deal with seriously conflicting desires and beliefs; patients would lack insight because the desires are accessible only through the fragmented self-models). Treating desires as unknown states to be matched by self-model in this way gives an independent rationale for the notion of 'narration' since a model is a kind of coherent story. This would have the attractive feature of not being a constructivist notion of narrative self since the self-model would be supervised by our real desires and intentions. Finally, it explains the reflective self in an agency-focused framework: we gain a reflective self as we try to make plans for action in a complex world.

The notion of a proto-narrative, pondering self coheres nicely with the notions of the predictive self in agency and perception since they all are manifestations of the same overall cognitive system, a system moreover that can be used to interpret Raichle and his colleagues' influential imaging studies of the default mode of brain function.

5. When Prediction and Pondering go Wrong

The cognitive system that explains mineness in agency has been used to explain the occurrence of delusions of control in schizophrenia (Frith 1992; Frith *et al*. 2000; Frith 2005). If there are problems with forward modeling of one's movements then there will not be enough attenuation of the incoming sensory consequences of the movement. The result is that the movement will be experienced in the same way as externally initiated movement is experienced. This experience will be unsettling and unusual because there is no normal explanation of it (such as, "someone pushed me" or "the floor suddenly gave way"). The subject would then explain this with the occurrence of supernatural powers or the like, saying things like "I felt like an automaton, guided by a female spirit who had entered me during it [an arm movement]" (Spence *et al*. 1997). Elsewhere, we have proposed why the supernatural explanation is favoured rather than the more probable, to other people, explanation concerning brain pathology (Hohwy and Rosenberg 2005).

Here I will speculate about how a parallel proposal would work in the case of perception (see also Hohwy 2004b). I have proposed that the framework using generative models and predictive coding can be interpreted as a perceptual reality testing system that grounds the cognitive distinction between the inner and the outer. It is therefore attractive to use this to explain other psychotic symptoms such as visual and auditory hallucinations that involve mistaken attributions of something inner to something outer.

The central principle for predictive coding frameworks is that the hypotheses with the highest posterior probability deliver the content of what is perceived—it is a "winner takes all" approach. One perceives what one best predicts, that is, what is associated with the lowest error signal. When this system works well one can track the real causes of sensory input in the external world. Hence it grounds reality testing and our ability to correctly attribute the external causes of our sensory data. The difference between this story and the sensorimotor story accords with the intuitive principle that the sensory system concerns how the world casually affects us and the motor system concerns how we causally affect the world. In the sensorimotor case, one correctly attributes causality to the inner ("It was me") when there is little error signal and to the outer ("you did it") when there is large error signal; misattribution to the outer of something that is really inner ("a female demon did it") then occurs when something inner or self-initiated is wrongly associated with relatively large error. In the case of perception, one correctly represents outer causes of sensory input ("this is how things are") when there is little error signal and correctly represents inner or noisy causes of sensory input ("this is merely how things seem") when the error signal is large. Misrepresentation of inner or noisy causes as outer causes—of what really just seems to be the case as what is really the case—would then occur when representation of inner or noisy causes are wrongly accompanied by relatively little error signal.

I therefore suggest that psychotic patients can have normal reality testing competence but that reality-testing performance is compromised (for this distinction as applied to delusions in general, see Gerrans (2001)). That is, they perceive what they best predict—what is associated with the globally smallest error signal. This is what gives them reality-testing competence. But since they may have some kind of trouble with predictive coding (with the detection of error signals and/or the formation and updating of generative models), the hypotheses that create the globally smallest error signal are in fact very bad hypotheses polluted with much internally generated noise. For healthy subjects they would be accompanied by relatively large error signals. However, since the globally best hypothesis is always favoured, patients cannot but believe that they represent real causes in the external world. They therefore develop, e.g., visual hallucinations or delusions.

A good heuristic for a probabilistic network such as predictive coding is that it concerns inference to the best explanation of sensory input (or, as I have suggested, of intentions and desires). Patients' trouble with reality testing performance manifests a very common problem for inferences to the best explanation, namely that the explanation one makes an inference to may be best out of a poor set of candidates. This proposal has the advantage that it combines into one cognitive system the sources of the deviant sensory content of psychotic states *and* malfunction in one's reality-testing performance. One forms representational content *as* one reality tests. It can therefore explain economically why deviant content goes with reality-testing problems. An important aspect is that, since this is the brain's basic representational and reality-testing system, there are no further avenues of representation and reality-testing to seek out in case one is told by family and carers that the system's deliverances are faulty (see also Hohwy and Rosenberg 2005). This corresponds with the clinical experience of patients' belief being unrevisable and the hallucinations recurrent. It may also play a role for this unrevisability that the advice from family and carers are also external causes of sensory input that needs to be predicted (Wolpert *et al*. 2003), which therefore also can be distorted when there is trouble with predictive coding. Likewise, cognitive trouble with generative models may afflict the integration of different pieces of evidence (e.g., testimony from carers with visual evidence) in the formation of a stable representation. The reason is that such integration can be seen as a type of reality-testing too, just as one may try to integrate haptic with visual evidence in order to create a better representation of something glimpsed.

On my suggestion that desires and longer term intentions also are unknown states that the self-model tries to match, other psychotic experiences, such as made emotions, begin to make sense. They are emotions that are not matched well by the generative self-model. Similarly for thought insertion, another passivity phenomenon, if thoughts are unknown states the evidence of which our generative models try to explain.

Notice that the suggestion is not that there is a total break-down of predictive coding in psychotic patients. This is not plausible since is would lead to gross impairments of representation and motor behaviour (Frith 2005). It seems plausible that there could be varying degrees and locations of impairments depending on the extent of damage to neural interconnectivity—much here depends on the etiology of psychotic disorders. For motor behaviour, for example, it also seems plausible that the experiences of external initiation are intensified when the movement is itself the goal of the motor behaviour rather than when it is a means to something else (Frith 2005; Hohwy and Rosenberg 2005). Similarly, it seems plausible that psychotic experiences are intensified when the experience itself is the object of attention, as in introspection, rather than a means to a behavioural goal such as locating a desired object (there may be a vicious, self-reinforcing cycle here, because less attenuated signals are attention grabbing in themselves and this may lead to them monopolising mental space and to hyperreflectivity, see (Sass and Parnas 2003; Gerrans forthcoming)).

On this proposal, it is also possible to explain how there can be some degree of loss of sense of minimal self in perception for people suffering from schizophrenia (Parnas *et al*. 1998; Sass and Parnas 2003). Even though they have perception, since we always perceive what is best predicted, their predictions are in fact rather poor. Therefore there will be less of a sense of already knowing or being ‘at home in’ the incoming signal simply because it is less well predicted than in healthy subjects that have an intact sense of mineness in perception.

6. Summary

I have described three cognitive tasks that seem relevant to the notion of the self, and argued how these cognitive tasks can be mapped on to three phenomenological aspects of the conscious sense of having a self. Moreover, the cognitive tasks, and the senses of self, can be united under one general, and independently motivated, cognitive framework based on generative models and predictive coding. In terms of neurobiology, the tasks and the senses of self are united under the idea that incoming sensory data modulate rather than determine neural activity. We have independent reason to believe that this is the right kind of cognitive system for the human brain, and we can see how just this kind of cognitive system would give rise to just these subjective experiences in an otherwise conscious subject. We thus get a parsimonious account in terms of cognitive and neurobiological unity of core aspects of the sense of self.

The sense of having a self is very hard to make explicit. Even though we can recognise the experiences in our own case, we are forced to resort to metaphors such as ‘narration’ or ‘autobiography’, or to neologisms such as ‘mineness’ in our attempts to describe them. I think the predictive coding account makes sense of whatever we can explicitly say about the sense of self and I think the account throws light on the nature of the self, the experience of which is so hard to capture in words. We learn something about what kind of thing the self is from this kind of account. The important thing in having a self is not so much the self itself, as the unknown states of the worlds around us and within us. That is to say, our sense of self is shaped by, and is the cognitive backdrop to, our need to move and act in the world, to represent the world and to plan reasonable courses of action in the light of the complex affordances of the world and our own preferences and desires about how the world should be.

I do not want to claim that generative models and predictive coding map perfectly onto the Husserlian phenomenological notion of the minimal self or the notion of the narrative self. Rather, properties of those cognitive mechanisms transpire as worthy deservers of the label 'minimal self' and 'narrative self'. Even in the absence of the Husserlian phenomenological analysis, consideration of the cognitive mechanisms could have lead one to predict that an otherwise conscious system implementing the mechanism would have a feeling of mineness associated with its agency and visual perception.

Notice finally, that the kind of reductive explanation offered here, in terms of a 'mapping' between cognitive and phenomenological properties is not as ambitious as it might at first seem. The claim is not that every system that realises generative models and predictions will have conscious experiences of mineness. This seems clearly wrong since rather primitive computers could realise such systems and yet intuitively lack completely in phenomenology. The claim is instead that in a system that is otherwise conscious, generative models and predictive coding can explain why there are conscious experiences with the content or representational mode we try to capture with 'narration' and 'mineness', rather than other kinds of conscious content (such as contents associated with inferential patterns or purely reflective access, or the contents characteristic of forms of psychopathology). It is thus a type of contrastive explanation aimed, not at consciousness itself, but at why specific aspects of conscious contents are they way they are, rather than another way (this strategy is discussed in Hohwy and Frith 2004). [3]

References

Bargh, J. and T.L. Chartrand (1999). The unbearable automaticity of being. *American Psychologist* 54: 462-479.

Blakemore, S.-J. and C. Frith (2003). Self-awareness and action. *Current Opinion in Neurobiology* 13(2): 219-224.

Blakemore, S.-J., D.W. Wolpert & C. Frith (2002). Abnormalities in the awareness of action. *Trends in Cognitive Sciences* 6: 237-242.

Block, N. (2005). Review of Alva Noë, "Action in Perception". *The Journal of Philosophy* CII(5): 259-272.

Boyer, P., P. Robbins & A.I. Jack (2005). Varieties of self-systems worth having. *Consciousness and Cognition* 14(4): 647.

de Vignemont, F. and P. Fourneret (2004). The sense of agency: A philosophical and empirical review of the "Who" system. *Consciousness and Cognition* 13(1): 1.

Dennett, D. (1992). The Self as a Center of Narrative Gravity. *Self and Consciousness: Multiple Perspectives*. F. Kessel, P. Cole and D. Johnson. Hillsdale, NJ: Erlbaum.

Eliasmith, C. (2003). Moving beyond metaphors: Understanding the mind for what it is. *Journal of Philosophy* C(10): 493-520.

Fox, M.D., A.Z. Snyder, J.L. Vincent, M. Corbetta, D.C. van Essen, M.E. Raichle (2005). The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *PNAS* 102(27): 9673-9678.

Friston, K. (2002). Functional integration and inference in the brain. *Progress in Neurobiology* 68: 113-143.

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

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions: Biological Sciences* 369(1456): 815 - 836.

Frith, C. (2005). The self in action: Lessons from delusions of control. *Consciousness and Cognition* 14(4): 752.

Frith, C., S.-J. Blakemore, D.M. Wolpert (2000). Explaining the symptoms of schizophrenia: Abnormalities in the awareness of action. *Brain Research Reviews* 31: 357-363.

Frith, C. (1992). *The Cognitive Neuropsychology of Schizophrenia*. Hillsdale, NJ, Lawrence Erlbaum Ass.

Frith, C. and D.J. Done (1989). Experiences of alien control in schizophrenia reflect a disorder in the central monitoring of action. *Psychological Medicine* 19: 359-363.

Gallagher, S. (2000). Philosophical conceptions of the self: Implications for cognitive science. *Trends in Cognitive Sciences* 4(1): 14.

Gallagher, S. (2005). *How the Body Shapes the Mind*. Oxford, Oxford University Press.

Gallagher, S. and D. Zahavi. (2005). Phenomenological approaches to self-Consciousness. *The Stanford Encyclopedia of Philosophy* Spring 2005. from http://plato.stanford.edu/archives/spr2005/entries/self-consciousness-phenomenological/.

Gerrans, P. (2001). Delusions as performance failures. *Cognitive Neuropsychiatry* 3: 161-173.

Gerrans, P. (forthcoming). Hostage to experience: how delusions monopolise the mental economy.

Gregory, R.L. (1980). Perceptions as hypotheses. *Phil. Trans. R. Soc. Lond., Series B, Biological Sciences* 290(1038): 181-197.

Grush, R. (2004). The emulation theory of representation: motor control, imagery, and perception. *Behavioral and Brain Sciences* 27: 377-442.

Grush, R. (forthcoming). How to, and how not to, bridge computational cognitive neuroscience and Husserlian phenomenology of time consciousness.

Gusnard, D.A. (2005). Being a self: Considerations from functional imaging. *Consciousness and Cognition* 14(4): 679.

Gusnard, D.A., E. Akbudak, G.L. Shulman & M.E. Raichle (2001). Medial prefrontal cortex and self-referential mental activity: Relation to a default mode of brain function. *PNAS* 98(7): 4259-4264.

Gusnard, D.A. and M.E. Raichle (2001). Searching for a baseline: functional imaging and the resting human brain. *Nature Reviews Neuroscience* 2: 685.

Haggard, P. and M. Eimer (1999). On the relation between brain potentials and awareness of voluntary movements. *Experimental Brain Research* 126: 128-133.

Helmholtz, H. V. (1860). *Treatise on Physiological Optics*. New York, Dover.

Henson, R.N.A. (2003). Neuroimaging studies of priming. *Progress in Neurobiology* 70(1): 53.

Hohwy, J. (2004a). The experience of mental causation. *Behaviour and Philosophy* 32: 377-400.

Hohwy, J. (2004b). Top-down and bottom-up in delusion formation. *Philosophy, Psychiatry and Psychology* 11(1): 65-70.

Hohwy, J. and C.D. Frith (2004). Can neuroscience explain consciousness? *Journal of Consciousness Studies* 11(7-8): 180-198.

Hohwy, J. and R. Rosenberg (2005). Unusual experiences, reality testing, and delusions of control. *Mind & Language* 20(2): 141-162.

Jeannerod, M. and E. Pacherie (2004). Agency, simulation and self-identification. *Mind and Language* 19(2): 113-146.

Kersten, D., P. Mamassian, A. Yuille (2004). Object perception as Bayesian inference. *Annual Review of Psychology* 55(1): 271-304.

Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *The Behavioral and Brain Sciences* 8: 529-566.

Libet, B., C.A. Gleason, E.W. Wright, D.K. Pearl (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness potential): the unconscious initiation of a freely voluntary act. *Brain* 106: 623-642.

Maccotta, L. and R.L. Buckner (2004). Evidence for neural effects of repetition that directly correlate with behavioral priming. *Journal of Cognitive Neuroscience* 16(9): 1625-1632.

MacKay, D. M. (1956). The epistemological problem for automata. *Automata studies*. C. E. Shannon and J. McCarthy. Princeton, Princeton University Press: 235–251.

Marchetti, C. and S. Della Salla (1998). Disentangling the alien and anarchic hand. *Cognitive Neuropsychiatry* 3: 191–208.

Mumford, D. (1992). On the computational architecture of the neocortex. *Biological Cybernetics* 66(3): 241.

Murray, S.O., D. Kersten, B.A. Olshausen, P. Schracter, D.L. Woods (2002). Shape perception reduces activity in human primary visual cortex. *Proceedings of the National Academy of Science* 99(23): 15164-15169.

Neisser, U. (1967). *Cognitive psychology*. New York, Appleton-Century-Crofts.

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

Noë, A. and K. O'Regan (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences* 24: 939-973.

Parnas, J., L. Jansson, L.A Sass, P. Handest (1998). Self-experience in the prodromal phases of schizophrenia: A pilot study of first admissions. *Neurology, Psychiatry and Brain Research* 6: 97–106.

Raichle, M.E., A.M. MacLeod, A.Z. Snyder, W.J. Powers, D.A. Gusnard, G.L. Shulman (2001). A default mode of brain function. *PNAS* 98(2): 676-682.

Ramnani, N. and A.M. Owen (2004). Anterior prefrontal cortex: insights into function from anatomy and neuroimaging. *Nature Reviews Neuroscience* 5(3): 184-94.

Rao, R.P. and D.H. Ballard (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience* 2: 79.

Rao, R.P., B.A. Olshausen, M.S. Lewicki (eds). (2002). *Probabilistic Models of the Brain*. Cambridge, Mass., MIT Press.

Sass, L.S. and J. Parnas (2003). Schizophrenia, consciousness and the self. *Schizophrenia Bulletin* 29(3): 427-444.

Shoemaker, S. (1963). *Self-Knowledge and Self-Identity*. Ithaca, NY, Cornell University Press.

Shulman, G. L., J.A. Fiez, M. Corbetta, R.L. Buckner, F.M. Miezin, M.E. Raichle, S.E. Petersen (1997). Common blood flow changes across visual Tasks: II. Decreases in cerebral cortex. *Journal of Cognitive Neuroscience* 9(5): 648-663.

Spence, S.A., D.J. Brooks, S.R. Hirsch, P.F. Liddle, J. Meehan, P.G. Grasby (1997). A PET study of voluntary movement in schizophrenic patients experiencing passivity phenomena (delusions of alien control). *Brain* 120: 1997-2011.

Stephens, G. L. and G. Graham (2000). *When Self-Consciousness Breaks: Alien Voices and Inserted Thoughts*. Cambridge, Mass., MIT Press.

Wegner, D.M. (2002). *The Illusion of Conscious Will*. Cambridge, Mass., MIT Press.

Wolpert, D.M., K. Doya, & M. Kawato (2003). A unifying computational framework for motor control and social interaction. *Philosophical Transactions of the Royal Society London B* 358: 593-602.

Wolpert, D.M., Z. Ghahramani, & J.R. Flanagan (2001). Perspectives and problems in motor learning. *Trends in Cognitive Sciences* 5(11): 487.

Zahavi, D. (2006). *Subjectivity and Selfhood: Investigating the First-Person Perspective*. Cambridge, Mass., MIT Press.

Notes

1. In discussions of psychopathology (Hohwy 2004b), I have previously suggested that this kind of cognitive system unifies the phenomenology of agency and perception. Writers more in the Husserlian phenomenological tradition (in particular, Gallagher 2005: Ch 8) have also explored the link between the minimal self and prediction (or so-called

‘protentional’ anticipation of events), though, as I shall show below, such anticipation alone doesn’t explain the sense of self.

2. Other people have emphasised the role of agency in perception (Noë & O'Regan 2001; Noë 2004). It is characteristic of this so-called ‘enactive’ view that it denies the need for explicit mental representation of the world; a core idea is thus that the world is its own model that is explored in action. This marks a fundamental difference to the use of generative models and predictive coding where the individual must maintain mental representations in order to perceive. This has to do with the ill-posed nature of the problem of perceptual inference noted above. In representing a partly occluded object (a cat behind a fence, for example), the system must overcome the noninvertible nature of the nonlinear mixing of the object and the occluder. This is done by maintaining and updating an explicit recognition model of the world in the brain, on the basis of which better and better predictions can be made (Friston 2005: 820). Put differently, it is unclear how the enactive view can account for something as fundamental as recognition of partly occluded objects without accepting mental representation. In fact, it has been argued that the enactive view collapses to a more orthodox view once problems such as these (concerning expectations) are dealt with, see Block’s (2005) review of Noë.

3. This paper was presented at a meeting of the Society for Philosophy and Psychology at Copenhagen University. I thank the audience for many valuable comments. Thanks also to Susanna Siegel for valuable comments on an earlier draft.