Published: In *ASCS09: Proceedings of the 9th Conference of the Australasian Society for Cognitive Science*. Edited by Wayne Christensen, Elizabeth Schier, and John Sutton. Sydney: Macquarie Centre for Cognitive Science. [This version has some updated references]

The hypothesis testing brain: some philosophical applications

Jakob Hohwy
Monash University

Abstract
According to one theory, the brain is a sophisticated hypothesis tester: perception is Bayesian unconscious inference where the brain actively uses predictions to test, and then refine, models about what the causes of its sensory input might be. The brain's task is simply continually to minimise prediction error. This theory, which is getting increasingly popular, holds great explanatory promise for a number of central areas of research at the intersection of philosophy and cognitive neuroscience. I show how the theory can help us understand striking phenomena at three cognitive levels: vision, sensory integration, and belief. First, I illustrate central aspects of the theory by showing how it provides a nice explanation of why binocular rivalry occurs. Then I suggest how the theory may explain the role of the unified sense of self in rubber hand and full body illusions driven by visuotactile conflict. Finally, I show how it provides an approach to delusion formation that is consistent with one-deficit accounts of monothematic delusions.

1. Introduction

From inside the skull, the brain must figure out what out in the world, or in the body, cause its sensory input. This is a difficult, indeed intractable, problem because it requires an inference from effects, the sensory input, to causes, the states of affairs in the world. Causes in the world occur in many contexts and interact in many ways so there is no easy mapping from the world to the input. The same input can be caused by many different things, and the same things may cause different kinds of input.

One way to deal with this kind of problem is to turn things around such that, instead of attempting an inference from effect to cause, one makes assumptions about what the cause could be and use a model of those causes to generate an estimate of what the sensory effects should be, if indeed those are the causes. One can then compare actual and expected input and, if the fit is good, infer that those were probably the causes.

On this kind of approach, prediction is crucial. The system in question, for example the brain, would constantly be trying to look ahead and predict what the sensory input will be like. Perception is then the currently best prediction. Perhaps the system could utilise this to overcome the processing delays that would occur if it could only begin make inferences about the causes in the environment after having received the sensory input (Helmholtz 1860; Gregory 1980; Gregory 1998).

This appeal to generative models and prediction – hypothesis testing – as the guiding principle for the brain has been around for a long time but is, I think, gaining in popularity. It is, in various guises, becoming more and more mainstream in machine learning, AI, and computational neuroscience (Mumford 1992). It is also beginning to coalesce into a dominant stream in cognitive neuroscience and areas of psychology (Kersten, Mamassian et al. 2004). Here, I mainly use work by Karl Friston and

Published: In *ASCS09: Proceedings of the 9th Conference of the Australasian Society for Cognitive Science*. Edited by Wayne Christensen, Elizabeth Schier, and John Sutton. Sydney: Macquarie Centre for Cognitive Science. [This version has some updated references]
colleagues to exemplify this approach and then discuss its various applications to areas of philosophical interest (Friston 2003; Friston 2005; Friston and Stephan 2007).

2. Prediction error minimisation – a simplified account

Viewed in a very simplistic manner, the brain has bottom-up signals and top-down signals. On the approach which emphasises generative models, bottom-up signals embody sensory input and top-down processing embodies predictions about the input generated on the basis of models of probable causes. The basic mechanism for this system is to minimise prediction error. Prediction error is the part of the bottom-up signal which is not predicted by the current model's top-down predictions. The more prediction error, the worse the predictions; and if no predictions occur at all, then the entire incoming signal is conceived as prediction error. If the brain had a procedure for minimising prediction error, its predictions would continually improve (barring evil demon scenarios), and it would be perceiving the world aright. This is perception. Models can then be revised in the light of prediction error. Model parameters are continually updated until prediction error is minimised, up to expected levels of noise. This is learning (Friston 2003; Friston and Stephan 2007). Once the right parameters have been found, they can be precisified such that more fine-grained prediction error leads to revision of the model. This can be viewed as attention (Friston 2009). The basic message is that the brain performs one main task: it minimises prediction error. This can, it is claimed, account for perception, learning and attention.

On this account it is wrong to say that perception is a matter of top-down processing (perhaps as in some "New Look" approaches in psychology). It is also wrong to say that it is a matter of bottom-up processing (perhaps as in some responses to the New Look). The truth is in the middle: perception is what happens in the "meeting" of top-down and bottom-up processing in the brain. The best way of looking at this involves turning the usual labelling of top-down signals as feed-back on bottom-up signalling on its head: "Cortical hierarchies are trying to generate sensory data from high-level causes. This means the causal structure of the world is embodied in the backward connections. Forward connections simply provide feedback by conveying prediction error to higher levels. In short, forward connections are the feedback connections." (Friston 2005: 825).

3. Core aspects of prediction error minimisation in the brain: hierarchy, agency, explaining away

A good heuristic for appreciating this approach to the brain is Bayesian probability theory. The current generative model is the model with the highest prior probability. This model is used to generate predictions about what the next sensory input will be. Being good at this translates to having a high likelihood. High prior and likelihood means high posterior probability. The model with the highest posterior wins and determines perceptual inference. This is a very intellectualist notion of perception but the suggestion is of course not that networks in the brain directly know and apply Bayes' rule or that perceptual inference is in any sense conscious. More research is needed to determine whether and how the brain implements such a Bayesian scheme of unconscious perceptual inference and there is indeed interesting work being done in this regard. One intriguing but also controversial idea is to assimilate the information theoretical notion of free energy (the sum of squared prediction errors) to the thermodynamic notion of free energy, which would make probabilistic prediction error minimisation a matter of dynamics of far from equilibrium open systems

Published: In *ASCS09: Proceedings of the 9th Conference of the Australasian Society for Cognitive Science*. Edited by Wayne Christensen, Elizabeth Schier, and John Sutton. Sydney: Macquarie Centre for Cognitive Science. [This version has some updated references]
(Friston and Stephan 2007; Hohwy, Roepstorff et al. 2008). Here I take no stance on this kind of high level issue. I will assume a Bayesian understanding of generative models implemented with predictive coding. I will however briefly describe some general characteristics of the PEM scheme and also attempt to tie this to some general aspects of Bayesian causal networks concerning the notion of 'explaining away'. Later, I will use these in considering various areas of philosophical interest.

Hierarchy

So far, the story is very simple: as if there is just an input level connected to a model level. It is however a crucial part of the PEM scheme that the brain is ordered in a hierarchy of overlapping pairs of bottom-up and top-down levels. To illustrate (and oversimplify): level 1 is the basic input level, level 2 tries to predict activity at level 1 and is also input level to level 3. Level three predicts activity in a top-down manner for level 2 and is paired upwards with level 4, and so on. Different levels allow the brain to build up representations of environmental causes from basic stimulus attributes to more and more abstract and perspective invariant properties. Importantly, this happens at varying time scales beginning with milliseconds, over hundreds of milliseconds, to seconds, minutes and on to months and stable rules. The hierarchy allows more subtle and hidden causal chains to be represented and in turn work as control parameters for lower level, faster moving attributes. This hierarchical structure is plausibly borne out anatomically in the brain with time scales getting progressively longer as one moves forwards along cortex from the occipital towards the frontal lobe (Friston 2008; Kiebel, Daunizeau et al. 2008). If prediction error minimisation works in this way then the brain must recapitulate the causal structure of the world (Friston 2008). Intriguingly, this means that the brain is indeed, and must be, a mirror of nature (excluding sceptical scenarios of evil demons).

Two salient aspects are noted. First, the level of detail and fineness of temporal grain decreases as one goes up in the hierarchy. Second, the prediction horizon shrinks as one goes down the hierarchy. High level causal models can predict things a long time in advance but are unable to directly predict fast low level sensory input in any great detail. Conversely, low level models can predict sensory input with great detail but only a few milliseconds or seconds in advance. This is commonsense too: you may know in rough outline what tomorrow will be like with lots of familiar causal interactions but you don't know exactly what you will perceive, from what perspective, and precisely when.

Agency

Minimising prediction error is minimising surprise since the better you can predict things the less you will be surprised. So the task for the brain is to make life less surprising – in particular to avoid dangerous, man-eating surprises. At this point a common objection is that this makes a mockery of thrill seekers and fear of death since they seem to involve cases where surprise is actively sought. The answer is that thrill seekers predict the rush of hard-to-predict sensory data they get when, say, bungy jumping. In particular they expect a surge in the autonomic system and will go to considerable lengths to ensure this surge happens (and this kind of predictive system is even more strongly high-jacked in addiction). Likewise, it is mistaken to think that death minimises prediction error. Using the thermodynamic idiom, death comes with a massive increase in prediction error since a dead organism is no longer able to attenuate input – and decomposes.

Published: In *ASCS09: Proceedings of the 9th Conference of the Australasian Society for Cognitive Science*. Edited by Wayne Christensen, Elizabeth Schier, and John Sutton. Sydney: Macquarie Centre for Cognitive Science. [This version has some updated references]

In fact, a strength of recent versions of this type of theory is that the incorporation of agency as a crucial element of prediction error minimisation. Agency is just a matter of putting the organism in a situation where prediction error is minimised without necessarily having to revise model parameters (Friston and Stephan 2007). Notice that such prediction of sensory input given agency depends on a *current state estimate*. If you don't know where you are and what your configuration is, then you cannot plan how to act such as to minimise prediction error. Notice also that, since perceptual inference depends on agency, we should expect interference with normal perceptual inference when there is interference with an organism's agency.

Explaining away

One morning you observe your lawn is wet. You consider two explanatory models: that it has been raining or that the sprinklers have been on. Each model explains the evidence of your wet lawn equally well. If one of them has a higher prior, then you might infer it is more probable but for now we can assume priors are balanced. Now you observe your neighbour's lawn is wet too. The Rain model explains that well, but the Sprinkler model doesn't since the sprinklers being on would only make your lawn wet. The Rain model accounts for all the evidence leaving no evidence behind for the Sprinkler model to explain. Even though the Sprinkler model did increase its probability in the light of the first observation, it seems intuitive right to say that its probability is now returned to near its prior value. The model has been explained away. This is commonsense but also puzzling since the two models are conditionally independent (given the first observation) but become dependent given the second observation. It is an aspect of causality which is particularly difficult to model quantitatively (Wellman and Henrion 1993; Jensen and Lauritzen 2000; Pearl 2000).

It seems natural to expect this kind of explaining away pattern in a brain governed by PEM. Top-down predictions attenuate the predicted input and leave as bottom-up signal only the unpredicted part. This means that alternative models will not be able to account for the attenuated parts of the input such that, as the winning model gets stronger, alternative models will weaken even though in principle they could account for the evidence. (Slightly confusingly, the 'explaining away' idiom is sometimes used in the PEM literature for the attenuating or accounting for the input).

4. Prediction error minimisation and philosophy.

I have described, in simple terms, a general approach to how the brain solves the problem of perception: it minimises prediction error. This explains not only perception but also learning and attention, and can plausibly be applied to emotion and bodily sensation as well. I find this approach extremely promising and I expect it to become dominant in the years to come. I am not alone in this belief as is indicated by the prominent neuroscientist Stan Dehaene, who says "It is the first time that we have had a theory of this strength, breadth and depth in cognitive neuroscience…Most other models, including mine, are just models of one small aspect of the brain, very limited in their scope. This one falls much closer to a grand theory." (*New Scientist*, 2658: 30-33, 2008).

Here I will work within this general framework. It is close to a universal theory of the mind and brain so we should expect it to have wide-ranging implications. This includes implications for domains of philosophical interest. My project here is

Published: In *ASCS09: Proceedings of the 9th Conference of the Australasian Society for Cognitive Science*. Edited by Wayne Christensen, Elizabeth Schier, and John Sutton. Sydney: Macquarie Centre for Cognitive Science. [This version has some updated references]
therefore to explore how the framework can be extended to philosophically interesting areas. There are at least two worries associated with this. The first one is that, as described here, the theory is at risk of trivializing what it is meant to explain. For any behaviour, it is possible to find a context in which it can be seen as minimizing prediction error (for example, being a drug addict is hardly something that minimises prediction error over time, since it often leads to death. But it is easy to explain nevertheless: drug addicts predict *huge* rewards from taking drugs and minimize prediction error by searching out dugs at any cost). I think this is close to the truth but it is not on its own, without more specific computational and empirical evidence, a very good response to the objection. The second, related, worry is that there is no direct, easy link from the prediction minimisation framework to philosophical problems. If this is ignored, then the application to philosophy becomes too general and bland. Some intermediate steps are required in order to make a more substantial and hopefully fruitful connection with the philosophical.

My method here is to use a broadly neurophilosophical approach to make contact between neuroscience and some philosophically interesting areas. Specifically, I consider how a range of empirical studies, which are philosophically relevant, can be understood in the light of prediction error minimization. I consider three areas: perception, where my case study is binocular rivalry; self-awareness, where my case study is rubber hand and full body illusions; and belief, where my case study is delusion formation. To make things concrete rather than general, I make most use of the three aspects of PEM I noted above: (a) we minimise prediction error, and thereby explain away competing hypotheses; (b) PEM is implemented in a cortical hierarchy in which there is a trade-off between detail and time scale; (c) agency is crucial to PEM, it aids perceptual inference and depends on reliable estimates of the current state of the system.

5. Binocular rivalry and explaining away

Binocular rivalry is an extremely stable visual effect where conscious perception alternates between two stimuli, one presented to each eye (Alais and Blake 2005). It is especially important philosophically because it is one of the dominant paradigms in consciousness science (Frith, Perry et al. 1999; Blake and Logothetis 2002; Hohwy 2007). It is thus a crucial opportunity for gaining detailed understanding of how consciousness is being studied scientifically. It is also, more broadly, an important tool for understanding how the brain represents the world, quite apart from the thorny issue of consciousness: it is a phenomenon that allows us to understand representation by seeing how the system works in unusual situations.

The mechanism behind binocular rivalry is however not known, despite scientific research stretching back over centuries. There are now a number of very sophisticated and interesting computational models of rivalry (Noest, van Ee et al. 2007; Wilson 2007; Gigante, Mattia et al. 2009). These models are built with the explicit purpose of accounting for the alternation typical of rivalry as well as some of the more detailed psychophysical findings associated with the phenomenon. The results are very impressive but are not the upshot of a general theory of cortical representation. This means that none of them throw much light on *why* a representational system such as the brain should display binocular rivalry.

Published: In *ASCS09: Proceedings of the 9th Conference of the Australasian Society for Cognitive Science*. Edited by Wayne Christensen, Elizabeth Schier, and John Sutton. Sydney: Macquarie Centre for Cognitive Science. [This version has some updated references]
PEM has some promise for providing such a broader perspective on rivalry because PEM is the kind of broad theory of how the brain represents. If it can be shown that rivalry is a natural upshot of PEM, then not only do we understand something about the nature of rivalry, we also support PEM itself. In recent work, we suggested that the notion of explaining away indeed does predict binocular rivalry (Hohwy, Roepstorff et al. 2008). The basic idea is that PEM works by attenuating the incoming, bottom-up signal. This can only happen for the current winner of perceptual inference. If the eyes are presented with a house and a face respectively and the face model is the winner, only the face input will be attenuated. This results in an unusual, "unecological" situation where a very large prediction error is unaccounted for but cannot be explained by any good model. This, we speculate, destabilises a system which is geared for minimising just such prediction error. This could lead to perceptual alternations and also helps explain a wide range of psychophysical phenomena observed in rivalry (a mathematical model based on explaining away substantiates this, see (Dayan 1998); our research group is currently expanding this model and we apply it to a wide range of rivalry data in work in preparation).

Our account has a couple of further aspects. We add that in addition to the destabilisation by the large prediction error from the suppressed stimulus, it is likely that the visual system expects change in sensory attributes such that the currently dominant model, or attractor, experiences a decreasing prior over time. We explain the absence of fusion (when stimuli are clearly different and overlapping in space and time) in terms of the low prior probability of the same thing occupying the same location in space and time. Thus we have never experienced, and never will, a face and a house in the same spatiotemporal location. This means that there is no model for the fused percept and there is reason to believe that no obvious revision to existing models can allow this (much here depends on stimulus properties, of course, but this hypothesis is consistent with the empirical findings in the field). It follows from this that two stimuli that can plausibly co-occur spatiotemporally will not lead to rivalry.

There is more to be said about this suggestion regarding rivalry (Hohwy, Roepstorff et al. 2008); the suggestion itself has received some support (Song and Yao 2009), and is currently being tested both in our lab and elsewhere. The strength of the suggestion is, in my view, that it goes beyond the usual models of rivalry. A priori, it is obvious that any account of rivalry will have to have, on the one hand, an element of selection combined with reciprocal inhibition of the suppressed stimulus, and on the other hand, an element of fatigue, such that alternation can ensue. This is in fact not hard to achieve and the best models are the ones that can, in addition to reciprocal inhibition and fatigue, demonstrate a degree of biological plausibility in terms of cell function and small network properties. Our model shows why a representational system as a natural upshot has reciprocal inhibition and something like fatigue. It does this in what seems to me an attractively surprising way, by turning things around such that the *suppressed* stimulus is in fact associated with *increased* activity, conceived as prediction error.

6. Rubber hand and out of body illusions: explaining away the body

One of the most striking and exciting illusions is the rubber hand illusion (Botvinick and Cohen 1998). The participant's right hand is concealed from view. A rubber hand is placed in a plausible position in front of her. During synchronous tapping or

Published: In *ASCS09: Proceedings of the 9th Conference of the Australasian Society for Cognitive Science*. Edited by Wayne Christensen, Elizabeth Schier, and John Sutton. Sydney: Macquarie Centre for Cognitive Science. [This version has some updated references]
stroking to the concealed hand and the visible rubber hand the participant will experience that the touch she can feel is produced on the rubber hand by the hand tapping it. In an exciting turn of events this illusion was extended to the whole body. With the help of virtual reality goggles, a touch delivered to the body can be felt as if produced away from the body, either outside of peripersonal space (Lenggenhager, Tadi et al. 2007), or in peripersonal space (Ehrsson 2007). The presence and strength of the illusion is gauged by introspective report and by a number of more objective measures such as skin conductance response, displacement and other more sophisticated psychophysical measures (Lenggenhager, Mouthon et al. 2008; Aspell, Lenggenhager et al. 2009).

Work in preparation in our own lab uses a combination of the original rubber hand paradigm but augmented with VR goggles (Hohwy and Paton 2010). This allows us to strengthen the illusion because there is then no displacement of the concealed hand relative to the rubber hand. This means that in creating the illusion of touch to the rubber hand, the brain does not have to overcome proprioceptive information.

Aspects of PEM are promising for understanding what goes on in the rubber hand illusion. The initial situation is that there is conflicting visual and tactile sensory input: vision suggests touch is delivered to a visible foreign arm, tactile sensation suggests touch is delivered to a concealed own arm. The brain needs to decide which model best accounts for this evidence. The *true* model is that touch and vision occur in different locations. The *false* model is that touch and vision occur in the same location. Something in the situation makes the false model win. One could appeal fairly directly to Bayes to explain this but such explanations are not always very satisfactory. It is always seems possible to construe a context for application of Bayes rule which will give the desired result. In this case it seems that both models account equally well for the sensory evidence, so likelihoods are matched. Perhaps one could then say that there is a higher prior for touch being where vision suggests it is, based on prior learning of this association. The problem is that we could probably find an equally plausible Bayesian account of cases where the illusion fails to occur. The simple Bayesian suggestion also ignores the obvious fact the participants who experience the illusion are very well aware that the arm on which they feel a touch is not their real arm, meaning that there is a very low prior for touch being felt there. One way to go is a quantitative Bayesian account (Schwabe and Blanke 2008). I appeal to some of the other aspects of PEM to suggest why the illusion may occur.

A central aspect of the illusion is that the brain overrides the prior knowledge that one's arm is not made of rubber in favour of the model on which one can feel a touch on the rubber hand. There may be a number of factors, deriving from PEM, which make this happen. The first factor is that synchronous tapping is something that calls for explanation: it is very unlikely that seen and felt tapping can be in synchrony by coincidence. A good explanation would normally be that vision and touch are have co-located effects. An equally good but computationally more demanding solution is that the synchrony is explained by a common cause producing effects in different locations. The latter explanation would be correct in this case because in fact the experimenter is a common cause of the observed effects. It seems likely that there is a bias in favour of co-location explanations over common cause explanation, at least in the sensory domain. And this bias would be a factor in explaining the illusion. A similar bias would be at play in the well-known ventriloquist effect where a common

Published: In *ASCS09: Proceedings of the 9th Conference of the Australasian Society for Cognitive Science*. Edited by Wayne Christensen, Elizabeth Schier, and John Sutton. Sydney: Macquarie Centre for Cognitive Science. [This version has some updated references]
cause, the ventriloquist, is not experienced as the cause, even if she is known to be the cause (Körding, Beierholm et al. 2007).

The second factor that may be implicated stems from PEM's reliance on agency. The two competing models differ in respect of unity of self. On the true model there is uncertainty about the current state of the body – the self-model (Metzinger 2009) – because it is divided between multiple locations. On the false, illusory model, there is less uncertainty about it: the model says where the body is, even if it is a weird rubbery body. It may be that there is a deep-seated prior in favour of having a unified self-model, if at all possible because a unified self-model is what best allows computation of the system's current state such that action can be undertaken. In this respect it may be that the brain's motto, as it were, it that it is better to have a rubber hand than a disunified self.

The rubber hand illusion suggests that, at this level of multisensory integration, the prior model of one's actual body – one's body image – can relatively easily be explained away. Prior research on the rubber hand illusion is conflicted on the depth of this. Some studies show that it is easy to dispense altogether with the body such that touch can be felt on a bare table top or outside peripersonal space (Armel and Ramachandran 2003). Other studies show that touch cannot really be felt on non-hand rubber objects (Tsakiris and Haggard 2005). With PEM and the notion of explaining away in mind, the prediction would be that as the illusion progresses in time, the prior body schema would be explained away more and more. Specifically, we would expect illusions that incorporate touch on non-body objects, touch without a visible body being touched, and touch in empty space. We would in particular expect this in the paradigm employed in our lab, where there is overlap in personal space between the seen and real arm. Preliminary data suggest that this is indeed the case and we suggest a resolution to the previous conflicting data such that non-body objects can indeed be incorporated into the illusion but primarily after participants have experienced the standard illusion (and possibly also primarily when illusion onset does not have to overcome divergent proprioceptive information; Hohwy Paton 2010).

If we step back a bit we can see that the rubber hand illusion arises in multisensory integration of touch and vision where vision captures touch on the fake hand. I have suggested a number of factors, which go with PEM, that contribute to the occurrence of the illusion. However, it is also important that the experimental paradigm restricts the opportunity for reality testing: participants are not allowed to move or touch their own arm and the real arm remains concealed from them. This contributes to the maintenance of the illusion because as soon as they move the arm the illusion breaks (unless the rubber hand moves too (Slater, Marcos et al. 2009)). Again, moving the arm would introduce more evidence as sensory input, which the false rubber hand model cannot explain. The illusion occurs, that is, under restricted reality testing conditions.

The resulting picture we have is then this: Within restricted reality testing conditions a unified self-model matters (progressively) more than a specific body-representation. I think this gives new insight into the notion of bodily self-awareness, captured well in a kindred, and impressive, study of bodily self-location: "[O]nline processing of body-related multisensory information in the brain is more like ongoing puzzle solving of which the normally experienced embodied self-location is just a fragile and

Published: In *ASCS09: Proceedings of the 9th Conference of the Australasian Society for Cognitive Science*. Edited by Wayne Christensen, Elizabeth Schier, and John Sutton. Sydney: Macquarie Centre for Cognitive Science. [This version has some updated references]
only temporarily stable solution, which is a setting that is naturally suited for the Bayesian approach to sensory information processing." (Schwabe and Blanke 2008).

This account of rubber hand illusions and full body illusions is striking because it reveals the flimsiness of a very familiar aspect of our experience, namely our bodily self-awareness. To experience the world – to engage in perceptual inference – we do not need a firm grasp of a material body. Indeed, once subjects are deep into the illusion, many feel that touch can be felt by an invisible hand on an invisible arm, when their real arm is touched and they see nothing but an empty table top in the VR goggles (Hohwy & Paton 2010).

It is tempting to conclude that the illusions arise when the participant is wholly passive and that therefore being able to engage in agency is fundamental to having a normal bodily self-awareness, or self-model (Metzinger 2009). From the perspective of PEM, this is not quite right. It is true that there is no overt agency, since we prohibit movement. But there is active prediction minimisation going on: the movement of the finger touching the rubber hand is constantly being predicted; the eyes move around, searching the visual field and thus conditioning inference on agency (and participants want to move as well – many to the point where they have to hold their real arm down with their other hand to prevent it from moving). It seems better to say then that there is restricted but not abolished agency, which again restricts PEM and maintains the illusion. It is a nice question what would happen if agency is altogether abolished such that no perceptual inference is conditioned on agency. One intriguing possibility is that all sense of self – the entire self-model – would cease.

The illusion presumably arises because in the rubber hand set-up, vision and touch are split apart such that an otherwise optimally functioning Bayesian processing mechanism spits out the wrong output. It is crucial here that other sensory modalities are rendered uninformative because further, independent sources of evidence (such as new proprioceptive input) can allow the Bayesian mechanism to correct the inference. This is methodologically challenging but involves in part the prohibition on taking off the VR goggles or moving the real arm.

From this perspective we can view the unusual experience in the rubber hand illusion and full body illusions as (momentarily) subjectively inescapable perceptual inference, augmented with explaining away effects. At this rather general level of description the illusion is best understood as arising when sensory input is somehow 'wrong' and therefore leads normal Bayesian reality-testing astray. Further, once led astray such reality testing can lead to very odd results as the unusual experience feeds into additional perceptual inference. The illusion is quite easy to break, by allowing further reality-testing. Imagine, however, what things would be like if there were no appropriate further avenues of reality-testing. Then the illusion could be more permanently inescapable. In this way, the Bayesian approach to bodily self-awareness, helped on the way by some of the specific aspects of PEM, begins to look like a useful way to think about delusions. This is because one account of delusion formation is that they begin with inescapable, unusual experiences.

7. Delusion formation

Published: In *ASCS09: Proceedings of the 9th Conference of the Australasian Society for Cognitive Science*. Edited by Wayne Christensen, Elizabeth Schier, and John Sutton. Sydney: Macquarie Centre for Cognitive Science. [This version has some updated references]

Delusions have become a dominant theme in interdisciplinary philosophy of psychiatry (Bortolotti 2009). A delusion is a false belief that the patient holds on to tenaciously, and in the face of counterevidence from carers and family. Models of delusion formation assume delusions arise when normal belief formation breaks down in specific ways and can in that way provide insights into the normal case too. Much of the discussion focuses on how unusual experiences could play a role in the formation of some types of delusions including what is sometime known as 'monothematic' delusions. Patients with these delusions do not need to have other pathological beliefs, and it is often the case that these delusions are relatively circumscribed without much impact on other domains of belief and behaviour. They include for example Capgras delusion where subjects believe loved ones are impostors (they might claim "my wife is an impostor"). This delusion is not uncommon and is often seen after stroke and in dementia. Another delusion in this class is delusions of alien control, which is mostly seen in schizophrenia (where patients might say "the force moved my lips. I began to speak. The words were made for me", "I felt like an automaton, guided by a female spirit who had entered me [when I moved my arm]" (Spence 1997; Frith, Blakemore et al. 2000). It belongs only uneasily to the monothematic delusions as they mostly occur in a larger symptomatic context). I will focus on these two types of delusions here (for a longer list, see (Davies and Coltheart 2000).

Delusions of alien control can be seen as the upshot of faulty prediction error minimisation (Hohwy and Rosenberg 2005; Fletcher and Frith 2009). The idea is that the motor system compares actual and predicted re-afferent sensory input as part of motor planning and execution of motor commands. Roughly speaking, my brain compares the predicted and real sensory consequences of my arm movement to make sure that the movement is the right one for the purpose and that it can be corrected as the movement is executed. The computational problem is analogous to the problem of perception: given a goal state there is an indefinite number of ways in which all the different muscles in one's body could be manipulated to achieve that goal. Rather than trying to figure out a direct solution to this problem, the brain assumes a certain series of motor commands will work and predicts the consequences of that series: if they are predicted to lead to the goal state, and if they actually fit the sensory input then it is a useful movement to perform. Crucially for delusion formation, good predictions of sensory re-afferents lead to attenuation of the actually incoming sensory input, which is just a form of prediction error minimisation. The prediction error itself can be used as a signal: if there is little, then I was probably the one who initiated this movement – since I could predict it with great precision (down to 100s of milliseconds (Blakemore, Wolpert et al. 1998)). On the other hand, if there is much then I was probably not the one who initiated the movement – when I cannot predict how my body is moving someone else must have moved it.

This is the key thought in the PEM account of delusions of alien control: when comparison of predicted and actual reafference goes wrong, the system may receive a signal that is the same as when someone pushes the patient even though she was in fact the one who initiated the movement (Hohwy and Rosenberg 2005; Fletcher and Frith 2009). This constitutes an unusual experience which higher level models concerning agents in the environment must try to account for. The idea is that the patient knows she had the intention to move, that she acted on the intention, probably she also knows that no-one physically pushed her around, and she knows how her

Published: In *ASCS09: Proceedings of the 9th Conference of the Australasian Society for Cognitive Science*. Edited by Wayne Christensen, Elizabeth Schier, and John Sutton. Sydney: Macquarie Centre for Cognitive Science. [This version has some updated references]
body actually moved. The best explanation of the unusual experience, under these circumstances, is that some supernatural force, like a demon, initiated the movement. And this is then adopted as belief.

There are at least three things that an account of delusion formation, such as the one just sketched, needs to explain: why is the wrong, supernatural higher-level model of the world prioritised, why is it elevated to belief, and why is the belief impervious to the counterevidence offered by carers and family?

We have earlier proposed a one-factor account of delusion formation on which delusions arise as a rational response to the unusual experience (Hohwy and Rosenberg 2005). This account is similar to that proposed by Maher (Maher 1974) but works out the detail for particular delusions and incorporates the PEM account as well as various neuroscience and neurological evidence. The main tasks for a one-factor account is to answer the three questions and to show how it is plausible that subjects endowed with normal rationality will also develop such delusions were they to have these unusual experiences. The account must also demonstrate in some way that subjects who report having the unusual experiences without having the delusions are not in fact counterexamples to the account. The main alternative to one-factor theories is two-factor theories, which posit a further role for biases or malfunctions of rationality. These accounts must explain why patients do not seem to develop delusions in response to all kinds of unusual experiences (Coltheart 2007; Aimola Davies and Davies 2009).

Here, I briefly appeal to aspects of PEM, and a few related sources, to provide a fuller understanding of delusion formation. This approach turns out to sit well with a one-factor account.

The initial situation I will consider goes like this. The unusual experience in the case of delusions of control stems from an unexpected prediction error. That is, the sensory consequences of own action is, due to a comparator fault, not attenuated and so is propagated upwards in the cortical system as a prediction error. A model must now be found that can account for this prediction error. Consider two models: the *Spirit* model that the movement was initiated by an invisible spirit, and the *Brain* model that the movement was the patient's own and the sense of other-initiation is caused by brain illness. The delusion arises because the Spirit model wins. The task is to explain why it wins.

Prioritising

The first problem is why the Spirit model is even prioritised as a credible candidate model. It is plausible that the content of the unusual experience triggers this kind of model, even in patients that would not normally subscribe to supernatural hypotheses. In causal reasoning, inference is often guided by what is known as property transmission: in trying to figure out the causes of a certain effect we tend to look for similarities in the properties of the effect and the cause. For example, if the effect is an indentation in some clay with a certain shape, then we expect the causal object to have the same shape (White 2009). There are of course many situations where this principle does not hold (e.g., the cause of death was the faulty breaks), so property transmission can easily lead causal inference astray. Nevertheless it seems to be a heuristic, which operates especially under conditions of uncertainty. If we are

Published: In *ASCS09: Proceedings of the 9th Conference of the Australasian Society for Cognitive Science*. Edited by Wayne Christensen, Elizabeth Schier, and John Sutton. Sydney: Macquarie Centre for Cognitive Science. [This version has some updated references]
relatively uncertain about the causes of an event then it is not unreasonable to at least begin causal inference with considering property transmission. In the case of the unusual experience that triggers delusions of control there is a high degree of uncertainty because it feels like someone else initiated the movement and yet there is no obvious candidate initiator. Movement initiation is an agent-based property so property transmission predicts that the patient will prioritise agent-based causes. Given that no-one visible is there the Spirit model seems appropriate. This appeal to property transmission is not specific to PEM but it does sit well with a Bayesian approach where prior probabilities are extracted from prior learned associations.

If this is right, then a prediction is that healthy individuals should also prioritise supernatural models when they have unusual experiences involving agency. We find this too in a version of the rubber hand illusion where subjects feel a touch but see a finger waving about some distance from the rubber hand. Participants freely explain the experience in terms of black magic, ESP, force fields or invisible extensions of the finger (Hohwy & Paton 2010).

One question now concerns the Brain model. It is possible that it too is prioritised, as would be natural at least for patients seen in clinic where the context very much is that something is wrong with one's brain. It is plausible therefore to expect competition between the Spirit and Brain models.

Believing

Assuming then that the Brain and Spirit models are in competition as accounts of the predictions error, why does the Spirit model win?

First I appeal to the temporal characteristics of the cortical hierarchy, mentioned as an aspect of PEM above (Kiebel, Daunizeau et al. 2008). The prediction error concerns causal interactions at relatively fast time scales (100s of milliseconds to seconds as appropriate for movement related predictions). The Brain model may be at a disadvantage in predicting brain dynamics at that time scale because that model does not seem to have the requisite fineness of temporal grain. It is very difficult to see how the rather general hypothesis that one's brain is sick can generate predictions about prediction error at such fast time scales. In contrast, the Spirit model is agent based and thus taps into stored knowledge of how other agents can interfere with movement. This is a more likely candidate for minimising the prediction error.

If we assume that there are only these two candidate models, and that the winner is whichever model best minimises prediction error, then it is plausible that the Spirit model could win and thus that the patient could believe that the movement was initiated by a demon.

I think there could be a deep-seated bias in favour of the Spirit model, which further boosts its chances of winning. This relates to the importance of agency-driven predictions in PEM. The Brain model says that the agent responsible for the movement was a spirit, so not the patient herself. The Brain model says that the agent responsible was the patient herself but it doesn't offer any clue to how the agent caused the movement. This difference suggests that the current state of the patient's movement control is under a cloud of uncertainty. Perhaps this makes the Brain model less attractive since it would render future agency-based predictions uncertain too.

Published: In *ASCS09: Proceedings of the 9th Conference of the Australasian Society for Cognitive Science*. Edited by Wayne Christensen, Elizabeth Schier, and John Sutton. Sydney: Macquarie Centre for Cognitive Science. [This version has some updated references]
This could thus be a factor that biases in favour of the Spirit model. The motto here could be that it is better to relinquish control than to have control uncertainty.

Imperviousness
Even if the Spirit model wins as the best prediction error minimiser, the question is why it is not dismissed when family and carers, as well as perhaps the patient's own background knowledge, begins to challenge the Spirit model by pointing out, for example, that there is no independent evidence of movement interfering Spirits. Why is the Spirit model impervious to counterevidence?

In terms of PEM, the question is why other models, which contradict the Spirit model, are not able to explain it away. The first thing to notice is that as long as the Spirit model is best able to minimise the prediction error it will not be explained away. Above I gave some reasons why that model is in fact better than the Brain model, which doesn't work well as a control parameter for fast time scale processing, it doesn't offer itself easily for property transmission, and it doesn't make it easy to construct a unified current state for agency-based inference. It is also possible that the Spirit model could be amended to deal with some types of counterevidence. If carers insist they cannot see any Spirits, the model can be revised to add that the Spirits are sinister and are hiding themselves from other people.

Even so, it is likely that the counterevidence would accumulate and yet it doesn't conquer the Spirit model. The question is therefore what it would take for a true high-level model like the Brain model, to be control parameter for fast time-scale dynamics at lower levels? Put like this it is really a question about what it would take for the more domain general model, according to which there are no spirits, to cognitively penetrate down into more domain specific areas and modulate sensory processing there. This is of course a core concern in cognitive science (Fodor 1983). I will not here try to tackle this massive issue head on. I will just indicate how the story about cognitive penetration might go on a PEM account and note that this supports imperviousness for delusional belief.

So, with PEM in mind, the issue of cognitive penetration is how prediction error minimisation can occur from high-levels to low levels. We have already noted that high-level models cannot be expected to predict low-level dynamics so this in itself seems to prevent much cognitive penetrability, the high-level model simply cannot "catch" the low-level bottom-up signal. I think the best hope for a notion of cognitive penetrability is one on which prediction error minimisation occurs by exploiting prediction error noise or ambiguities in the low-level signal. If the low-level signal is very noisy then it is not so important to exactly match the temporal dynamics – prediction error need only be predicted to expected levels of noise. In such a situation it would perhaps be easier for a high-level model to modulate low-level activity (i.e., to minimise low-level prediction error).

We found some evidence for this in our rubber hand study, mentioned earlier. If a stationary toy spider is placed on the visible rubber hand and movement is felt on the real hand, participants do not modulate the visual input in a top-down fashion to actually see the spider moving. But participants very often exploit noise and ambiguity to modulate their experience. They say, for example, that the spider's partially obscured far legs are moving, or that it breathes or have moving baby spiders

Published: In *ASCS09: Proceedings of the 9th Conference of the Australasian Society for Cognitive Science*. Edited by Wayne Christensen, Elizabeth Schier, and John Sutton. Sydney: Macquarie Centre for Cognitive Science. [This version has some updated references]
on its belly (unpublished data). In this case, though, the 'high-level" model is not very high: there is temporally fine-grained tactile information about how the spider should be moving. This kind of detail is still missing from most types of counterevidence to the Spirit model and it is therefore not easy to see how the Spirit model could cognitively penetrate even by exploiting noise and ambiguity.

This approach does however suggest that the chance of cognitive penetrability and thereby susceptibility to counterevidence would, everything else being equal, be best when uncertainty is high. This predicts that the Brain model would have its best chance at being the winning model very early on, while the Spirit model have not been able to suppress prediction error and comprehensively explain away its competitors. There is indeed some evidence that early intervention in the form of both medication and at first psychosis onset is more efficient than intervention later on, as the illness and the delusions have solidified and progressively explained away true competing models.

It remains to address the issue of cases of successful cognitive penetration, that is, reports of unusual experiences but without the subject developing the delusion (Davies, Coltheart et al. 2001). This may be subjects who report a degree of alienation or dissociation from their movements instead of the delusion of alien control, or report a certain strangeness in their experience of loved ones instead of the delusion that the loved one is an impostor. These are the kinds of cases that motivate a two-factor account on which a deficit to the second factor, domain general rationality, is needed to generate the delusions.

I think there is hope that these cases can be explained within the PEM framework, and that they can thus be amenable to a one-factor treatment. Specifically I think there are three ways in which there can be something like the unusual experience without the delusion. First, it is possible that though the experience is unusual its character is somewhat different than the one that leads to the delusion. It may be for example that the prediction error is somewhat less than in the cases that generate the delusion. There may then still be a feeling of strangeness but the prediction error may be so close to expected levels of noise that a higher level model is not needed to account for it. Second, there may be individual differences in levels of expected noise such that the same level of prediction error, generated by faulty comparisons of predicted and actual input, is processed differently in different individuals (Fletcher and Frith 2009). People with high levels of expected noise will not need to engage in prediction error minimisation to the same degree as those with low levels of expected noise. Third, there may be individual differences in how the property transmission heuristic is engaged. It may be that some people are less inclined to adopt this heuristic, and quick to abandon it. Such people may be less susceptible to the delusion even though they can have the unusual experience. This third suggestion is getting close to a two-factor account inasmuch as it focuses on individual biases in reasoning patterns.

There is therefore some reason to think that a one-factor account of delusion formation, based on PEM, can work. That is to say, a second factor does not seem necessary for the account even though of course there may be cases where a second factor is involved, such as a bias or perhaps high-level anatomical damage.

8. Concluding remarks

Published: In *ASCS09: Proceedings of the 9th Conference of the Australasian Society for Cognitive Science*. Edited by Wayne Christensen, Elizabeth Schier, and John Sutton. Sydney: Macquarie Centre for Cognitive Science. [This version has some updated references]

In this paper I have first given a brief presentation of an approach to brain functioning that is gaining more influence. This is the idea that the brain is basically involved in prediction error minimisation. It does this by using generative models to predict what the next sensory input will be and then comparing these predictions with the actual input. The difference between predicted and actual input is the prediction error, which is a quantity that can be used to gauge how good the model is. If the prediction error is large, then the model parameters are updated such that better predictions can be generated. I noted how this approach comes with a notion of explaining away of competing models, how it comes with a notion of temporally characterised cortical hierarchy, and how it incorporates agency as a central component in perceptual inference.

I then applied this framework to three philosophically interesting domains: vision via binocular rivalry, self via the rubber hand illusion, and belief via delusion formation. In each case the prediction error minimisation framework seems able to provide new insights. Binocular rivalry can be understood as the natural upshot of a cognitive system based on prediction error minimisation, rather than merely a matter of reciprocal inhibition and fatigue. Bodily self-awareness can be understood in terms of Bayesian multisensory integration with a basis in agency, such that the role of the body-image for the sense of self is surprisingly fragile. The prediction error minimisation scheme can be used to strengthen a one-factor account of delusion formation; this happens by suggesting a way to understand the notion of cognitive penetrability in terms of the dynamics of prediction error minimisation between cortical levels and across different processing time scales.

I think this discussion has helped show the attractiveness of the prediction error minimisation idea and also that it outlines interesting lines of research for those three areas of philosophical interest.

References

Aimola Davies, A. M. and M. Davies (2009). Explaining pathologies of belief. *Psychiatry as Cognitive Neuroscience: Philosophical Perspectives*. M. R. M.R. Broome and L. Bortolotti. Oxford, Oxford University Press**:** 285–326.

Alais, D. and R. Blake, Eds. (2005). *Binocular Rivalry*. Cambridge, Mass., MIT Press.

Armel, K. C. and V. S. Ramachandran (2003). Projecting sensations to external objects: evidence from skin conductance response. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 270(1523): 1499-1506.

Aspell, J. E., B. Lenggenhager and O. Blanke (2009). Keeping in Touch with One's Self: Multisensory Mechanisms of Self-Consciousness. *PLoS ONE* 4(8): e6488.

Blake, R. and N. K. Logothetis (2002). Visual competition. *Nature Rev. Neurosci.* 3: 13-21.

Blakemore, S.-J., D. M. Wolpert and C. D. Frith (1998). Central cancellation of self-produced tickle sensation. *Nature Neuroscience* 1(7): 635-640.

Bortolotti, L. (2009). *Delusions and Other Irrational Beliefs*. Oxford, Oxford University Press.

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

Published: In *ASCS09: Proceedings of the 9th Conference of the Australasian Society for Cognitive Science*. Edited by Wayne Christensen, Elizabeth Schier, and John Sutton. Sydney: Macquarie Centre for Cognitive Science. [This version has some updated references]
Coltheart, M. (2007). Cognitive neuropsychiatry and delusional belief. *Q. J. Exp. Psychol. (Colchester)* 60: 1041-1062.

Davies, M. and M. Coltheart (2000). Introduction: pathologies of belief. *Mind and Language* 15(1): 1-46.

Davies, M., M. Coltheart, R. Langdon and N. Breen (2001). Monothematic delusions: Toward a two-factor account. *Philosophy, Psychiatry, Psychology* 8(2-3): 133-158.

Dayan, P. (1998). A Hierarchical Model of Binocular Rivalry. *Neural Computation* 10(5): 1119 - 1135.

Ehrsson, H. H. (2007). The Experimental Induction of Out-of-Body Experiences. *Science* 317(5841): 1048-.

Fletcher, P. C. and C. D. Frith (2009). Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. *Nat Rev Neurosci* 10(1): 48-58.

Fodor, J. A. (1983). The Modularity of Mind.

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

Friston, K. (2005). A theory of cortical responses. *Phil. Trans. R. Soc. B* 360: 815-836.

Friston, K. (2008). Hierarchical Models in the Brain. *PLoS Computational Biology* 4(11): e1000211.

Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences* 13(7): 293-301.

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

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

Frith, C., R. Perry and E. Lumer (1999). The neural correlates of conscious experience: an experimental framework. *Trends in Cognitive Sciences* 3(3): 105.

Gigante, G., M. Mattia, J. Braun and P. Del Giudice (2009). Bistable Perception Modeled as Competing Stochastic Integrations at Two Levels. *PLoS Comput Biol* 5(7): e1000430.

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

Gregory, R. L. (1998). *Eye and Brain*. Oxford, Oxford University Press.

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

Hohwy, J. (2007). The search for neural correlates of consciousness. *Philosophy Compass* 2(3): 461–474.

Hohwy, J. and B. Paton (2010). Explaining away the body: experiences of supernaturally caused touch and touch on non-hand objects within the rubber hand illusion *PLoS ONE* 5(2): e9416.

Hohwy, J., A. Roepstorff and K. Friston (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition* 108(3): 687-701.

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

Jensen, F. V. and S. L. Lauritzen (2000). Probabilistic networks. *Handbook of Defeasible Reasoning and Uncertainty Management Systems: Algorithms for*

Published: In *ASCS09: Proceedings of the 9th Conference of the Australasian Society for Cognitive Science*. Edited by Wayne Christensen, Elizabeth Schier, and John Sutton. Sydney: Macquarie Centre for Cognitive Science. [This version has some updated references]
uncertainty and defeasible reasoning/volume J. Kohlas and M. S. Dordrecht, Kluwer**:** 289-320.

Kersten, D., P. Mamassian and A. Yuille (2004). Object perception as Bayesian inference. *Annu. Rev. Psychol.* 55: 271.

Kiebel, S. J., J. Daunizeau and K. J. Friston (2008). A Hierarchy of Time-Scales and the Brain. *PLoS Computational Biology* 4(11): e1000209.

Körding, K. P., U. Beierholm, W. J. Ma, S. Quartz, J. B. Tenenbaum and L. Shams (2007). Causal inference in multisensory perception. *Plos One* 2(9).

Lenggenhager, B., M. Mouthon and O. Blanke (2008). Spatial aspects of bodily self-consciousness. *Consciousness and Cognition* In Press, Corrected Proof.

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

Maher, B. A. (1974). Delusional thinking and perceptual disorder. *Journal of Individual Psychology* 30: 98-113.

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

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

Noest, A. A. J., R. R. van Ee, M. M. M. Nijs and R. R. J. A. van Wezel (2007). Percept-choice sequences driven by interrupted ambiguous stimuli: a low-level neural model. *Journal of vision* 7(8): 10.

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

Schwabe, L. and O. Blanke (2008). The vestibular component in out-of-body experiences: a computational approach. *Frontiers in Human Neuroscience* 2 (Article 17): 1-10.

Slater, M., D. P. Marcos, H. Ehrsson and M. V. Sanchez-Vives (2009). Inducing illusory ownership of a virtual body. *Frontiers in Neuroscience*.

Song, C. and H. Yao (2009). Duality in Binocular Rivalry: Distinct Sensitivity of Percept Sequence and Percept Duration to Imbalance between Monocular Stimuli. *PLoS ONE* 4(9): e6912.

Spence, S. A. (1997). A PET study of voluntary movement in schizophrenic patients experiencing passivity phenomena (delusions of alien control). *Brain* 120: 1997-2011.

Tsakiris, M. and P. Haggard (2005). The rubber hand illusion revisited: visuotactile integration and self-attribution. *Journal of experimental psychology. Human perception and performance* 31(1): 80.

Wellman, M. P. and M. Henrion (1993). Explaining 'explaining away'. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15(3): 287-292.

White, P. A. (2009). Property Transmission: An Explanatory Account of the Role of Similarity Information in Causal Inference. *Psychological Bulletin* 135(5): 774-793.

Wilson, H. R. (2007). Minimal physiological conditions for binocular rivalry and rivalry memory. *Vision Research* 47(21): 2741-2750.