Why should any body have a self?*

Jakob Hohwy[1] & John Michael[2,3]

[1]Cognition & Philosophy Lab, Faculty of Arts, Monash University, Melbourne Australia.

[2]Department of Philosophy, University of Warwick, Coventry, UK.

[3]Department of Cognitive Science, Central European University, Budapest, Hungary.

1. Introduction

We use a general computational framework for brain function to develop a theory of the self. The theory is that the self is an inferred model of endogenous, deeply hidden causes of behavior. The general framework for brain function on which we base this theory is that the brain is fundamentally an organ for prediction error minimization.

There are three related parts to this project. In the first part (Sections 2-3), we explain how prediction error minimization must lead to the inference of a network of deeply hidden endogenous causes. The key concept here is that prediction error minimization in the long term approximates hierarchical Bayesian inference, where the hierarchy is critical to understand the place of the self, and the body, in the world.

In the second part (Sections 4-5), we discuss why such a set of hidden endogenous causes should qualify as a self. We show how a comprehensive prediction error minimization account can accommodate key characteristics of the self. It turns out that, though the modelled endogenous causes are just some among other inferred causes of sensory input, the model is special in being, in a certain sense, a model of itself.

* Acknowledgements. We wish to thank the editors Adrian Alsmith and Frederique de Vignemont, and Jennifer Windt as well as participants in The Body and the Self workshop in Copenhagen 2015, for comments on earlier versions. JH is funded by Australian Research Council FT100100322 and DP160102770. JM is funded by the European Research Council ERC STG 679092.

The third part (Sections 6-7) identifies a threat from such self-modelling: how can a self-model be accurate if it represents itself? We propose that we learn to be who we are through a positive feedback loop: from infancy onward, humans apply agent-models to understand what other agents are up to in their environment, and actively align themselves with those models. Accurate self-models arise and are sustained as a natural consequence of humans' skill in modeling and interacting with each other.

The concluding section situates this inferentialist yet realist theory of the self with respect to narrative conceptions of the self.

2. Bayesian inference and prediction error minimization

On the picture of the human brain that we shall work with, organisms like us construct internal models of the world in order to interact meaningfully in our environment. The inputs that our senses receive are caused by objects and states of affairs in the world, including other people and endogenous states of our own bodies. Sensory input is all the brain has to go by in constructing its model of the world (Helmholtz 1867). A series of sensory inputs is a sample of some size, for example, samples for an estimate of the location of a flash of light. Such a sample will be subject to noise, and we can assume it will be normally distributed (i.e., a Gaussian distribution). This is what the brain has to work with in its attempt to learn the true cause (the location of the flash of light). Assume for simplicity that samples come in one at a time. The brain needs to learn something for each sample, so that it can arrive at an estimate of the true, underlying cause. The first sample comes in and suggests location *a*, so the brain represents the location as location *a*. The next sample comes in and suggests location *b*. Now the brain could stick to location *a,* but would then have learned nothing from the second sample; this strategy is tantamount to assuming falsely that a first sample is veridical and the rest misleading. The brain could switch to location *b*, but would then have thrown out all information carried by the first sample; this strategy is an example of overfitting. Instead, a location between the two samples should be picked: a location, *m,* that results from weighting the existing evidence against the new evidence.

This is an exceedingly simple example, but it is instructive. Notice that the estimate of location *m* is a *model* fitted to the samples, and that the brain never receives direct

confirmation of this inferred location. In other words, if the sensory input is noisy, then the brain must be entertaining internal models that are in some sense removed from the states of affairs – the hidden causes – that are being modeled.

Bayes' rule specifies the optimal location of *m*, namely in terms of the probability of location *a*, also known as the prior – what is already believed – and the probability of the new sample given location *a*, also known as the likelihood. Bayesian *inference* is the process of updating the prior in the light of the new evidence to arrive at a new estimate – the posterior. In inference, the prior works as a *prediction* ("the next input will be *a*"), and the likelihood encodes the *prediction error*, which would be zero if the next sample were indeed *a*. In our example, the new prediction error is *b*, so the prediction error is (*a* – *b*). Crucially, the amount learned from this prediction error should depend on how robust the prior and likelihoods are. If it is not sample one and two, but sample 1001 and 1002, then even a large prediction error should not change the posterior very much; similarly, if the prediction error itself is very noisy then it should not be allowed to update the posterior very much. Bayes' rule takes care of this weighting. It turns out that for a normal distribution, this estimate turns out to be the mean, hence '*m'*.

Consider now what happens to the long-term average of prediction error as Bayesian inference converges on the mean. There will still be prediction error for every new, noisy sample. Most samples will be within a couple of standard deviations, and there will be some outliers, but overall, by sticking to *m*, error will be minimized. In contrast, any other, non-Bayesian, estimate will tend to create greater error in the long run (e.g., if the true mean is 0, and the samples -1 and 1, then a non-Bayesian estimate of -3 will generate a prediction error of 6 rather than 2) (for this approach to Bayes in terms of relative precisions, and subsequent application to hierarchical learning, see (Mathys, Daunizeau et al. 2012, Mathys, Lomakina et al. 2014)).

If Bayesian inference is prediction error minimization, then (assuming normal distributions) any system that minimizes prediction error in the long run will be approximating Bayesian inference. This is crucial to appreciate how the brain solves the problem of building a model of the world on the basis of sensory input.

It is not plausible to say that the brain knows and applies Bayes' rule – neurons don't know probability theory. But it is plausible to say that neuronal populations harbor expectations about what the sensory input should be, the strength of which is based on past learning. These expectations are compared against the actual input to extract the prediction error, and the activity of the neuronal population is adjusted in the light of the size and variance of the ensuing prediction error.

In other words, it is plausible that the brain can minimize prediction error, even if it does not explicitly apply Bayes' rule. Moreover, if the brain simply aims to keep prediction error as low as possible (taking into account irreducible noise and assuming normal distributions) on average and over the long run, then it will be guaranteed to approximate Bayesian inference (Friston 2003).

This is the fundamental idea of predictive processing: a system, like the brain, that minimizes prediction error on the long-term average will approximate Bayesian inference.

The challenges to prediction error minimization become much more prevalent in a world as complex as ours, where there is not just a single source of visual input but multiple interacting causes. In such an environment, the system must constantly assess hypotheses and compare them against each other to arrive at the best overall model. For example, the light, represented with the mean location *m*, may periodically disappear and thus create a situation where the absence of a predicted input causes a prediction error (cf. den Ouden, Friston et al. 2009). The system may then try to introduce a further hidden cause into its model, also normally distributed, that represents an occluder (e.g., someone waving their hand in front of the light). Then the system may be able to get prediction error minimization back on track, when the causal interaction of the occluder and the light is correctly modeled.

A prediction error minimization system in the real world, exposed to its manifold of causes, would build up a vast repertoire of beliefs about all sorts of hidden causes and their interactions. This would be hierarchically ordered in a spatiotemporal sense, such that small receptive fields and regularities operating at fast time scales would be at the bottom (e.g., the intensity and location of the light), and larger fields and longer

term causal regularities higher up (e.g., the movement of a hand periodically occluding the light). The hierarchy would also include expectations not just about means but also of the variances (or *precisions*) of the prediction errors it may encounter, in order for it to steer an informed route between underfitting and overfitting.

The claim now is that *we* are like this (Hohwy 2013, Clark 2013, 2016). We manage to represent the world by being prediction error minimizers – neural mechanisms that realize Bayesian inference in the long run. Equivalently, we minimize uncertainty about our model of the world – accumulate evidence for it – by minimizing error. On some versions of the prediction error minimization framework, this is *all* we ever do (Friston 2010, Hohwy 2013, Hohwy 2015).

If all we ever do is minimize prediction error, what could it mean to have a *self*? It seems prediction error minimization is just a matter of inferring causes in the world, so expecting there to be a self could appear as unreasonable as expecting the mean location of a series of visual samples, arrived at through Bayesian inference, to have a self. However, by working with the notion of a prediction error minimizing system we will show how it can in fact accommodate many of our preconceptions about the self.

Interestingly, there are now a number of studies attempting to connect different aspects of body and self with the predictive processing framework. Our approach here builds upon an early approach developed in Hohwy 2007 and in Hohwy 2013 (Ch. 12). It also resonates with and draws upon several other recent approaches, including Limanowski and Blankenburg (2013), who connect the framework to minimal phenomenal selfhood; Apps and Tsakiris (2014), who explain bodily self-awareness in terms of the framework; Seth, Suzuki et al. (2011), who focus on interoceptive predictive coding and self; Moutoussis, Fearon et al. (2014), who focus on predictive processes in the relation of self and other; Frith and Friston (2015), who focus on social understanding in terms of coupled oscillations of selves; Fotopolou (2012) and Carhart-Harris and Friston (2010), who connect the framework to psychodynamical notions of self; finally, an important precursor to the Bayesian approach is Thomas Metzinger's (2004, 2009) work on self-models (see Limanowski and Blankenburg 2013 for discussion).

In the simple example of a flash of light, we considered the possibility of causal interaction with other causes, such as an occluder making the light disappear periodically. Interactions create non-linearity in the sensory input that cannot be explained away well on the basis of a model of a simple, normal distribution. Instead, it becomes necessary to postulate multiple causes that, when convolved, are able to fit the sensory input. Of course, a good explanation for why a light periodically disappears could be that the perceiver is blinking their eyes, which technically speaking occludes the light source. If the perceiver models this, then they are modeling part of what happens to be their *own body* in order to minimise prediction error. This can be generalised and extended to any part of the body or the trajectory of the body as a whole (e.g., why are the flashes of light moving at an accelerating pace? You have begun a sprint; why is the chair creaking? You have put on weight). In this simple sense, modeling the body is the automatic upshot of prediction error minimisation, which allows perception of the world. A full body model will be finessed over time, as we accumulate evidence, learn about precisions, and learn how to actively test hypotheses through our own behavior (for discussion, see Hohwy 2013: Ch. 1-4; Seth 2015).

By taking the perspective of long-term prediction error minimization, it is thus quite easy to see how internal models of agents will end up including representations of the agent's own body and its trajectories and interactions with other causes in the world. Agents will represent their own bodies simply because prediction error minimization of necessity leads to representing the worldly causes of the agent's sensory input, and the agent's body is in fact itself one of these causes.

3. The self in the body

Perception is perfused with the interaction of the body with the environment, mandating internal representation of the body. The body is nothing special, it is just one among many causes interacting with each other in the environment, and in the course of this impacting on the senses. Representation of the body is nothing special either; it is just one among many causes that get represented in the internal model used for prediction error minimization.

As we noticed, internal models need to be hierarchical to be efficient at minimizing prediction error in the long run. For example, in the short time scale, you might be well-served to use an umbrella whenever it rains, but it also helps to model how the seasons influence the probability of rainfall so you know better when to bring your umbrella. The long-term representation of seasons is higher in the hierarchy in the sense that it helps control the predictions of rainfall, and of course, perception of actual rainfall is evidence that filters upwards and helps to shape the long-term representations of seasons. Seasons are in this sense more deeply hidden causes than rainfall itself, they (or the tilt of the Earth's rotational axis in its orbital plane) are causes that operate behind, but are perceivable in, the more directly observable causes, such as rainfall.

As it is for rain and seasons, so also for the body and its more deeply hidden causes. For example, an agent who blinks at a normal rate will blink more when exposed to an unexpected, startling auditory stimulus. In this case, fear is an internal mental state that causally interacts with the body and thereby modulates the evolution of sensory input. Eye-blinks can also be brought about by a puff of air to the eye. Here we have an ambiguous situation where two different hypotheses could explain the same visual input: something scary or a mere puff of air. Contextual information such as additional auditory input or the presence of an air puffer can help disambiguate and arrive at the best hypothesis. Similar examples can be run for other cases of body-involving prediction error. If you begin experiencing repeatedly removing yourself from parties and social get-togethers before everyone else, then there is, assuming you are normally keen to socialise, prediction error concerning the trajectory of your body. This can be explained away by hypothesizing that you are beginning to develop introversion as a character trait, or, as alternative hypotheses to test over time, social phobia, depression, or, again, as something non-self related the deterioration of your social scene.

Over time, the overall internal model will be populated with states and parameters for not just the less hidden bodily causes of sensory input but also the more deeply hidden internal causes of the agent, which interact with each other (e.g., fear plus hunger gives one trajectory of sensory input, fear plus pain gives another) and in turn with worldly influences (e.g., fear and presence of tigers vs. fear and no presence of tigers).

The model of causes hidden within the body captures an integrated net of character traits, biases, reaction patterns, affections, standing beliefs, desires, intentions, base level internal states, and so on. This representation comes about through prediction error minimization just as naturally as representations of deep hidden causes in the wider environment, such as seasons. Without representation of these deep causes we would be much worse off in our ability to minimize moment-to-moment prediction error over the long run. Causes within the body can be thought of as *endogenous* (i.e., as changes initiated within the organism in a sense we will specify more below), and causes in the environment as *exogenous* (i.e., as changes initiated outside of the organism).

Our proposal is to conceive of this internal model of endogenous causes as a representation of *the self*. The suggestion, then, is that agents model the self as a hierarchy of hidden, endogenous causes, and further, that the self is identical to these causes. As we shall explain later on, this modeling in turn becomes an endogenous cause serving to stabilize the hierarchy of causes constituting the self, so the self-model is part of the self. Our proposal is motivated by the basic idea that who we are is determined by the regularities in our constitution that determine what we feel, think and decide, and how we act. This proposal sits well with earlier notions of self-models, in particular Thomas Metzinger's approach: "A self-model, an inner image of the organism as a whole [is] built into the world-model, and this is how the consciously experienced first-person perspective develop[s]" (2009: 64; see also 2004).

This proposal presents the self as more than a mere bundle of causes. The self-model is a hierarchical construct whose levels are linked by message-passing as top-down predictions are generated and bottom-up prediction errors minimized. For example, the agent's desire for ice cream manifests every day but its magnitude is modulated by longer-term regularities associated with the desire to lose weight, which manifest predominantly at the beginning of the calendar year. All the parts of the self-model are thus tied together and interact causally with each other. This captures the idea that one's self is comprised of both long-term characteristics and the expression of these in

shorter-term behaviours, and, vice versa, that our long term characteristics must depend on relatively stable patterns in our shorter-term behaviours.

Over time, noise and uncertainty are filtered out of the self-model, just as they are in the modeling of the broader environment in which the self is causally enmeshed. This means that more coincidental events do not make it as parts of the self-model. For example, if one summer day the ice-cream loving agent by happenstance refrains from eating ice cream, this doesn't mean the parameters of the self-model should be revised. If such coincidental events are included, then the self-model risks being overfitted to coincidentally varying behaviour, which is not so clearly self-related. However, if this kind of variability reduction happens too much, then the model becomes uninformative (underfitted) and fails to capture actual patterns of behaviour. This balancing act between overfitting and underfitting corresponds to what happens in any kind of Bayesian modeling, as described above.

A balanced self-model of endogenous causes may be paraphrased in simple terms as a *theory* or *narrative* that appeals to regularities or plotlines at different, interlocked time scales. Such a theory or narrative can be seen as an answer to the question: which kind of agent am I? For a simplistic agent, this might be "I am an ice-cream loving person who worries about my weight, and this manifests in the way I go through life, where I am often eating ice-cream though I manage to refrain in the couple of months after my New Year's resolution". For realistic agents, the narrative will be very much longer and complicated. Crucially – and in contrast to narrative accounts of the self (as we will explain in more detail in Section 8) – this account does not imply that agents generally make such narratives fully explicit, nor that the narratives in questions take a linguistic form, nor that they are coherent in the sense of forming a coherent plot or story (cf. Dennett 1991, 1992, Velleman 2006, Hardcastle & Flanagan 1996). Rather, it is a narrative account, in the first instance, in the sense that the subsumption of events under higher-level regularities (i.e., hierarchical Bayes) structures and constrains our interpretation of those events. As we shall see later, in Section 6-7, the self-model is also narrative in the sense that it actively shapes itself over time to align with those higher-level regularities.

Above, we briefly noted that *action* is a useful tool for disambiguation of models of the body. The reliance on action is unsurprising on a prediction error minimization scheme since action is an efficient tool for reducing uncertainty in causal inference. Action, or intervention in the world, allows us to go beyond mere associative evidence (e.g., *A* is statistically associated with *B*) and on to learning causal relations (e.g., *A* causes *B*) (Pearl 2000, Woodward 2003). Accordingly, action is also crucial for forming the self-model – in shaping the narrative. For example, if you suspect that your fear of heights is subsiding then you may seek out tall buildings to accumulate evidence for this hypothesis. James Russell neatly ties action in with forming a conception – or model – of the self in causal terms:

> The freer we are to alter our perceptual inputs, the more we learn of the *refractory* nature of the world and, correlatively, the richer the conception we gain of ourselves as determiners of our immediate mental life. This refractoriness, therefore, sets limits on what our agency can achieve in determining our experiences, thereby engendering a conception in us of something as setting these limits, as causing them to be set (Russell 1995: 134).

Through action, the self-model stands out more sharply and with less uncertainty. Of course, we have also argued that it is action itself that leads to the need for a self-model in the first place. The agent's causal interaction with the world causes changes to sensory input of a kind that is best explained away by positing a body and a self. We thus represent the self because we in general are active creatures, and we act so that we can represent the self. This circularity is a springboard for discussing some of the deeper aspects of the prediction error minimization approach to the self (Sections 4-5), as well as the developmental dynamics in the shaping of the self (Sections 6-7).

4. Body, self, and existence

The self-model is hierarchical, which means that it encompasses causal regularities governing events at many time scales. At the low levels of this hierarchy are regularities that represent the body and its movement, which happens at relatively fast time scales (e.g., 1-2 seconds for reaching, a few hours for moving around town to do the shopping). At intermediate levels are regularities that represent medium-term

endogenous causes (e.g., days for reading up for an exam, months for a new learning objective). At higher levels are stable regularities such as character traits (e.g., being fastidious).

What we label the 'self' is constituted by more deeply hidden causes than what is represented in the body-model specifically. Given the hierarchical ordering of the overall model, this does not imply that body and self are separate entities. They are tied up in one causal nexus, just as daily rainfall and seasons are tied up together. We can certainly talk about the movement of the body on its own, and we can talk about dynamics of the self on its own, but it is very hard to understand or predict the movements of the body or the dynamics of the self without connecting them to each other in just the way hierarchical Bayes suggests. For example, knowing character traits enables predictions of bodily movement, and character traits that never manifest in bodily movement will quickly begin to look spurious.

By treating body and self as just more hierarchical modeling of the causes in the world we thus get an appealing picture of body and self as separate and yet related. This makes good sense of core intuitions about body and self. When we speculate about the self it is hard to avoid involving thoughts about the body, and conversely, it is difficult to divorce thoughts about one's own body from thoughts about one's self. Yet our conceptions of body and self allow that they can be referred to separately (e.g., "my body may be old but I still feel young inside").

There is nothing special about this hierarchically mediated relation between body and self. The self-model represents them together, just as the model of the world represents snowfall and seasons together. One analogy here would be biological cells. Body and self are as connected as subcellular components are to the detailed activity at the cell-membrane. There is no great mystery in either case; the main difference would be that in the case of the body and self, there are causes at more levels of hierarchical depth than in the cell (even though cells are rather complex biological entities; see Friston 2013 for a prediction error minimization account of basic organic entities).

In his delightful thought experiment 'Where am I?', Dan Dennett (1981) imagines that his endogenous causes, harboured in the brain, are dislodged from the body but connected to it wirelessly. Dennett notes the profound uncertainty he feels in trying to locate himself either where his brain is or where his body is. We would feel the same kind of uncertainty if a similar thought experiment were made for the cell: is the cell where the membrane is or where the subcellular components are? We feel something has been left out of the story if either is ignored. The Bayesian explanation is simple: the causal hierarchy ties body and self (or membrane and subcellular components) together such that the processes in one only makes sense in the context of the other. The thought experiment works because on the one hand the mathematical detail is left unchanged such that the causal hierarchy is intact (Dennett's body still is the lower level of the causal hierarchy and the endogenous causes in his brain at the higher levels). On the other hand the self-model is challenged because we have learned over time that the self and the body are co-located. In this sense, this type of philosophical reflection about the self is very similar to self-related illusions like the rubber hand illusion or full body illusion (Botvinick and Cohen 1998, Lenggenhager, Tadi et al. 2007; see also the theoretical background for this work on self-models, developed by Metzinger 2004; 2009, which also challenge the core belief of spatial co-location associated with the self-model).

We next offer a consideration about the self-model, which helps tie it to a very traditional notion of the self, namely that having or being a self is in some inchoate sense tied to *existence* or *being* – to the fact or thought *that I am*. At the low level of the overall self-model we have parameters representing the body. These parameters are modulated in part by higher-level causes, the set of which we identified with the self. We said that at the higher levels of the hierarchy we have stable regularities concerning, for example, character traits. At the very highest levels, details are stripped out and only more general beliefs are found. This leads, in the end, to just the belief that *I exist*. In the most ambitious versions of prediction error minimization schemes, which are based on free energy minimization, the belief that I exist is equivalent to the belief that I act to maintain myself in certain states (Friston and Stephan 2007; for further discussion, see Limanowski and Blankenburg 2013). The reasoning is that I cease to exist if I cannot maintain myself in a limited number of states, that is, if my body begins to disperse across many states. I must act to prevent

dispersion: inaction leaves me open to the unfettered impact of the second law of thermodynamics. So if I exist I must be acting, and if I act I must exist. Put differently, the highest-level belief is that I am a cause in the world, or an agent. This belief permeates all the levels below in the sense that they would all generate large prediction errors in the long run if agency ceased. Conversely, that highest level belief in being an agent is itself based on just the kind of Bayesian inference exemplified above: the best prediction error minimization is obtained by modeling the self and its body as partial causes of changes in sensory input. Here we see how the circularity mentioned earlier (I act to model myself and model myself because I act) is grounded in the free energy principle (for this kind of approach, see the free energy principle, Friston and Stephan 2007, Friston 2011).

This speculative link to the free energy minimization account enables substantial characteristics of our conception of self to emerge from hierarchical Bayesian processing. Being a self is in a fundamental sense tied to existence and to agency, rather than merely to having a certain narrative, certain character traits, certain desires, or a certain bodily configuration.

The hierarchical Bayesian perspective can be illustrated through David Hume's famous failure to identify the self through introspection:

> For my part, when I enter most intimately into what I call myself, I always stumble on some particular perception or other, of heat or cold, light or shade, love or hatred, pain or pleasure. I never catch myself at any time without a perception, and never can observe any thing but the perception (Hume 1739-40: B1.4.6)

When he stumbles upon his perceptions he is reporting on low-level causes, and failing to report that the trajectory of sensory experience is governed by more deeply hidden causes, which are not directly perceivable in the same sense and yet are represented. Hume might as well have complained that he never perceives seasons but only snowfall, blooming flowers, ripening fruit and falling leaves. He is looking in the right place but focusing on the wrong timescale. As a result, Hume succumbs to what might be termed the *short-timescale fallacy*, failing to pick up the real patterns which

unfold at longer timescales and which underlie more rapidly fluctuating perceptual states. In particular, he fails to appreciate the more deeply hidden causes that pertain to his own interaction with the world and which are informed by the highest-level belief that he exists and is an agent in the world.

5. The model that models itself

The self is just one set of causes among other causes of sensory input. By minimizing prediction error, the self can be modeled just as other causes can be modeled. If the cause is a leafy tree in the environment, then the internal model represents that leafy tree. Seen from an outside, third-person perspective, there is clear separation of the causal relata – of the model and what is modeled. The model is harboured in the brain and the tree is outside the brain. The less the prediction error, the more the mutual information between them.

The situation is less straightforward with the self-model and the self. Seen from the outside, third-person perspective, it is not immediately clear what the causal relata are. In some sense, the self-model must be representing itself rather than things external to itself. But this notion of self-representation is somewhat mystifying. Metzinger discusses this, explaining a weak version of the self in terms of "a self-organising and self-sustaining physical system that can represent itself on the level of global availability" 2009: 208. Metzinger denies the existence of a self in a stronger sense, based on the idea that the self-model is a mere process, a "biological data format" (ibid.; see also Metzinger 2011). We agree with much of this but below we offer a seemingly more metaphysically robust account of self-representation in terms of inferred hidden causes.

Consider first that inference of the deeply hidden causes constituting the self is mediated through the impact of these causes on less deeply hidden non-self causes in the body and the environment – these are the states the agent can control through action. Blinking or moving a cup are examples of this. For inference of other hidden causes such as seasons, migrations, and economic downturns, the causal flow can be pictured as relatively straight, going through deeply hidden to less deeply hidden causes, and then together impacting on the sensory organs of the organism in question and shaping its internal model. In the case of the self, this flow simply bends back on

itself such that the deeply hidden causes stem from the organism itself. This is part of what is modeled: the self-model represents the circular character of the causal flow.

The notion of self-involving circular causation does not wholly demystify self-modeling. This is because it is clear that *action* must be central to understanding what is going on: without action there would be no need to posit endogenous causes to explain changes in prediction error. It will therefore help to clarify what action is, according to the prediction error minimization approach.

Above we noted that a system that minimizes prediction error on average and in the long run will approximate Bayesian inference. Prediction error can be minimized by changing the model to fit the data, as happens in perception. But prediction error can obviously also be minimized by acting in the world. For example, you move the source of sound to the location you expect it to be at. Since average long-run prediction error minimization will approximate Bayesian inference, it follows that considered in the long term, action is inference just as much as perception is; it is *active inference* (Friston, Daunizeau et al. 2009). Inference is the job of the model, so active inference too stems from the model. It follows that the part of the model that is involved in active inference is the self: this part of the model (the active states and their more deeply hidden causes) are the very endogenous causes that can be inferred in perceptual inference, which therefore become part of the self-model that in turn, in a dynamic downstream manner, shape active inference.

Self-modeling can now be demystified further. The self is modeled in perceptual inference, as the system learns what its own self is. What is learnt are hierarchical patterns of active inference: the deeply hidden expectations that the system relies upon in active inference (for example, a pattern may be that you consistently move classical music sound sources closer to yourself, but other types of music further away, revealing something about your deeper desires). The self-model's learnt patterns of active inference in turn inform the next volley of active inference (e.g., relocating to Vienna to make classical music encounters more likely). If this process manages to keep prediction error low on average and over the long run, then uncertainty about the model is gradually reduced. The model thus comes to know itself – represent itself – through action (as anticipated in the quote by Russell above).

It is worth taking a moment to let the strangeness of this idea sink in. It may seem that there must be ways of reducing prediction error about the self other than revising the self-model. In particular, it seems that I could simply move to a new environment that does not elicit behavior from me which generates the prediction error. For example, if I notice that I have been attending lots of Schönberg concerts even though I don't believe myself to be a fan of atonal music, do I really need to revise my self-model in order to do away with the discrepancy? Can't I just move to a place where there is no possibility of going to Schönberg concerts (Wyoming?). In fact, however, this would be a form of active inference, and its successful execution (i.e. choosing a place where there really is no Schönberg to tempt me) would require me to be sensitive to just the right hidden cause in myself (when offered the opportunity to listen to Schönberg, I tend to take it). In other words, it turns out that this move, too, would constitute a form of Bayesian inference about my self.

Self-modeling as a process of coming to learn one's own model through action implies that there are two related stages in the flow of activity in the brain of a prediction error minimizing agent from sensory states, impacted by the worldly causes, through the internal states and on to the active states that enslave bodily movement, which in turn affect changes in the world's hidden causes. The key is that active states are the downstream effects of what we have called deeply hidden endogenous states, and these endogenous states are the downstream effects of sensory states. Self-modeling is thereby a process that can be described in causal terms, as first finessing a model of the self, and then engaging that model in action, leading to further finessing of the model, and more action. (This approach applies the notion of a Markov blanket, which is essential to the free energy principle (cf. Friston 2013), to the case of the self).

For this dynamic inferential process to make sense, it cannot be that any inference about the patterns of active inference is as good as any other; the self-model would be cluttered with all sorts of more or less plausible beliefs, and prediction error would not be minimized. But since some actions are more likely than others to minimize prediction error, the system is constrained insofar as it needs to learn which patterns of active inference are associated with long-term prediction error minimization. That

information is available to the model, since prediction error is the one quantity the brain can compute. This means that patterns of active inference can be ranked according to their ability to minimize error.

A ranking of patterns of active inference will facilitate decision-making, on the assumption that agents believe they will occupy low prediction error states in the long run (i.e., that they will exist). This in turn implies that the highly ranked patterns will reflect how probable it is that the agent harbours these patterns, that is, that those patterns go into constituting the self. In other words, the better the pattern of active inference is at minimizing error, the more likely it is that it belongs to the agent. This is because the agent has learnt one overarching belief, as we mentioned above, namely that it exists as an agent – as an active prediction error minimizer in the long run. The mere belief in existence therefore makes it probable that the agent harbours patterns of active inference that are good at minimizing error in the long run. Self-related prediction error minimization will infer to those well-performing patterns over poor performing patterns, and they will then become part of what we here call the self. This means the self-modeling approximates Bayesian inference about the self, through action, because it minimizes error in the long run, fundamentally with respect to the hypothesis that the agent exists.

Of course, a self-model will only minimize prediction error in the long run if it can maintain a fairly high degree of *accuracy*. Given the circular causality associated with self-modelling, it may seem that the process can float free of any constraints. To address this concern, we will (in the next two sections) explain in some detail how to think about accuracy for such a self-modelling system. The core idea is that self-models and selves are fitted together as a natural consequence of humans' skill in modeling and interacting with each other. More concretely, infants and young children apply agent-models to others during the course of development, and use these agent-models to guide imitation and other forms of cultural learning. From the perspective of the prediction error minimization framework, this appears as a form of active inference: infants and young children shape their selves progressively to match the agent models that they have been using to interpret others. Thus, the predictive processing framework is able not just to underpin key notions of body and self but also to provide us with a novel way of thinking about recent findings from

developmental psychology pertaining to cultural learning and the development of self-understanding.

6. Agent-Modeling, and Cultural Learning as Active Inference

There is a wealth of research in developmental psychology suggesting that infants differentiate between agents, on the one hand, and non-agents on the other. In particular, they are sensitive to core features of agents, such as faces and gaze direction (Senju and Csibra 2008; Farroni et al. 2002), as well as contingent motion (Gergely and Watson 1996; Carey 2009).

Over the course of the first year of life, this sensitivity to such features comes increasingly to inform and be informed by representations of agents' goals (Woodward 1998), of their strategies for attaining those goals (Gergely and Csibra 2003), and of basic mental states such as attention (Reddy 2003, emotions (Stern 1985) and intentions (Behne et al. 2005). Indeed, there is an ever-increasing body of evidence suggesting that infants are able to represent and reason about other agents' beliefs by the second year of life (Onishi and Baillargeon, 2005; Surian et al., 2007; for reviews, see Christensen & Michael, 2015; Michael & Christensen, in press; Apperly, 2011, chap. 3; Baillargeon et al., 2010), and perhaps as early as the middle of the first year (Kovacs et al., 2010; Southgate et al., 2014).

What this evidence indicates is that infants and young children not only identify agents as a subset of the objects around them, but that they are predicting the behavior of those agents by applying an increasingly hierarchical model of what agents are like: Agents have particular desires and preferences, which they act upon in ways that are informed and constrained by doxastic representations which in turn are acquired through perception and through inferences.

We will now explain how the application of an agent-model, in particular in infancy and early childhood, supports imitation and other forms of *cultural learning*. For present purposes, we will adopt Tomasello, Kruger & Ratner's (1993) conception of cultural learning as a form of learning which depends upon learners and teachers understanding each other as beings who 'have intentional and mental lives like their own' (Tomasello 1999, 7).

One important type of cultural learning is imitative learning. Following Tomasello (1999; cf. also Tomasello et al. 2005) we will regard imitation (which is sensitive to an observed agent's goals and strategies), but not emulation (where the learner focuses on the environmental effect of an observed action) as a type of cultural learning. In other words, imitation relies upon a deeper model of the observed agent's mental states in their relation to her movements and to the environmental affordances and constraints. When engaged in imitative learning, the learner understands the agent as rationally selecting an appropriate sequence of actions to realize a goal. From about 18 months, infants tend to imitate incomplete but intended actions rather than replicating the exact behavior they have seen, e.g. when an agent tries but fails to close a drawer (Meltzoff 1995). This demonstrates that they are modeling other agents' intentions and using those intentions to structure and guide their own learning. Similarly, Gergely et al. 2002 have found evidence that, by around 14 months, infants are able to selectively imitate features of an action that are relevant to the goal of the action, and to ignore extraneous or idiosyncratic features. In other words, they interpret observed agents' behavior in terms of goals and rational strategies to attain those goals, and their imitative learning is guided by this interpretation.

These findings indicate that, in order to understand what the adults in their environment are up to, children apply agent-models which are structured by coherent relations among mental states, environmental constraints and affordances, and bodily movements. In addition, however, it also reveals that children have a tendency to align themselves actively with what those adults are up to – at least with what those adults appear to be up to when viewed through the lense of the agent-models which children apply to them.

This is perhaps most obviously the case for imitation, since it involves children manifestly acting to align themselves with a model. But it is also just as true, though perhaps more subtly, for other forms of cultural learning. In social-referencing situations, for example, which occur by around 9 months (Baldwin and Moses 1994) and perhaps as earlier as 5-6 months (Vaillant-Molina et al 2012), infants are sensitive to the objects of caregivers' emotional expressions and adopt those attitudes toward the same objects. That is, infants treat as dangerous or disgusting objects toward

which the caregiver has expressed fear or disgust. Here we see how the emotional evaluation of the object is influenced by a sophisticated understanding of the adult's visual attention in order to track not only the type of emotional state but also its intentional object, and to adopt an emotion of the same type toward the same intentional object.

Furthermore, there has been a great deal of research documenting how children's understanding of adults' attentional states and intentions is crucial for language acquisition. For example, if an adult announces her intention to 'find the toma' and then searches in a number of locations, scowling upon seeing some objects and smiling upon seeing one object, children will learn the new word 'toma' for the object the adult smiles at (Tomasello and Barton 1994). In fact, Southgate et al. (2010) found that 17 month-olds learned to apply a novel word ('sefu') to a toy that an adult falsely believed was hidden in a box if the adult pointed at that box and pronounced the word (after being out of the room while the toy was moved from the box to a different location) (for a review of several similar studies, see Tomasello 1999: 114-116). Moreover, as Tomasello (1999) has emphasized, language acquisition gets going around 12 months, at which time children engage in triadic interactions with an adult and an object, and exhibit pointing behavior to inform others of events they do not know about or to share an attitude about mutually attended events others already know about (Liszkowski et al 2007). As with the case of imitation, then, we see that language-learning also requires children to apply agent-models to others, and to actively align themselves with those agent-models.

These central forms of cultural learning also provide the foundation for additional, more nuanced and elaborate forms of cultural learning further downstream. Language, for example, enables a multitude of further effects upon cognitive development – "from exposing children to factual information to transforming the way they understand and cognitively represent the world by providing them with multiple, sometimes conflicting, perspective upon phenomena" (Tomasello 1999:163). In acquiring a natural language, children learn to partition the world into objects and events in a way specific to their culture, and to categorize the objects and events so partitioned, and to take different perspectives upon them. Thus, one object can be described as "the dog", "fido", "the dog over there", "the golden retriever", etc., and

one event can be described as "The dog bit the man", "The man was bitten by the dog", "Fido bit Daddy", etc. Which of these descriptions is appropriate depends upon the speaker's communicative goals and upon her evaluation of the listener's interests, knowledge, etc. The ability to switch among these perspectives and to deploy them flexibly is entrenched through early experiences of disagreeing with others who take different perspectives and in re-formulating utterances that have not been understood. Language also enables children to internalize rules, to memorize information and procedures, to talk about their own reasoning processes and other experiences, and to re-describe previously implicit procedural knowledge in explicit symbolic terms, thus enabling greater flexibility and systematicity.

Moreover, sensitivity to others' mental states is also of crucial importance in understanding all manner of norms that structure human sociality, since these derive their binding force not from physical facts but from agreement in people's attitudes about the statuses of entities, the entitlements and obligations they entail, etc. (cf. Searle 1995; Gilbert 1990). And there is evidence that children as young as two are sensitive to conventional ways of doing things or using objects, and treat these conventions as normatively binding (Rakoczy et al. 2008; Schmidt et al. 2010).

Finally, language is also important in shaping children's emerging understanding of themselves as unique individuals with their own preferences, as well as their own distinctive perspectives on and evaluations of events. In particular, as Fivush and Nelson (2004) have emphasized, talking about events together with adults helps children to become more fluent in linking their own emotions and other mental states with external events (e.g. I was sad because my balloon flew away). Moreover, shared remembering of past events, and of their own attitudes towards those events, helps children to develop a more reflective understanding of themselves as agents with distinctive perspectives on the world, some of whose thoughts and feelings can change over time, whereas some other mental states are much more lasting, such as some preferences, some beliefs, and many memories (Nelson 2003).

The upshot of these various forms of cultural learning is that infants and young children progressively refine their agent-models through interaction with others, and also, through active inference, increasingly conform to those models themselves.

7. Active inference and the reciprocal shaping of self-models

We have seen in the previous section that cultural learning can be conceptualized as a form of active inference by which children progressively converge upon the agent-models which they initially apply to interpret *other* agents. One effect of this is that children become increasingly similar to the adults in their culture. More precisely, they become increasingly similar to the adults around them *as those adults appear to them on the basis of the interpretations they generate by applying agent-models*. This effect ensures that the application of agent-models during development increases the predictive power of those agent-models. For children's use of agent-models will have shaped their own development such that they themselves approximate the intentional agents that they take others to be. And, if so, they will themselves be more easily intelligible for other interpreters who are also applying agent-models.

The flip-side of this is that *adults* also apply agent-models toward young children, and that this also plays a key role in structuring children's cognitive development, i.e. by setting up expectations for them to fulfill, and by acquainting them with culture-specific objects, practices, narratives, social roles, etc. For example, gender-specific interpretations of infant behavior (such as boys' cries more often being interpreted as expressions of anger as opposed to sadness) create expectations that children then conform to (Mameli 2001). Insofar as adults' interpretation of young children as potentially rational intentional agents facilitates children's enculturation, it also increases the predictive power of agent-models, and thus becomes, as Mameli (2001) puts it, a "self-fulfilling prophecy".

And of course this structuring effect[1] of agent-models continues into adulthood. Consider, for example, how we sanction others for departing from rational or social norms. Thus, as McGeer (2007) puts it, folk psychology is "a *regulative* practice, moulding the way individuals act, think and operate so that they become well-behaved folk-psychological agents: agents that can be well-predicted and explained

[1] Mameli (2001) coined the term "mindshaping" to denote this structuring effect of adults' intentional interpretations of children; Zawidzki (2013, chapter 2) generalizes it to include cases, like imitative learning, where an agent actively converges upon some external model – such as models of other agents, generated by taking the intentional stance toward them.

using both the concepts and the rationalizing narrative structures of folk psychology" (139, emphasis in original). Similarly, Alfano (2013) has recently argued that similar ideas can play an important role in shaping people's ethical behavior: labelling somebody as generous, for example, may lead her to think of herself as generous and/or to believe that others think of her this way, and motivate her to want to live up to this expectation and to maintain this image. As he puts it: "Trait attributions of the right sort function as self-fulfilling prophecies" (2013: 83).

Alfano's analysis also underscores the point that our interpretations of our own behavior and biographies have an influence on our own actions and choices, and are thus also sometimes self-fulfilling prophesies. Some research with split-brain patients serves as a dramatic illustration of this. One woman, whose right-hemisphere received the instruction that she should get up and leave the room, and who was then presented with a request to her left hemisphere to explain what she was doing (Gazzaniga, 1995: 1393), confabulated that she had gotten up in order to get a soda – and, crucially, she then really did go and get a soda. Thus, to borrow Zawidzki's gloss on this example: "whether or not our public self-interpretations are justified or true, we actively work to confirm them" (Zawidzki, 2013: 231).

The mechanism proposed here is circular, but not viciously so. It does not require that young children have the intentional states ascribed by interpreters applying agent-models, nor that agent-models have a high degree of predictive power when applied to very young children. It requires merely that young children have imprecise predictions to this effect about other agents and about themselves. If they do so, and if this provides them with role models to learn from and thereby to become more similar to, then they will develop in such a way that they subsequently become intelligible to others who also apply agent-models—including themselves. Moreover, becoming more like the agents in their culture they are modeling also adds to children's interpretive resources, which, in turn, enable more learning and thereby more similarity, etc. In this sense, agent-modeling and cultural learning constitute a positive feedback loop.

8. Is the Bayesian Self a Narrative Self?

As we have seen, the prediction error minimization framework presents an elegant and novel way of understanding how agents come to model the self, namely as a hierarchy of hidden endogenous causes, and of how human agents – through active inference in development and in social interaction throughout the lifespan – reciprocally align their selves with their agent-models. The upshot is increasing controllability and decreasing prediction error (for a fully developed view of alignment within the free energy principle, see Frith and Friston 2015, which essentially treats interaction as coupled oscillation that entails alignment). In this final section, we will flesh out this conception a little more by considering similarities it bears to narrative accounts of the self – but also some interesting differences.

The key feature of narrative accounts is that the self is conceptualized as an abstraction emerging from autobiographical narratives. As Dennett, for example, has expressed this view:

> A self…is an abstraction defined by the myriad of attributions and interpretations (including self-attributions and self-interpretations) that have composed the biography of the living body whose centre of narrative gravity it is. As such it plays a singularly important role in the ongoing cognitive economy of that living body, because, of all the things in the environment an active body must make mental models of, none is more crucial than the model the agent has of itself (Dennett 1991, p. 427).

And:

> And where is the thing your self-representation is about? It is wherever you are. And what is this thing? It's nothing more than, and nothing less than, your centre of narrative gravity (Dennett 1991, p. 429).

This approach is attractive insofar as it offers a way of conceptualizing the self as real and robust (like centers of gravity) without postulating mysterious immaterial entities. In fact, insofar as descriptions of the self and its traits can be used to identify behavioral and interactive patterns that facilitate predictions and to specify norms and ideals to conform to, it can also be a tool for structuring and controlling actions over time. Building on this point, Velleman (2006) articulates a narrative theory which emphasizes that the descriptions that constitute the self feed back into the decision-

making and planning of the agent, and thus have real causal powers, akin to self-fulfilling prophesies (see also Mackenzie, 2007; for discussion, see Menary 2008).

Despite the prima facie attraction of the basic idea of the self as a narrative center of gravity, the metaphor of the narrative also raises several important questions and challenges. First of all, the concept of a narrative implies linguistic representation, and perhaps also explicit recounting. This implies that pre-linguistic children and non-linguistic creatures do not have selves. And if they do have some more primitive form of self, it is difficult to see how a narrative self would arise from or otherwise relate to this more primitive form of self (Menary 2008). Secondly, as some (e.g. Strawson 2015) have objected, the concept of a narrative appears to imply coherence, whereas in fact life (and one's autobiographical memories) may seem more like a desultory collection of fragments than a coherently crafted narrative. Third, narrative accounts make it difficult to make sense of the possibility of error. While narratives about fictional characters are unconstrained by objective facts, it is possible to be mistaken in thinking that one is, for example, kind or well-loved.

Like narrativist accounts, the account we have been articulating resolutely avoids postulating an immaterial entity as the bearer of psychological properties or of experiences. And, also like narrativist accounts, it stops short of eliminating altogether the self. Instead, it avoids Hume's short-timescale fallacy by understanding the self as an abstraction that is useful in recognizing deeply hidden, longer term patterns among endogenous psychological properties, experiences and sensory inputs.

In contrast to narrative accounts, however, this account does not imply that the form of representation in question is linguistic (cf. narratives). Instead, it is a hierarchical Bayesian model that passes messages amongst its levels to reduce prediction error in the long term. One important consequence of this modification is that the account does not rely on narrative *coherence* (see Strawson's objection above). This is because the self-model is not a representation of events or sequences of events in the life of the agent. Instead, it captures the hidden patterns of endogenous causes of those events. Different models can however have different depths. For example, a very shallow self-model will have trouble subsuming regular events under longer-term regularities; conversely, an overly deep model will tend to explain irregular

happenstance as manifestations of longer-term regularities. This speaks to differing degrees of narrative coherence and the long-term task of arriving at a model of appropriate depth: a shallow model will have a fragmented narrative and a deep model will be more structured. Individual differences in the depth of the self-model are to be expected, given the dynamic nature of active self-alignment.

However, no matter how deep or shallow the self-model, in order for it to contribute to the reduction of prediction error, a different kind of coherence is always required. Specifically, the hierarchically superordinate layers of the model can only have predictive power if the parameters that they include are really somehow connected to each other and to the parameters of hierarchically subordinate layers. For example, the postulation of clusters of psychological properties like preferences, beliefs, and personality traits must make it possible to generate predictions about behavior, and the error in predictions about behavior help shape the clusters of psychological properties. In other words, there must be a coherent link between these psychological properties, the behavior, other causes in the world, and back on to sensory input. The notion of coherence is here given a precise meaning, namely in terms of how a system that minimizes prediction error will approximate Bayesian inference ensuring that all levels are maximally predictive of each other (given expected levels of noise in the self-model).

Finally, the account is better positioned than narrative accounts to explain how we can be wrong in our self-representations. The self is not merely the fictitious subject of a narrative. Instead, it is the set of endogenous causes being referred to by self-models, which are constrained by their embeddedness in a positive feedback loop constituted by worldly causes, bodily states, sensory states, internal states at various levels of causal depth, and active states.

While this constitutes a departure from the way in which narrativist theorists such as Dennett have explicitly characterized the self, it is a departure that in fact builds upon core insights of narrative theorists. In particular, the developmental feedback loop we have described can be understood as the product of applying Dennett's intentional stance theory to the development of self-understanding and the self: young children's use of the intentional stance enables them to learn from and thereby to become more

similar to the adults in their culture. As a result, they themselves become increasingly intelligible to other people taking the intentional stance. Thus, the intentional stance and cultural learning constitute a feedback loop that (partially) explains the reliability of the intentional stance. Michael (2015) has even argued that this developmental perspective on the intentional stance – contra Dennett – provides grist to the mill for a *causal* realist interpretation of the reference of intentional terms and of the self, insofar the causal interaction between intentional interpretations of behavior and cognitive development provides an anchor that links intentional terms to the endogenous hidden causes constituting the self.

This realist interpretation of intentional states and of self also means that – contra narrativism – correctness does enter into the picture with respect to the self. In particular, the agent can accumulate more or less evidence for a self-model over different time spans. Of course, it is often not a straightforward matter to evaluate the probability of an attribution of a psychological property, such as whether I am generous or hot-tempered. But the attribution is constrained by all the evidence that I have accumulated about myself, and can be contested by others who have access to an overlapping set of evidence about me. Thus, attributions of properties to the self can be more or less consistent with evidence, and they are open to revision according to how well they predict the evidence.

In summary, the predictive processing approach to brain function has a central role for both body and self, as inferred entities in the world that are represented in the agent's internal model of the world. The inference is based on filtering out patterns of causation that are best explained away by the model inferring that it is itself a cause of changes to its sensory input. Various aspects of this inferential account speak to the conception of the self as something more than mere inference but indeed something connected to existence and agency. These abstract ideas are not only consistent with recent research on self-conceptions in infants and young children, they allow us to see social interaction and cultural learning as key elements in the dynamic process of shaping one's self through action and interaction. Overall, we arrive at an idea of the embodied self, which harkens back to but also significantly revises well-known narrative accounts of the self.

References

Alfano, M. (2013) *Character as Moral Fiction*, Cambridge University Press

Apperly, I. 2011. *Mindreaders: The Cognitive Basis of Theory of Mind*, New York: Psychology Press, 2011.

Apperly I., and S. Butterfill. 2009. Do humans have two systems for tracking beliefs and belief-like states? *Psychological Review* 116 (4): 953-970.

Apps, M. A. J. and M. Tsakiris (2014). "The free-energy self: A predictive coding account of self-recognition." *Neuroscience & Biobehavioral Reviews* **41**(0): 85-97.

Baillergeon, R. Scott, R., Z. He. 2010. False-belief understanding in infants. *Trends in Cognitive Sciences* 14 (3): 108-115.

Baldwin, D. A., Baird, J. A., Saylor, M. M. & Clark, M. A. (2001) Infants parse dynamic action. *Child Development* 72: 708–17.

Baldwin, D. A., & Moses, L. J. (1994). Early understanding of referential intent and attentional focus: Evidence from language and emotion. *Children's early understanding of mind: Origins and development*, 133-156.

Behne, T., Carpenter, M., Call, J. & Tomasello, M. 2005. Unwilling versus unable: Infants' understanding of intentional action. *Developmental Psychology* 41:328–37.

Bischof-Köhler, D. 1991. The development of empathy in infants. In M. E. Lamb & H. Keller (Eds.), *Infant development: Perspectives from German speaking countries* (pp. 245- 273). Hillsdale, NJ: Lawrence Erlbaum Associates.

Botvinick, M. and J. Cohen (1998). "Rubber hands 'feel' touch that eyes see." *Nature* **391**(6669): 756-756.

Carey, S. 2009. *The Origin of Concepts* (1st ed.). Oxford University Press, USA.

Carhart-Harris, R. L. and K. J. Friston (2010). The default-mode, ego-functions and free-energy: a neurobiological account of Freudian ideas. *Brain* **133**(4): 1265-1283.

Dennett, D. 1981. 'Where Am I?' In: *Brainstorms: Philosophical Essays on Mind and Psychology,* Cambridge, Mass.: MIT Press, pp. 310-323.

Dennett, D. 1991. Real Patterns, *Journal of Philosophy* 88 (1): 27-51).
-- (1991), *Consciousness Explained* (London: Allen Lane).

den Ouden, H. E. M., K. J. Friston, N. D. Daw, A. R. McIntosh and K. E. Stephan (2009). "A Dual Role for Prediction Error in Associative Learning." *Cereb. Cortex* **19**(5): 1175-1185.

Eisenberg, N., & Fabes, R. A. 1998. Prosocial development. In N. Eisenberg (Ed.), *Handbook of child psychology, Vol. 3: Social, emotional, and personality development* (5th ed., pp. 701-778). New York: Wiley & Sons.

Eisenberg, N., Spinrad, T. L., & Sadovsky, A. 2006. Empathy-related responding in children. In M. Killen & J. G. Smetana (Eds.), *Handbook of moral development* (pp. 517–549)

Fivush, Robyn, and Katherine Nelson. "Parent–child reminiscing locates the self in the past." *British Journal of Developmental Psychology* 24.1 (2006): 235-251.

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

Fotopoulou, A (2012). Towards psychodynamic neuroscience. In *From the couch to the lab: Trends in psychodynamic neuroscience*. (ed.) A. Fotopoulou, M. Conway, & D. Pfaff. New York: Oxford University Press. Pp. 25-47.

Friston, K. (2010). "The free-energy principle: a unified brain theory?" *Nat Rev Neurosci* **11**(2): 127-138.

Friston, K. (2011). Embodied Inference: or "I think therefore I am, if I am what I think". *The Implications of Embodiment*. W. Wolfgang Tschacher and C. Bergomi. Sussex, Imprint Academic.

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

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

Friston, K. J., J. Daunizeau and S. J. Kiebel (2009). "Reinforcement Learning or Active Inference?" *PLoS ONE* **4**(7): e6421

Frith, C., Friston, K. (2015). A duet for one. *Consciousness and Cognition* 36: 390-405.

Gazzaniga, M. 1995. Consciousness and the cerebral hemispheres. In: *The cognitive neurosciences*, ed. M. Gazzaniga. MIT Press.

Gergely, G., Bekkering, H. and I. Kiraly. 2002. Rational imitation in preverbal infants. *Nature* 415: 755.

Gergely, G., Szilvia Biro, S., Kiro, O. and M Brockbank. 1999. Goal attribution without agency cues: the perception of 'pure reason' in infancy. *Cognition* 72: 237–267.

Gergely, G. & Csibra, G. 2003. Teleological reasoning in infancy: The naïve theory of rational action. *Trends in Cognitive Sciences*, 7(7), 278-292.

Gilbert, M. 1990. Walking together: a paradigmatic social phenomenon Midwest Studies in Philosophy.

Helmholtz, H. v. (1867). *Handbuch der Physiologishen Optik*. Leipzig, Leopold Voss.

Hoffman, M. 2000. *Empathy and Moral Development: Implications for Caring and Justice*. New York: Cambridge University Press.

Hohwy, J. (2007). "The sense of self in the phenomenology of agency and perception." *Psyche* **13**(1).

Hohwy, J. (2013). *The Predictive Mind*. Oxford, Oxford University Press.

Hohwy, J. (2015). The neural organ explains the mind. *Open MIND*. T. Metzinger and J. M. Windt. Frankfurt am Main, MIND Group**:** 1-23.

Hume, D. (1739-40). *A Treatise of Human Nature*. Oxford, Oxford: Clarendon Press.

Knobe, J. 2006. The Concept of Intentional Action: A Case Study in the Uses of Folk Psychology. *Philosophical Studies*. 130: 203-231.

Lenggenhager, B., T. Tadi, T. Metzinger and O. Blanke (2007). "Video Ergo Sum: Manipulating Bodily Self-Consciousness." *Science* **317**(5841): 1096.

Lewis, M. & Brooks-Gunn, J. 1979. *Social cognition and the acquisition of self*. New York: Plenum Press. p. 296.

Limanowski, J. and F. Blankenburg (2013). "Minimal Self-Models and the Free Energy Principle." *Frontiers in Human Neuroscience* **7**.

Liszkowski, U., Carpenter, M. and M. Tomasello. 2007. Pointing out new news, old news and absent referents at 12 moths of age. *Developmental Science* 10 (2): 1-7.

Mackenzie C. (2007), 'Bare personhood? Velleman on selfhood', *Philosophical Explorations*, **10** (3), 263–81.

Mameli, M. 2001. Mindreading, mindshaping, and evolution. *Biology and Philosophy* 16: 597-628.

Mathys, C., J. Daunizeau, S. Iglesias, A. O. Diaconescu, L. A. E. Weber, K. J. Friston and K. E. Stephan (2012). "Computational modeling of perceptual inference: A hierarchical Bayesian approach that allows for individual and contextual differences in weighting of input." *International Journal of Psychophysiology* **85**(3): 317-318.

Mathys, C. D., E. I. Lomakina, J. Daunizeau, S. Iglesias, K. H. Brodersen, K. J. Friston and K. E. Stephan (2014). "Uncertainty in perception and the Hierarchical Gaussian Filter." *Frontiers in Human Neuroscience* **8**

Menary, R. 2008. Embodied narratives. *Journal of Consciousness Studies*, **15**, No. 6, 2008, pp. 63–84

Metzinger, T. (2004). *Being no one: The self-model theory of subjectivity*, MIT Press (MA).

Metzinger, T. (2009). *The Ego Tunnel*. New York, Basic Books.

Metzinger, T. (2011). The no-self alternative. *The Oxford Handbook of the Self*. (S. Gallagher (ed.). Oxford: Oxford University Press.

Meltzoff, A.N. 1995. Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology*, 31: 838–50.

Michael, J. (2015) Cultural learning and the reliability of the intentional stance, in: Munoz-Suárez, C. (Ed.) *Content and Consciousness 2.0: Four Decades After* (*Studies in Brain and Mind Series*, vol 7): Springer, pp: 163-184.

Moutoussis, M., P. Fearon, W. El-Deredy, R. J. Dolan and K. J. Friston (2014). "Bayesian inferences about the self (and others): A review." *Consciousness and Cognition* **25**(0): 67-76.

Nelson, K. (2003). Narrative and self, myth and memory: Emergence of the cultural self. In: Fivush, R., & Haden, C. A. (Eds.). (2003). *Autobiographical memory and the*

construction of a narrative self: Developmental and cultural perspectives. Psychology Press.

Pearl, J. (2000). *Causality*. Cambridge, Cambridge University Press.

Prinz, J. 2005. Imitation and moral development. In *Perspectives on Imitation*. Ed. Hurley, S. and N. Chater. Cambridge: MIT Press: 267-282. Psillos, S. 1999. *Scientific Realism: How Science Tracks Truth*, London: Routledge

Rakoczy, H., Warneken, F. and M. Tomasello. 2008. The Sources of normativity: children's understanding of the normative structure of games. *Developmental Psychology*, 44 (3): 875-881.

Reddy, V. 2003. On being the object of attention: implications for self–other consciousness. *Trends in Cognitive Sciences*, *7*(9), 397-402.

Russell, J. (1995). At two with nature: Agency and the development of self-world dualism. *The body and the self*. J. L. Bermudez, A. J. Marcel and N. M. Eilan. Cambridge, Mass., MIT Press**:** 127-151.

Schmidt, M. Rakoczy, H. and M. Tomasello. 2010. Young children attribute normativity to novel actions without pedagogy or normative language. *Developmental Science* 14 (3): 17 530-539.

Searle, J. 1995. *The Construction of Social Reality*. New York: Free Press.

Senju, A., and G. Csibra. 2008. Gaze-following in human infants depends on communicative Signals. *Developmental Science* 18 (9): 678-671.

Seth,A.K.,Suzuki,K.,andCritchley, H. D.(2011).An interoceptive predictive coding model of conscious presence. *Frontiers in Psychology* 2:395. doi:10.3389/fpsyg.2011.00395

Seth, A. K. (2015). The Cybernetic Bayesian Brain. *Open MIND*. T. K. Metzinger and J. M. Windt. Frankfurt am Main, MIND Group.

Shea, N. 2009. Imitation as an Inheritance System. *Philosophical Transactions of the Royal Society B*, 364: 2429-2443.

Song, H, Onishi, K., Baillargeon, R., Fisher, C. Can an agent's false belief be corrected through an appropriate communication? Psychological reasoning in 18-month-old infants. Cognition, 109 (2008), pp. 295–315

Southgate, V., Chevallier, C., and Csibra G. 2010. Seventeen-month-olds appeal to false beliefs to interpret others' referential communication. *Developmental Science* 13 (6): 917-912.

Sterelny, K. 2003. *Thought in a Hostile World*. Oxford: Blackwell. Stern , D. (1985/1998). *The interpersonal world of the infant: A view from psychoanalysis and developmental psychology*. New York: Basic Books.

Strawson G. (2004), 'Against narrativity', *Ratio (new series)*, **XVII** (4), pp. 428–52.

Surian, L., Caldi, S., and D. Sperber. 2007. Attribution of Beliefs by 13-Month-Old Infants. *Psychological Science 18* (7): 580-586.

Thornton, T. (2003), 'Psychopathology and two kinds of narrative account of the self', *Philosophy Psychiatry and Psychology*, **10**, pp. 361–67.

Velleman J.D. (2006), *Self to Self: Selected Essays* (New York: Cambridge University Press).

Velleman J.D. (2007), 'Reply to Mackenzie', *Philosophical Explorations*, **10** (3), pp. 283–90.

Tomasello, M. 1999. *The Cultural Origins of Human Cognition*. Cambridge: Harvard University Press.

Tomasello, M. and M. Barton. 1994. Learning words in non-ostensive contexts. *Developmental Psychology* 30: 639-650.

Tomasello, M. Carpenter, M., Call, J., Behne, T. and H. Moll. 2005. Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences* 28: 675–735.

Tomasello, M., Kruger, A., and H. Ratner. 1993. Cultural Learning. *Behavioral and Brain Sciences* 16, 495-552.

Topal, J., Gergely, G., Miklosi, A., Erdohegyi, A., Csibra, G. 2008. Infants' perseverative search errors are induced by pragmatic misinterpretation. *Science* 321: 1831-3.

Vaillant-Molina, M., & Bahrick, L. E. (2012). The role of intersensory redundancy in the emergence of social referencing in 5½-month-old infants. *Developmental psychology*, *48*(1), 1.

Vaish A, Carpenter M, Tomasello M (2009) Sympathy through affective perspective-taking and its relation to prosocial behavior in toddlers. *Developmental Psychology* 45(2):534– 543.

Wimmer, H., and J. Perner. 1983. Beliefs about beliefs: representation and constraining function of wrong beliefs in children's understanding of deception. *Cognition* 13 (1): 103-128.

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, 69(1), 1–34. doi:10.1016/S0010-0277(98)00058-4

Woodward, J. (2003). *Making Things Happen*. New York, Oxford University Press

Zahn-Waxler, C., Radke-Yarrow, M., Wagner, E., & Chapman, M. 1992. Development of concern for others. *Developmental Psychology, 28*(1), 126-136.

Zawidzki, T. 2013. *Mindshaping – A New Framework for Understanding Human Social Cognition*. Cambridge, MA: MIT Press.

Zawidzki, T. 2008. The function of folk psychology: mind reading or mind shaping? *Philosophical Explorations* 11(3): 193 – 210.