Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science.

Andy Clark

Abstract

Brains, according to an increasingly influential cluster of research programs in cognitive computational neuroscience, are essentially prediction machines. They are bundles of cells whose evolved role is to enable the right action to occur at the right time, and one key way this is achieved by using multi-layer generative models to minimize prediction errors within a hierarchy of cortical processing regions. This 'hierarchical prediction machine' approach suggests a unifying model of perception and action, illuminates the functional role of attention, and may neatly capture the special contribution of cortical processing to adaptive success. In this paper, I first lay out (sections 1 and 2) the key elements and implications of this approach. I then explore (section 3) a variety of pitfalls and challenges, spanning the evidential, the methodological, and the more properly conceptual. I end (sections 4 and 5) by asking how such approaches might impact our more general vision of mind, experience, and agency.

1. Prediction Machines

1.1 From Helmholtz to Action-Oriented Predictive Processing

"The whole function of the brain is summed up in: error correction"

So wrote W. Ross Ashby, the British psychiatrist and cyberneticist, some half a century ago[1]. Computational neuroscience has come a very long way since then. There is now increasing reason to believe that Ross Ashby's (admittedly somewhat vague) statement is correct, and that it captures something crucial about the way that spending metabolic money to build complex brains pays dividends in the search for adaptive success. In particular, one of the brain's key tricks, it now seems, is to implement dumb processes that correct a certain kind of error: error in the multi-layered prediction of input. In mammalian brains, such errors look to be corrected within a cascade of cortical processing events in which higher-level systems attempt to predict the inputs to lower level ones on the basis of their own emerging models of the causal structure of the world (i.e. the signal source). Errors in predicting lower level inputs cause the higher-level models to adapt so as to reduce the discrepancy. Operating

over a plethora of linked higher-level models, the upshot is a brain that encodes a rich body of information about the source of the signals that regularly perturb it. Such models follow Helmholtz (1860) in depicting perception as a process of probabilistic, knowledge-driven inference. From Helmholz comes the key idea that sensory systems are in the tricky business of inferring sensory causes from their bodily effects. This in turn involves computing multiple probability distributions since a single such effect will be consistent with many different sets of causes distinguished only by their relative (and context dependent) probability of occurrence.

Helmholz insight informed influential work by MacKay (1956), Neisser (1967), and Gregory (1980), as part of the cognitive psychological tradition that became known as 'analysis-by-synthesis' (for a review, see Yuille and Kersten (2006)). In this paradigm the brain does not build its current model of distal causes (its model of how the world is) simply by accumulating, from the bottom-up, a mass of low-level cues such as edge-maps etc. Instead (see Hohwy (2007)) the brain tries to predict the current suite of cues from its best models of the possible causes. In this way:

> "The mapping from low- to high-level representation (e.g. from acoustic to word-level) is computed using the *reverse* mapping, from high- to low-level representation" Chater and Manning (2006, p.340, their emphasis).

Helmholz' insight was also pursued in an important body of computational and neuroscientific work. Crucial to this lineage were seminal advances in machine learning that began with pioneering connectionist work on back-propagation learning (McClelland et al (1986), Rumelhart et al (1986)) and continued with work on the aptly named 'Helmholz machine' (Dayan et al (1995), Dayan and Hinton (1996) – see also Hinton and Zemel (1994))[2]. Importantly for our purposes, the Helmholtz Machine sought to learn new representations in a multi-layer system (thus capturing increasingly deep regularities within a domain) without requiring the provision of copious pre-classified samples of the desired input-output mapping. In this respect it aimed to improve (see Hinton (2010)) upon standard back-propagation driven learning. It did this by using its own top-down connections to provide the desired states for the hidden units thus (in effect) self-supervising the development of its perceptual 'recognition model' using a *generative* model that tried to create the sensory patterns for itself (in 'fantasy' as it was sometimes said)[3]. (For a useful review of this crucial innovation and a survey of many subsequent developments, see Hinton (2007a)).

A generative model, in this quite specific sense, aims to capture the statistical structure of some set of observed inputs by tracking (one might say, by schematically recapitulating) the causal matrix responsible for that very structure. A good generative model for vision would thus seek to capture the ways in which observed lower level visual responses are generated by an interacting web of causes - for example, the various aspects of a visually presented scene. In practice, this means that top-down connections within a multi-layer (hierarchical) system come to encode a probabilistic model of the activities of units and groups of units within lower level layers, thus tracking (as we shall shortly see in more detail) interacting causes in the signal source, which might be the body, or the external world – see e.g. Kawato et al (1993), Hinton and Zemel (1994), Mumford (1994), Hinton et al (1995), Dayan et al (1995), Olshausen & Field, (1996), Dayan (1997), Hinton and Ghahramani (1997).

It is this twist – the strategy of using top-down connections to try to generate, using high-level knowledge, a kind of 'virtual version' of the sensory data via a deep multilayer cascade – that lies at the heart of 'hierarchical predictive coding' approaches to perception (Rao and Ballard (1999), Lee and Mumford (2003), Friston (2005)). Such approaches, along with their recent extensions to action (Friston and Stephan (2007), Friston, Daunizeau et al (2009), Friston (2010), Brown et al (2011)) form the main focus of the present treatment. These approaches combine the use of top-down probabilistic generative models with a specific vision of one way in which such downward influence might operate. That way (borrowing from work in linear predictive coding – see below) depicts the top-down flow as attempting to predict and fully 'explain away' the driving sensory signal, leaving only any residual 'prediction errors' to propagate information forward within the system (Rao and Ballard (1999), Lee and Mumford (2003), Friston (2005), Hohwy et al (2008), Jehee and Ballard (2009), Friston (2010), Brown et al (2011) -– for a recent review, see Huang and Rao (2011)).

Predictive coding itself was first developed as a data compression strategy in signal processing (for a history, see Shi and Sun (1999)). Thus consider a basic task such as image transmission. In most images, the value of one pixel regularly predicts the value of its nearest neighbors, with differences marking important features such as the boundaries between objects. That means that the code for a rich image can be compressed (for a properly informed receiver) by encoding only the 'unexpected' variation: the cases where the actual value departs from the predicted one. What needs to be transmitted is thus just the difference (aka the 'prediction error') between the actual current signal and the

predicted one. This affords major savings on bandwith, an economy that was the driving force behind the development of the techniques by James Flanagan and others at Bell Labs during the 1950's (for a review, see Musmann (1979)). Descendents of this kind of compression technique are currently used in JPEGS, in various forms of lossless audio compression, and in motion-compressed coding for video. The information that needs to be communicated 'upwards' under all these regimes is just the prediction error: the divergence from the expected signal. Transposed (in ways we are about to explore) to the neural domain, this makes prediction error into a kind of proxy (Feldman and Friston (2010)) for sensory information itself.

Hierarchical predictive processing combines the use, within a multi-layer cascade, of 'top-down' probabilistic generative models with the core predictive coding strategy of efficient encoding and transmission. Such approaches, originally developed in the domain of perception, have been extended (by Friston and others – see 1.5 following) to encompass action, and to offer an attractive, unifying perspective on the brain's capacities for learning, inference, and the control of plasticity. Perception and action, if these unifying models are correct, are intimately related and work together to reduce prediction error by sculpting and selecting sensory inputs. In the remainder of this section I rehearse some of the main features of these models before highlighting (sections 2-5 following) some of their most conceptually important and challenging aspects.

1.2 Escaping the Black Box

A good place to start (following Rieke (1999)) is with what might be thought of as the "view from inside the black box". For the task of the brain, when viewed from a certain distance, can seem impossible: it must discover information about the likely causes of impinging signals without any form of direct access to their source. Thus consider a black box taking inputs from a complex external world. The box has input and output channels along which signals flow. But all that it 'knows' about, in any direct sense, are the ways its own states (e.g. spike trains) flow and alter. In that (restricted) sense, all the system has direct access to are its own states. The world itself is thus off-limits (though the box can, importantly, issue motor commands and await developments). The brain is one such black box. How, simply on the basis of patterns of changes in its own internal states, is it to alter and adapt its responses so as to tune itself to act as a useful node (one that merits its relatively huge metabolic expense) for the origination of adaptive responses? Notice how different this

conception is to ones in which the problem is posed as one of establishing a mapping relation between environmental and inner states. The task is not to find such a mapping but to infer the nature of the signal source (the world) from just the varying input signal itself.

Hierarchical approaches in which top-down generative models are trying to predict the flow of sensory data provide a powerful means for making progress under such apparently unpromising conditions. One key task performed by the brain, according to these models, is that of guessing the next states of its own neural economy. Such guessing improves when you use a good model of the signal source. Cast in the Bayesian mode, good guesses thus increase the posterior probability[4] of your model. Various forms of gradient descent learning can progressively improve your first guesses. Applied within a hierarchical predictive processing[5] regime, this will – if you survive long enough – tend to yield useful generative models of the signal source (ultimately, the world).

The beauty of the hierarchical structure is that it allows the system to infer its own priors (the prior beliefs essential to the guessing routines) as it goes along. It does this by using its best current model - at one level - as the source of the priors for the level below, engaging in a process of 'iterative estimation' (see Dempster, Laird, and Rubin (1977), Neal and Hinton (1998)) that allows priors and models to co-evolve across multiple linked layers of processing so as to account for the sensory data. This kind of procedure (which implements a version of 'empirical Bayes' Robbins (1956)) has an appealing mapping to known facts about the hierarchical and reciprocally connected structure and wiring of cortex (Lee and Mumford (2003), Friston (2005))[6].

A classic early example, combining this kind of hierarchical learning with the basic predictive coding strategy described in 1.1. above, is Rao and Ballard's (1999) model of predictive coding in the visual cortex. At the lowest level there is some pattern of energetic stimulation, transduced (let's suppose) by sensory receptors from ambient light patterns produced by the current visual scene. These signals are then processed via a multi-level cascade in which each level attempts to predict the activity at the level below it via backward[7] connections. The backward connections allow the activity at one stage of the processing to return as another input at the previous stage. So long as this successfully predicts the lower level activity all is well, and no further action needs to ensue. But where there is a mismatch 'prediction error' occurs and the ensuing (error-indicating) activity is propagated to the higher level, which automatically alters its own response tendencies so as to reduce that discrepancy next time around.

Forward connections between levels thus carry the 'residual errors' (Rao and Ballard (1999) p.79) separating the predictions from the actual lower level activity, while backward connections (which do most of the 'heavy lifting' in these models) carry the predictions themselves. Changing predictions corresponds to changing or tuning your hypothesis about the hidden causes of the lower level activity. The concurrent running of this kind of prediction error calculation within a loose hierarchy of cortical areas allows information pertaining to many different grain sizes (that is, different spatial and temporal scales) within the image to be played off against one another, with each such 'hypothesis' being used to tune the rest. As the authors put it:

> "prediction and error-correction cycles occur concurrently throughout the hierarchy, so top-down information influences lower-level estimates, and bottom-up information influences higher-level estimates of the input signal" Rao and Ballard (1999) p.80

In visual cortex, such a scheme suggests that backward connections from V2 to V1 would carry a prediction of expected activity in V1, while forward connections from V1 to V2 would carry forward the error signal indicating residual (unpredicted) activity.

To test these ideas, Rao and Ballard implemented a simple hierarchical network of such 'predictive estimators' and trained it on image patches derived from five natural scenes. Using learning algorithms that progressively reduce prediction error across the linked cascade and after exposure to thousands of image patches, the system learnt to use responses in the first level network to extract features such as oriented edges and bars, while the second level network came to capture combinations of such features corresponding to patterns involving larger spatial configurations. The model also displayed (see section 3.1 following) a number of interesting 'extra-classical receptive field' effects suggesting that such non-classical surround effects (and, as we'll later see, context effects more generally) may be a rather direct consequence of the use of hierarchical predictive coding. For immediate purposes, however, what matters is that the predictive coding approach, given only the statistical properties of the signals derived from the natural images, was able to induce a kind of generative model of the structure of the input data: it learnt about the presence and importance of features such as lines, edges, and bars, and about combinations of such features, in ways that enable better predictions concerning what to expect next, in space or in time. The cascade of processing induced by the progressive reduction of prediction error in the hierarchy reveals the world outside the black box. It maximizes the posterior probability

of generating the observed states (the sensory inputs), and in so doing induces a kind of internal model of the source of the signals: the world hidden behind the veil of perception.

1.3 Dynamic Predictive Coding by the Retina

As an example of the power (and potential ubiquity) of the basic predictive coding strategy itself, and one that now moves context centre stage, consider Hosoya et al's (2005) account of dynamic predictive coding by the retina. The starting point of this account is the well-established sense in which retinal ganglion cells take part in some form of predictive coding, insofar as their receptive fields display centre-surround spatial antagonism, as well as a kind of temporal antagonism. What this means, in each case, is that neural circuits predict, on the basis of local image characteristics, the likely image characteristics of nearby spots in space and time (basically, assuming that nearby spots will display similar image intensities) and subtract this predicted value from the actual value. What gets encoded is thus not the raw value but the differences between raw values and predicted values. In this way "Ganglion cells signal not the raw visual image but the departures from the predictable structure, under the assumption of spatial and temporal uniformity." (Hosoya et al (2005) p.71). This saves on bandwidth, and also flags what is (to use Hosoya et al's own phrase) most 'newsworthy' in the incoming signal[8].

These computations of predicted salience might be made solely on the basis of average image statistics. To do so, however, would be a hostage to fortune in many ecologically significant environments. For example, consider an animal that moves between a watery environment and dry land, or between a desert landscape and a verdant oasis. The spatial scales at which nearby points in space and time are typically similar in image intensity vary markedly between such cases, because the statistical properties of the different types of scene vary. Hosoya et al thus predicted that, in the interests of efficient, adaptively potent, encoding, the behaviour of the retinal ganglion cells (specifically, their receptive field properties) should vary as a result of adaptation to the current scene or context, exhibiting what they term 'dynamic predictive coding'.

Putting salamanders and rabbits into varying environments, and recording from their retinal ganglion cells, Hosoya et al confirmed their hypothesis: within a space of several seconds about 50% of the ganglion cells altered their behaviours to keep step with the changing image statistics of the varying environments. A mechanism was then proposed and tested using a simple feedforward neural network that performs a form of anti-Hebbian learning.

Anti-Hebbian feedforward learning, in which correlated activity across units leads to inhibition rather than to activation (see e.g. Kohonen (1989)) enables the creation of 'novelty filters' that learn to become insensitive to the most highly correlated (hence most 'familiar') features of the input. This, of course, is exactly what is required in order to learn to discount the most statistically predictable elements of the input signal in the way dynamic predictive coding suggests. Better yet, there are neuronally plausible ways to implement such a mechanism using amacrine cell synapses to mediate plastic inhibitory connections that in turn alter the receptive fields of retinal ganglion cells (for details, see Hosoya et al (2005) p.74) so as to suppress the most correlated components of the stimulus. In sum, retinal ganglion cells seem to be engaging in a computationally and neurobiologically explicable process of dynamic predictive recoding of raw image inputs, whose effect is to "strip from the visual stream predictable and therefore less newsworthy signals" (Hosoya et al (2005) p.76).

1.4 Another Illustration: Binocular Rivalry

So far, our examples have been restricted to relatively low-level visual phenomena. As a final illustration, however, consider Hohwy, Roepstorff, and Friston's (2008) hierarchical predictive coding model of binocular rivalry. Binocular rivalry (see e.g. essays in Alais and Blake (2005), and the review article by Leopold and Logothetis (1999)) is a striking form of visual experience that occurs when, using a special experimental set-up; each eye is presented (simultaneously) with a different visual stimulus. Thus, the right eye might be presented with an image of a house, while the left receives an image of a face. Under these (extremely – and importantly - artificial) conditions, subjective experience unfolds in a surprising 'bi-stable' manner. Instead of seeing (visually experiencing) a confusing all-points merger of house and face information, subjects report a kind of perceptual alternation between seeing the house and seeing the face. The transitions themselves are not always sharp, and subjects often report a gradual breaking through (see e.g. Lee at al (2005)) of elements of the other image, before it dominates the previous one, after which the cycle repeats.

Such 'binocular rivalry', as Hohwy et al remind us, has been a powerful tool for studying the neural correlates of conscious visual experience, since the incoming signals remain constant while the percept switches to and fro (Frith, Perry, and Lumer (1999)). Despite this attention, however, the precise mechanisms at play here are not well understood. Hohwy et al's strategy is to take a step back, and to attempt to explain the phenomenon from first

principles in a way that makes sense of many apparently disparate findings. In particular, they pursue what they dub an 'epistemological' approach: one whose goal is to reveal binocular rivalry as a reasonable (knowledge-oriented) response to an ecologically unusual stimulus condition.

The starting point for their story is, once again, the emerging unifying vision of the brain as an organ of prediction using a hierarchy of generative models. Recall that, on these models, the task of the perceiving brain is to account for (to 'explain away') the incoming or 'driving' sensory signal by means of a matching top-down prediction. The better the match, the less prediction error then propagates up the hierarchy. The higher level guesses are thus acting as priors for the lower level processing, in the fashion of so-called 'empirical Bayes' (such methods use their own target data sets to estimate the prior distribution: a kind of bootstrapping that exploits the statistical independencies that characterize hierarchical models).

Within such a multi-level, hierarchical setting a visual percept is determined by a process of prediction operating across many levels of a processing hierarchy, each concerned with different types and scales of perceptual detail. All the communicating areas are locked into a mutually coherent predictive coding regime, and their interactive equilibrium ultimately selects a best overall (multi-scale) hypothesis concerning the state of the visually presented world. This is the hypothesis that "makes the best predictions and that, taking priors into consideration, is consequently assigned the highest posterior probability" (Hohwy, Roepstorff, and Friston (2008) p.690). Other overall hypotheses, at that moment, are simply crowded out: they are effectively inhibited, having lost the competition to best account for the driving signal.

Notice, though, what this means in the context of the predictive coding cascade. Top-down signals will explain away (by predicting) only those elements of the driving signal that conform to (and hence are predicted by) the current winning hypothesis. In the binocular rivalry case, however, the driving (bottom-up) signals contain information that suggests two distinct, and incompatible, states of the visually presented world – e.g. face at location X/house at location X. When one of these is selected as the best overall hypothesis, it will account for all and only those elements of the driving input that that hypothesis predicts. As a result, prediction error for that hypothesis decreases. But prediction error associated with the elements of the driving signal suggestive of the alternative hypothesis is not suppressed, and is now propagated up the hierarchy. To suppress *those* prediction errors the system needs to find another hypothesis. But having done so (and hence, having

flipped the dominant hypothesis to the other interpretation), there will again emerge a large prediction error signal, this time deriving from those elements of the driving signal not accounted for by the flipped interpretation. In Bayesian terms, this is a scenario in which no unique and stable hypothesis combines high prior and high likelihood. No single hypothesis accounts for all the data, so the system alternates between the two semi-stable states. It behaves as a bi-stable system, minimizing prediction error in what Hohwy et al describe as an energy landscape containing a double well.

What makes this account different from its rivals (such as Lee, Blake, and Heeger (2005)) is that where they posit a kind of direct, attention-mediated but essentially feedforward, competition between inputs, these accounts (see also Lee and Mumford (2003)) posit 'top-down' competition between linked sets of hypotheses. The effect of this competition is to selectively suppress the prediction errors associated with the elements of the driving (sensory) signals suggesting the current winning hypothesis. But this top-down suppression leaves untouched the prediction errors associated with the remaining elements of the driving signal. These errors are then propagated up the system. To explain them away the overall interpretation must switch. This pattern repeats, yielding the distinctive alternations experienced during dichoptic viewing of inconsistent stimuli[9].

Why, under such circumstances, do we not simply experience a combined or interwoven image: a kind of house/face mash-up for example? Although such partially combined percepts do apparently occur, for brief periods of time, they are not sufficiently stable, as they do not constitute a viable hypothesis given our more general knowledge about the visual world. For it is part of that general knowledge (which acts as a so-called 'hyperprior' in the system) that, for example, houses and faces are not present in the same place, at the same scale, at the same time. That is to say "the prior probability of both a house and face being co-localized in time and space is extremely small" (op cit p.5). This, indeed, is the deep explanation of the existence of competition between certain higher-level hypotheses in the first place. They compete because the system has learnt that 'only one object can exist in the same place at the same time' (op cit p.8). (This obviously needs careful handling, since a single state of the world may be consistently captured by multiple high level stories that ought not to compete in the same way: e.g. seeing the painting as valuable, as a Rembrandt, as an image of a cow, etc).

1.5 Action-Oriented Predictive Processing

Recent work (Friston, Daunizeau et al (2009), Friston (2010), Brown et al (2011)) generalizes this basic 'hierarchical predictive processing' model to include action. According to what I shall now dub 'action-oriented predictive processing'[10] perception and action both follow the same deep 'logic' and are even implemented using the same computational strategies. A fundamental attraction of these accounts thus lies in their ability to offer a deeply unified account of perception, cognition, and action.

Perception, as we saw, is here depicted as a process that attempts to match incoming 'driving' signals with a cascade of top-down predictions (spanning multiple spatial and temporal scales) that aim to cancel it out. Motor action exhibits a surprisingly similar profile, except that:

> "In motor systems error signals self-suppress, not through neuronally mediated effects, but by eliciting movements that change bottom-up proprioceptive and sensory input. This unifying perspective on perception and action suggests that action is both perceived and caused by its perception" Friston (2003) p. 1349

This whole scenario is wonderfully captured by Hawkins and Blakeslee who write that:

> "As strange as it sounds, when your own behaviour is involved, your predictions not only precede sensation, they determine sensation. Thinking of going to the next pattern in a sequence causes a cascading prediction of what you should experience next. As the cascading prediction unfolds, it generates the motor commands necessary to fulfil the prediction. Thinking, predicting, and doing are all part of the same unfolding of sequences moving down the cortical hierarchy" Hawkins and Blakeslee (2004) p.158

A closely related body of work in so-called 'optimal feedback control theory' (e.g. Todorov and Jordan (2002), Todorov (In Press)) displays the motor control problem as mathematically equivalent to Bayesian inference. Very roughly – see Todorov (In Press) for a detailed account - you treat the desired (goal) state as observed and perform Bayesian inference to find the actions that get you there. This mapping between perception and action emerges also in

some recent work on planning (e.g. Toussaint (In Press)). The idea, closely related to these approaches to simple movement control, is that in planning we imagine a future goal state as actual, then use Bayesian inference to find the set of intermediate states (which can now themselves be whole actions) that get us there. There is thus emerging a fundamentally unified set of computational models which, as Toussaint (in press) comments "do not distinguish between the problems of sensor processing, motor control, or planning". At the very least, it now seems likely that perception and action are in some deep sense computational siblings and that:

> " the best ways of interpreting incoming information via perception, are deeply the same as the best ways of controlling outgoing information via motor action….so the notion that there are a few specifiable computational principles governing neural function seems plausible. " Eliasmith (2007) p. 380

Action-oriented predictive processing goes further, however, in suggesting that motor intentions actively elicit, via their unfolding into detailed motor actions, the ongoing streams of sensory (especially proprioceptive) results that our brains predict. This deep unity between perception and action emerges most clearly in the context of so-called 'active perception', where the agent moves its sensors in ways that amount (Friston (2009), Friston, Daunizeau et al (2011))) to actively searching out the sensory consequences that they (or rather, their brains) expect. Perception, cognition, and action - if this unifying perspective proves correct - work closely together to minimize sensory prediction errors by selectively sampling, and actively sculpting, the stimulus array. They thus conspire to move a creature through time and space in ways that fulfil an ever-changing and deeply inter-animating set of (sub-personal) expectations. According to these accounts, then:

> "Perceptual learning and inference is necessary to induce prior expectations about how the sensorium unfolds. Action is engaged to resample the world to fulfil these expectations. This places perception and action in intimate relation and accounts for both with the same principle" (Friston, Daunizeau, and Kiebel (2009) p. 12)

In some (I'll call them the 'desert landscape') versions of this story (see especially Friston (2011a), Friston, Daunizeau, et al (2011)) proprioceptive prediction errors act directly as motor commands. On these models it is our expectations about the proprioceptive consequences of moving and acting that

directly bring the moving and acting about[11]. I return briefly to these 'desert landscape' scenarios in section 5.1 following.

1.6 The Free Energy Formulation

That large-scale picture (of creatures enslaved to sense and act in ways that make most of their sensory predictions come true) finds fullest expression in the so-called 'free-energy minimization' framework (Friston and Stephan (2007), Friston (2003) (2009) (2010)). Free-energy formulations originate in statistical physics and were introduced into the machine learning literature in treatments that include Neal and Hinton (1998), Hinton and von Camp (1993), Hinton and Zemel (1994), and MacKay (1995). Such formulations can arguably be used (e.g. Friston (2010)) to display the prediction error minimization strategy as itself a consequence of a more fundamental mandate to minimize an information-theoretic isomorph of thermodynamic free-energy in a system's exchanges with the environment.

Thermodynamic free energy is a measure of the energy available to do useful work. Transposed to the cognitive/informational domain, it emerges as the difference between the way the world is represented as being, and the way it actually is. The better the fit, the lower the information-theoretic free energy (this is intuitive, since more of the system's resources are being put to 'effective work' in representing the world). Prediction error reports this information-theoretic free energy, which is mathematically constructed so as always to be greater than surprise. This is because surprise, used in this way (a better label, given this technical usage, is 'surprisal' – see Tribus (1961) and section 4.1 following) is simply the sub-personally computed implausibility of some sensory state given a model of the world. Reducing information-theoretic free energy thus amounts to improving the world model so as to reduce such errors, hence reducing surprisal[12] (since better models make better predictions). The overarching rationale (Friston (2010)) is that good models help us to maintain our structure and organization, hence to appear, (over extended but finite timescales) to resist entropy and the second law of thermodynamics. They do so by rendering us good predictors of sensory unfoldings, hence better poised to avoid damaging exchanges with the environment.

The 'free-energy principle' itself then states that "all the quantities that can change; i.e. that are part of the system, will change to minimize free-energy" (Friston and Stephan (2007) p. 427). Notice that thus formulated this is a claim about all elements of systemic organization (from gross morphology to the entire organization of the brain) and not just about cortical information

processing. Using a series of elegant mathematical formulations, Friston (2009) (2010) suggests that this principle, when applied to various elements of neural functioning, leads to the generation of efficient internal representational schemes and reveals the deeper rationale behind the links between perception, inference, memory, attention, and action scouted in the previous sections. Morphology, action tendencies (including the active structuring of environmental niches), and gross neural architecture are all claimed to fall – though at very different timescales - under this single umbrella.

The free energy account is of great independent interest. It represents a kind of 'maximal version' of the claims scouted in section 1.5 concerning the computational intimacy of perception and action, and it is suggestive of a general framework that might accommodate the growing interest (see e.g. Thompson (2007)) in understanding the relations between life and mind. A full assessment of the free energy principle is, however, far beyond the scope of the present treatment[13]. In the remainder of this treatment I turn instead to a number of issues and implications arising more directly from hierarchical predictive processing accounts of perception and their possible extensions to action.

2. Representation, Inference, and the Continuity of Perception, Cognition, and Action.

The hierarchical predictive processing account, along with the more recent generalizations to action represents, or so I shall now argue, a genuine departure from many of our previous ways of thinking about perception, cognition, and the human cognitive architecture. It offers a distinctive account of neural representation, neural computation, and the representation relation itself. It depicts perception, cognition, and action as profoundly unified and, in important respects, continuous. And it offers a neurally plausible and computationally tractable gloss on the claim that the brain performs some form of Bayesian inference.

2.1 Explaining Away

To successfully represent the world in perception, if these models are correct, depends crucially upon cancelling out sensory prediction error. Perception thus involves 'explaining away' the driving (incoming) sensory signal by matching it with a cascade of predictions pitched at a variety of spatial and temporal scales.

These predictions reflect what the system already knows about the world (including the body) and the uncertainties associated with its own processing. Perception here becomes 'theory-laden' in at least one (rather specific) sense: what we perceive depends heavily upon the set of priors (including any relevant hyperpriors) that the brain brings to bear in its best attempt to predict the current sensory signal. On this model, perception is nothing but the success of some mutually supportive stack of states of generative models (recall 1.1 above) at minimizing prediction error by hypothesizing an interacting set of distal causes that predict, accommodate, and (thus) "explain away" the driving sensory signal.

This appeal to 'explaining away' is important and central, but it needs very careful handling. It is important as it reflects the key property of hierarchical predictive processing models, which is that the brain is in the business of active, ongoing, input prediction and does not (even in the early sensory case) merely react to external stimuli. It is important also insofar as it is the root of the attractive coding efficiencies that these models exhibit, since all that need be passed forward through the system is the error signal, which is what remains once predictions and driving signals have been matched. In these models it is thus the backwards (recurrent) connectivity that carries the main information processing load. It is important, however, not to overplay this difference. In particular, it is potentially misleading to say that:

> "Activation in early sensory areas no longer represents sensory information per se, but only that part of the input that has not been successfully predicted by higher-level areas" (De-Wit, Machilsen, and Putzeys (2010) 8702

For this stresses only one aspect of what is (at least in context of the rather specific models we have been considering[14]) actually depicted as a kind of duplex architecture: one that at each level *combines* quite traditional representations of inputs with representations of error. According to the duplex proposal, what gets 'explained away' or cancelled out is the error signal, which (in these models) is depicted as computed by dedicated 'error units'. These are linked to, but distinct from, the so-called 'representation units' meant to encode the causes of sensory inputs. By cancelling out the activity of the error units, activity in some of the laterally interacting 'representation' units (which then feed predictions downwards and are in the business of encoding the putative sensory causes) can actually end up being selected and sharpened. The hierarchical predictive processing account thus avoids any direct conflict with accounts (e.g. biased competition models such as Desimone and Duncan

(1995)) that posit top-down *enhancements* of selected aspects of the sensory signal since:

> "High-level predictions explain away prediction error and tell the error units to 'shut up' [while] units encoding the causes of sensory input are selected by lateral interactions, with the error units, that mediate empirical priors. This selection stops the gossiping [among laterally competing representations in lower areas]" Friston (2005) p.829

The drive towards 'explaining away' is thus consistent, in this specific architectural setting, with both the sharpening and the dampening of (different aspects of) early cortical response[15]. Thus Spratling, in a recent formal treatment of this issue[16], suggests that any apparent contrast here reflects:

> "a mis-interpretation of the model that may have resulted from the strong emphasis the predictive coding hypothesis places on the *error-detecting nodes* and the corresponding *under-emphasis on the role of the prediction nodes in maintaining an active representation of the stimulus*" Spratling (2008) p. 8 (my emphasis)

What is most distinctive about this duplex architectural proposal (and where much of the break from tradition really occurs) is that it depicts the forward flow of information as solely conveying error, and the backwards flow as solely conveying predictions. The duplex architecture thus achieves a rather delicate balance between the familiar (there is still a cascade of feature-detection, with potential for selective enhancement, and with increasingly complex features represented by neural populations that are more distant from the sensory peripheries) and the novel (the forward flow of sensory information is now entirely replaced by a forward flow of prediction error).

This balancing act between cancelling out and selective enhancement is only made possible, it should be stressed, by positing the existence of "two functionally distinct sub-populations, encoding the conditional expectations of perceptual causes and the prediction error respectively" (Friston (2005), p.829). Functional distinctness need not, of course, imply gross physical separation. But a common conjecture in this literature depicts superficial pyramidal cells (a prime source of forward neuro-anatomical connections) as playing the role of error units, passing prediction error forwards, while deep pyramidal cells play the role of representation units, passing predictions (made on the basis of complex generative models) downwards (see e.g. Mumford (1992), Friston (2005), (2009)). However it may (or may not) be realized, some form of

functional separation is required. Such separation constitutes a central feature of the proposed architecture, and one without which it would be unable to combine the radical elements drawn from predictive coding with simultaneous support for the more traditional structure of increasingly complex feature detection and top-down signal enhancement. But essential as it is, this is a demanding and potentially problematic requirement, which we will return to in section 3.1 following.

2.2 Encoding, Inference, and the 'Bayesian Brain'.

Neural representations, should the hierarchical predictive processing account prove correct, encode probability density distributions in the form of probabilistic generative models, and the flow of inference respects Bayesian principles that balance prior expectations against new sensory evidence. This (Eliasmith (2007)) is a departure from traditional understandings of internal representation, and one whose full implications have yet to be understood. It means that the nervous system is fundamentally adapted to deal with uncertainty, noise, and ambiguity, and it requires some (perhaps several) concrete means of internally representing uncertainty. (Non-exclusive options here include the use of distinct populations of neurons, varieties of 'probabilistic population codes' (Pouget, Dayan, and Zemel (2003)), and relative timing effects (Deneve (2008))   - for a very useful review see Vilares and Kording (2011)). Predictive processing accounts thus share what Knill and Pouget (2004, p.713) describe as the "basic premise on which Bayesian theories of cortical processing will succeed or fail", viz that:

> "The brain represents information probabilistically, by coding and computing with probability density functions, or approximations to probability density functions" (op cit, P. 713)

Such a mode of representation implies that when we represent a state or feature of the world, such as the depth of a visible object, we do so not using a single computed value but using a conditional probability density function that encodes "the relative probability that the object is at different depths Z, given the available sensory information" (op cit 712). The same story applies to higher-level states and features. Instead of simply representing "CAT ON MAT" the probabilistic Bayesian brain will encode a conditional probability density function, reflecting the relative probability of this state of affairs (and any somewhat-supported alternatives) given the available information. This information-base will include both the bottom up driving influences from multiple sensory channels, and top-down context-fixing information of various

kinds. During the early stages of processing, the system will avoid committing itself to any single interpretation, and there will often be an initial flurry of error signals (which are said to constitute a major component of early evoked responses- see e.g. Friston (2005) p.829) as competing 'beliefs' propagate up and down the system. This is typically followed by rapid convergence upon a dominant theme (CAT, MAT), with further details (STRIPEY MAT, TABBY CAT) subsequently negotiated. The set-up thus favours a kind of recurrently negotiated 'gist-at-a-glance' model, where we first identify the general scene (perhaps including general affective elements too – for a fascinating discussion see Barrett and Bar (2009)) followed by the details. This affords a kind of 'forest first, trees second' approach (Hochstein and Ahissar (2002), Friston (2005) p.825).

This does not mean, however, that context effects will always take time to emerge and propagate downwards[17]. For in many (indeed, most) real-life cases, substantial context information is already in place when new information is encountered. An apt set of priors is thus often already active, poised to impact the processing of new sensory inputs without further delay. This is important. The brain, in ecologically normal circumstances, is not just suddenly 'turned on' and some random or unexpected input delivered for processing. So there is plenty of room for top-down influence to occur even before a stimulus is presented. In the event that we already know we are in a forest (perhaps we have been hiking for hours) there has still been prior settling into a higher level representational state. But such settling need not occur within the temporal span following each new sensory input[18]. Over whatever time-scale, though, the endpoint (assuming we form a rich visual percept) is the same. The system will have settled into a set of states that make mutually consistent bets concerning many aspects of the scene (from the general theme all the way down to more spatio-temporally precise information about parts, colours, and orientations etc.). At each level, the underlying mode of representation will remain thoroughly probabilistic, encoding a series of intertwined bets concerning all the elements (at the various spatio-temporal scales) that make up the perceived scene.

In what sense are such systems truly Bayesian? According to Knill and Pouget:

> "The real test of the Bayesian coding hypothesis is in whether the neural computations that result in perceptual judgments or motor behaviour take into account the uncertainty available at each stage of the processing" Knill and Pouget (2004) 713

That is to say, reasonable tests will concern how well a system deals with the uncertainties that characterize the information it actually manages to encode and process, and (I would add) the general shape of the strategies it uses to do so. There is increasing (though mostly indirect- see 3.1 following) evidence that biological systems approximate, in multiple domains, the Bayesian profile thus understood. To take just one example (for others, see section 3.1 following) Weiss et al (2002) - in a paper revealingly titled "Motion illusions as optimal percepts" - used an optimal Bayesian estimator (the so-called 'Bayesian ideal observer') to show that a wide variety of psychophysical results, including many motion 'illusions', fall naturally out of the assumption that human motion perception implements just such an estimator mechanism[19]. They conclude that:

> "Many motion 'illusions' are not the result of sloppy computation by various components in the visual system, but rather a result of a coherent computational strategy that is optimal under reasonable assumptions" Weiss et al (2002) p. 603

Examples could be multiplied (see Knill and Pouget (2004) for a balanced review). At least in the realms of low-level, basic, and adaptively crucial, perceptual and motoric computations, biological processing may quite closely approximate Bayes' optimality. But what researchers find in general is not that we humans are - rather astoundingly - 'Bayes' optimal' in some absolute sense (i.e. responding correctly relative to the absolute uncertainties in the stimulus) but rather that we are often optimal, or near optimal, at taking into account the uncertainties that characterize the information that we actually command: the information that is made available by the forms of sensing and processing that we actually deploy (see Knill and Pouget (2004, p. 713)). That means, taking into account the uncertainty in our own sensory and motor signals, and adjusting the relative weight of different cues according to (often very subtle) contextual clues. Recent work confirms and extends this assessment, suggesting that humans act as rational Bayesian estimators, in perception and in action, across a wide variety of domains (Yu (2007), Berniker and Körding (2008), Körding, Tenenbaum, and Shadmehr (2007)). Of course, the mere fact that a system's response profiles take a certain shape does not itself demonstrate that that system is implementing some form of Bayesian reasoning. In a limited domain, a look-up table could (Maloney and Mamassian (2009)) yield the same behavioural repertoire as a 'Bayes' optimal' system. Nonetheless, the hierarchical predictive processing story, if correct, would rather directly underwrite the claim that the nervous system approximates, using tractable computational strategies, a genuine version of Bayesian inference. The

computational framework of hierarchical predictive processing realizes, using the signature mix of top-down and bottom-up processing, a robustly Bayesian inferential strategy, and there is mounting neural and behavioural evidence (again, see 3.1 following) that such a mechanisms is implemented in the brain. Experimental tests have also recently been proposed (Maloney and Mamassian (2009), Maloney and Zhang (In Press)) which aim to 'operationalize' the claim that a target system is (genuinely) computing its outputs using a Bayesian scheme, rather than merely behaving 'as if' it did so. This, however, is an area that warrants a great deal of further cognitive scientific thought and investigation.

Hierarchical predictive processing models also suggest something about the nature of the representation relation itself. To see this, recall (1.2 above) that hierarchical predictive coding, in common with other approaches deploying a cascade of top-down processing to generate low-level states from high-level causes, offers a way to get at the world from 'inside' the black box. That procedure (which will work in all worlds where the next state in space or time is usefully predicted by the present one - see Hosoya (2005), Schwarz, Hsu and Dayan (2007)) allows the learner reliably to match her internal generative models to the statistical properties of the signal source (the world) yielding contents that are, I submit, as 'grounded' (Harnad (1990)) and 'intrinsic' (Adams and Aizawa (2001)) as any philosopher could wish for. Such models thus deliver a novel framework for thinking about neural representation and processing, and a compelling take on the representation relation itself: one that can be directly linked (via the Bayesian apparatus) to rational processes of learning and belief fixation.

2.3 The Delicate Dance Between Top-Down and Bottom-Up

In the context of hierarchical models of brain function, action-oriented predictive processing yields a new account of the complex interplay between top-down and bottom-up influences on perception and action, and perhaps ultimately of the relations between perception, action, and cognition. As noted by Hohwy (2007, p.320) the generative models providing 'top-down' predictions are here doing much of the more traditionally 'perceptual' work, with the bottom-up driving signals really providing a kind of ongoing feedback on their activity (by fitting or failing to fit, the cascade of downward-flowing predictions). This procedure combines 'top-down' and 'bottom–up' influences in an especially delicate and potent fashion, and leads to the development of neurons that exhibit a "selectivity that is not intrinsic to the area but depends

on interactions across levels of a processing hierarchy" (Friston (2003) p.1349). Hierarchical predictive coding delivers, that is to say, a processing regime in which context-sensitivity is fundamental and pervasive. To see this, we need only reflect that the neuronal responses that follow an input (the 'evoked responses') may be expected to alter quite profoundly according to the contextualizing information provided by a current winning top-down prediction. The key effect here (itself familiar enough from earlier connectionist work using the 'interactive activation' paradigm – see e.g. Rumelhart and McClelland and Rumelhart (1981), Rumelhart, McClelland et al (1986)) is that "when a neuron or population is predicted by top-down inputs it will be much easier to drive than when it is not" Friston (2002) p. 240. This is because the best overall fit between driving signal and expectations will often be found by (in effect) inferring noise in the driving signal and thus recognizing a stimulus as for example, the letter 'm' (say, in the context of the word 'mother') even though the same bare stimulus, presented out of context or in most other contexts, would have been a better fit with the letter 'n'[20]. A unit normally responsive to the letter 'm' might, under such circumstances, be successfully driven by an n-like stimulus. Such effects are pervasive in hierarchical predictive processing, and have far-reaching implications for various forms of neuroimaging. It becomes essential, for example, to control as much as possible for expectations when seeking to identify the response selectivity of neurons or patterns of neural activity. Strong effects of top-down expectation have also recently been demonstrated for conscious recognition, raising important questions about the very idea of any simple (i.e. context independent) 'neural correlates of consciousness. Thus Melloni et al (2011) show that the onset time required to form a reportable conscious percept varies substantially (by around 100ms) according to the presence or absence of apt expectations, and that the neural (here, EEG) signatures of conscious perception vary accordingly - a result they go on to interpret using the apparatus of hierarchical predictive processing. Finally, in a particularly striking demonstration of the power of top-down expectations, Egner, Monti and Summerfield (2010) show that neurons in the fusiform face area (FFA) respond every bit as strongly to non-face (in this experiment, house) stimuli under high expectation of faces as they do to face-stimuli. In this study:

> "FFA activity displayed an interaction of stimulus feature and expectation factors, where the differentiation between FFA responses to face and house stimuli decreased linearly with increasing levels of face expectation, with face and house evoked signals being indistinguishable under high face expectation" Egner, Monti, and Summerfield (2010) p. 16607.

Only under conditions of low face expectation was FFA response maximally different for the face and house probes, suggesting that "[FFA] responses appear to be determined by feature expectation and surprise rather than by stimulus features per se" op cit p.16601. The suggestion, in short, is that FFA (in many ways the paradigm case of a region performing complex feature detection) might be better treated as a face-expectation region rather than a face-detection region: a result that the authors interpret as favoring a hierarchical predictive processing model. The growing body of such results leads Muckli to comment that:

> "Sensory stimulation might be the minor task of the cortex, whereas its major task is to […] predict upcoming stimulation as precisely as possible" Muckli (2010) p.137

In a similar vein, Rauss et al (in press) suggest that:

> "neural signals are related less to a stimulus per se than to its congruence with internal goals and predictions, calculated on the basis of previous input to the system" Rauss et al (in press)

Attention fits very neatly into this emerging unified picture, as a means of variably balancing the potent interactions between top-down and bottom-up influences by factoring in their precision (degree of uncertainty). This is achieved by altering the gain (the 'volume' to use a common auditory analogy) on the error-units accordingly. The upshot of this is to "control the relative influence of prior expectations at different levels" (Friston (2009) p. 299). In recent work effects of the neurotransmitter dopamine are presented as one possible neural mechanism for encoding precision (see Fletcher and Frith (2009) p. 53-54, who refer the reader to work on prediction error and the mesolimbic dopaminergic system such as Holleman and Schultz (1998), Waelti et al (2001)). Greater precision (however encoded) means less uncertainty, and is reflected in a higher gain on the relevant error units (see Friston (2005), (2010), Friston, Daunizeau, and Kiebel (2009)). Attention, if this is correct, is simply one means by which certain error-unit responses are given increased weight, hence becoming more apt to drive learning and plasticity, and to engage compensatory action.

More generally, this means that the precise mix of top-down and bottom-up influence is not static or fixed. Instead, the weight that is given to sensory prediction error is varied according to how reliable (how noisy, certain, or

uncertain) the signal is taken to be. This is (usually) good news, as it means we are not (not quite) slaves to our expectations. We are still able to see very surprising things, at least in conditions where the brain assigns high reliability to the driving signal. But (importantly) that then requires that other high-level theories, though of an initially agent-unexpected kind, then win out so as to explain away the highly-weighted sensory evidence. In extreme and persistent cases (more on this in 4.2 following) dealing with the error may require new generative models gradually to be formed, in what Fletcher and Frith (2009, p.53) nicely describe as a 'reciprocal interaction between perception and learning'.

All this makes the lines between perception and cognition fuzzy, perhaps even vanishing. In place of any real distinction between perception and belief we now get variable differences in the mixture of top-down and bottom-up influence, and differences of temporal and spatial scale in the generative models that are making the predictions. Top level (more 'cognitive') generative models intuitively correspond to increasingly abstract conceptions of the world and these tend to capture or depend upon regularities at larger temporal and spatial scales. Lower level (more 'perceptual') ones capture or depend upon the kinds of scale and detail most strongly associated with specific kinds of perceptual contact. But it is the precision-modulated constant content-rich interactions between these levels, as mediated by ongoing motor action of one kind or another, that now emerges as the heart of intelligent, adaptive response.

These accounts thus appear to dissolve, at the level of the implementing neural machinery, the superficially clean distinction between perception and knowledge/belief. To perceive the world just is to use what you know to explain away the sensory signal across multiple spatial and temporal scales. The process of perception is thus inseparable from rational (broadly Bayesian) processes of belief fixation, and context (top down) effects are felt at every intermediate level of processing. As thought, sensing, and movement here unfold, we discover no stable or well-specified interface or interfaces between cognition and perception. Believing and perceiving, although conceptually distinct, emerge as deeply mechanically intertwined. They are constructed using the same computational resources, and (as we shall see in 4.2) are mutually, reciprocally, entrenching.

Summary So Far

Action-oriented (hierarchical) predictive processing models promise to bring cognition, perception, action, and attention together within a common framework. This framework suggests probability density distributions induced by hierarchical generative models as our basic means of representing the world, and prediction-error minimization as the driving force behind learning, action-selection, recognition, and inference. Such a framework offers new insights into a wide range of specific phenomena including non-classical receptive field effects, bi-stable perception, cue integration, and the pervasive context-sensitivity of neuronal response. It makes rich and illuminating contact with work in cognitive neuroscience while boasting a firm foundation in computational modelling and Bayesian theory. It thus offers what is arguably the first truly systematic bridge[21] between three of our most promising tools for understanding mind and reason: cognitive neuroscience, computational modelling, and probabilistic Bayesian approaches to dealing with evidence and uncertainty.

3. From Action-Oriented Predictive Processing to an Architecture of Mind.

Despite that truly impressive list of virtues, both the hierarchical predictive processing family of models and their recent generalizations to action face a number of important challenges, ranging from the evidential (what are the experimental and neuroanatomical implications, and to what extent are they borne out by current knowledge and investigations?) to the conceptual (can we really explain so much about perception and action by direct appeal to a fundamental strategy of minimizing errors in the prediction of sensory input?) to the more methodological (to what extent can these accounts hope to illuminate the full shape of the human cognitive architecture?) In this section I address each challenge in turn, before asking (section 4) how such models relate to our conscious mental life.

3.1 The Neural Evidence

Direct neuroscientific testing of the hierarchical predictive coding model, and of its action-oriented extension, remains in its infancy. The best current evidence tends to be indirect, and comes in two main forms. The first (which is highly indirect) consists in demonstrations of precisely the kinds of optimal sensing and motor control that the 'Bayesian brain hypothesis' (2.2 above) suggests. Good examples here include compelling bodies of work on cue

integration (see also 2.2 above and 4.3 following) showing that human subjects are able optimally to weight the various cues arriving through distinct sense modalities, doing so in ways that delicately and responsively reflect the current (context-dependent) levels of uncertainty associated with the information from different channels (Ernst and Banks (2002), Knill and Pouget (2004) – and for further discussion, see Mamassian et al (2002), Rescorla (In Press)). This is beautifully demonstrated, in the case of combining cues from vision and touch, by Bayesian models such as Helbig and Ernst (2007). Similar results have been obtained for motion perception, neatly accounting for various illusions of motion perception by invoking statistically-valid priors that favor lower and smoother motions – see Weiss, Simoncelli, and Adelson (2002), Ernst (2010)). Another example is the Bayesian treatment of color perception (see Brainerd (2009)) which again accounts for various known effects (here, color constancies and some color illusions) in terms of optimal cue combination.

The success of the Bayesian program in these arenas (for some more examples, see Rescorla (In Press) and section 4.4 following) is impossible to doubt. It is thus a major virtue of the hierarchical predictive coding account that (1.2 and 2.2 above: see also Friston (2003), (2005)) it effectively implements a computationally tractable version of the so-called 'Bayesian Brain Hypothesis' (Knill and Pouget (2004), Doya et al (eds) 2007)). But behavioral demonstrations of Bayesian-like performance, though intrinsically interesting and clearly suggestive, cannot establish strong conclusions about the shape of the mechanisms generating those behaviors.

More promising in this regard are other forms of indirect evidence, such as the ability of computational simulations of predictive coding strategies to reproduce and explain a variety of observed effects. These include non-classical receptive field effects, repetition suppression effects, and the bi-phasic response profiles of certain neurons involved in low-level visual processing. Thus consider (Rao and Sejnowski (2002)) non-classical receptive field effects. In one such effect, an oriented stimulus yields a strong response from a cortical cell but that response is suppressed when the surrounding region is filled with a stimulus of identical orientation, and is enhanced when the orientation of the central stimulus is orthogonal to those of the surrounding region. This is a surprising set of features. A powerful explanation of this result, Rao and Sejnowksi (2002) suggest, is that the observed neural response here signals *error* rather than some fixed content. It is thus smallest when the central stimulus is highly predictable from the surrounding ones, and largest when it is actively counter-predicted by the surroundings. A related account (Rao and Ballard (1999), based on the simulation study sketched in 1.2 above)) explains 'end-

stopping' effects, in which a lively neural response to a preferred stimulus such as an oriented line segment ceases or becomes reduced when the stimulus extends farther than the neuron's standard receptive field. Here too, computational simulations using the predictive coding strategy displayed the same effect. This is because is the natural images used to train the network contained many more instances of these longer line segments, facilitating prediction in (and only in) such cases. Extended line segments are thus more predictable, so error-signaling responses are reduced or eliminated. In short, the effect is explained once more by the assumption that activity in these units is signaling error/mismatch. Similarly, Jehee and Ballard (2009) offer a predictive processing account of 'biphasic response dynamics' in which the optimal stimulus for driving a neuron (such as certain neurons in LGN - lateral geniculate nucleus) can reverse (e.g. from preferring bright to preferring dark) in a short (20 ms) space of time. Once again the switch is neatly explained as a reflection of a unit's functional role as an error or difference detector rather than a feature detector as such. In such cases, the predictive coding strategy (1.1 above) is in full evidence since:

> "Low-level visual input [is] replaced by the difference between the input and a prediction from higher-level structures….higher-level receptive fields…represent the predictions of the visual world while lower-level areas…signal the error between predictions and the actual visual input" Jehee and Ballard (2009) p.1

Finally, consider the case of 'repetition suppression'. Multiple studies (for a recent review, see Grill-Spector et al (2006)) have shown that stimulus-evoked neural activity is reduced by stimulus repetition. Summerfield et al (2008) manipulated the local likelihood of stimulus repetitions, showing that the repetition-suppression effect is itself reduced when the repetition is improbable/unexpected. The favored explanation is (again) that repetition normally reduces response because it increases predictability (the second instance was made likelier by the first) and thus reduces prediction error. Repetition suppression thus also emerges as a direct effect of predictive processing in the brain, and as such its severity may be expected to vary (just as Summerfield et al found) according to our local perceptual expectations. In general then, the predictive coding story offers a very neat and unifying explanation, of a wide variety of such contextual effects.

Can we find more direct forms of evidence too? Functional imaging plays an increasing role here. For example, an fMRI study by Murray et al (2002)

revealed just the kinds of relationships posited by the predictive processing (hierarchical predictive coding) story. As higher level areas settled into an interpretation of visual shape, activity in V1 was dampened, consistent with the successful higher level predictions being used to explain away (cancel out) the sensory data. More recently, Alink et al (2010) found decreased responses for predictable stimuli using variants on an apparent motion illusion, while den Ouden et al (2010) report similar results using arbitrary contingencies that were manipulated rapidly during the course of their experiments[22]. Finally, the study by Egner, Monti, and Summerfield (2010) (and described in 2.3 above) went on to compare, in simulation, several possible models that might be used to account for their results. The authors found a predictive processing regime involving the co-presence of representation and error units (2.1 above) to offer by far the best fit for their data. In that best-fit simulation, error ('face-surprise') units are modeled as contributing twice as much to the fMRI signal as representation ('face-expectation') units, leading the authors to comment that:

> "the current study is to our knowledge the first investigation to formally and explicitly demonstrate that population responses in visual cortex are in fact better characterized as a sum of feature expectation and surprise responses than by bottom-up feature detection…" Egner et al (2010) 16607

The predictive processing model also suggests testable hypotheses concerning the ways in which interfering (e.g. using TMS - transcranial magnetic stimulation - or other methods) with the message passing routines linking higher to lower cortical areas should impact performance. To take one specific example, the model of binocular rivalry rehearsed in section 1.4 above predicts that:

> "LGN and blind spot representation activity measured with fMRI will not suggest that rivalry is resolved before binocular convergence, if deprived of backwards signals from areas above binocular convergence" Hohwy, Roepstorff, and Friston (2008) p.699

In general, if the predictive processing story is correct, we expect to see powerful context effects propagating quite low down the processing hierarchy. The key principle – and one that also explains many of the observed dynamics of evoked responses - is that (subject to the caveats mentioned earlier concerning already active expectations) "representations at higher levels must emerge before backward afferents can reshape the response profile of neurons in lower areas" (Friston (2003) p.1348). In the case of evoked responses, the

suggestion (Friston (2005) section 6) is that an early component often tracks an initial flurry of prediction error: one that is soon suppressed (assuming the stimulus is not novel or encountered out of its normal context) by successful predictions flowing backwards from higher areas. Such temporal delays, which are exactly what one would expect if perception involves recruiting top-level models to explain away sensory data, are now widely reported in the literature (see e.g. Pack and Born (2001), Born et al (2009)).

One extremely important and as yet not well-tested implication of the general architectural form of these models is (recall 2.1 above) that each layer of processing should contain two functionally distinct sub-populations of units (neurons, or neuronal features such as the distinct apical and basal dendrites of pyramidal cells – see Spratling and Johnson (2006)). One sub-population, recall, is doing the 'real' work of representing the current sensory cause: these units ('representational neurons' or 'state units') encode the area's best guess, in context as processed so far, at the current stimulus. They thus encode what Friston (2005, p.829) describes as the area's 'conditional expectations of perceptual causes'. The other sub-population is in the business of encoding precision-weighted prediction errors: these units (so-called 'error units') fire when there is a mismatch between what is predicted and what is apparently being observed. The two sets of units are assumed to interact in the manner prescribed by the hierarchical predictive coding model. That is to say, the error units process signals from the representation units both at their own level and at the level above, and the representation units send signals to the error units both at their own level and at the level below. Forward connections thus convey error, while backwards connections are free to construct (in a more complex non-linear fashion) predictions that aim to cancel out the error. Unfortunately, direct, unambiguous neural evidence for these crucial functionally distinct sub-populations is still missing. Hence:

> "...one limitation of these models – and of predictive coding in general-is that to date no single neuron study has systematically pursued the search for sensory prediction error responses" Summerfield and Egner (2009, p.408)

The good news is that there is, as we saw, mounting and converging indirect evidence for such a cortical architecture in the form (largely) of increased cortical responses to sensory surprise. Crucially, there also exists (2.1 above) a plausible neuronal implementation for such a scheme involving superficial and deep pyramidal cells. Nonetheless much more evidence is clearly needed for the

existence of the clean functional separation (between the activity of different neuronal features or sub-populations) required by these models[23].

3.2 Scope and Limits

> "In the ultimate stable state, the deep pyramidals [conveying predictions downwards] would send a signal that perfectly predicts what each lower area is sensing, up to expected levels of noise, and the superficial pyramidals [conveying prediction errors upwards] wouldn't fire at all." Mumford (1992) p.247

> "In some sense, this is the state that the cortex is trying to achieve: perfect prediction of the world, like the oriental Nirvana, as Tai-Sing Lee suggested to me, when nothing surprises you and new stimuli cause the merest ripple in your consciousness" (op cit, p.247 footnote ♯ 5)

There is a very general worry that is sometimes raised in connection with the large-scale claim that cortical processing fundamentally aims to minimize prediction error, thus quashing the forward flow of information and achieving what Mumford (above) evocatively describes as the 'ultimate stable state'. It can be put like this:

> "How can a neural imperative to minimize prediction error by enslaving perception, action, and attention accommodate the obvious fact that animals don't simply seek a nice dark room and stay in it? Surely staying still inside a darkened room would afford easy and nigh-perfect prediction of our own unfolding neural states? Doesn't the story thus leave out much that really matters for adaptive success: things like boredom, curiosity, play, exploration, foraging, and the thrill of the hunt?"

The simple response (correct, as far as it goes) is that animals like us live and forage in a changing and challenging world, and hence 'expect' to deploy quite complex 'itinerant' strategies (Friston, Daunizeau, and Kiebel (2009), Friston (2010)) to stay within our species-specific window of viability. Change, motion, exploration, and search are *themselves* valuable for creatures living in worlds where resources are unevenly spread and new threats and opportunities continuously arise. This means that change, motion, exploration, and search themselves become predicted, and poised to enslave action and perception accordingly. One way to unpack this idea (see Friston (2010b) p.8) would be to look at the possible role of priors that induce motion through a state space

until an acceptable, though possibly temporary or otherwise unstable, stopping point (an attractor) is found. In precisely this vein Friston (2011b, p.113) comments that "some species are equipped with prior expectations that they will engage in exploratory or social play".

The whole shape of this space of prior expectations is specific to different species and may also vary as a result of learning and experience. Hence nothing in the large-scale story about prediction error minimization dictates any general or fixed balance between what is sometimes glossed as 'exploration' versus 'exploitation' (for some further discussion of this issue, see Friston and Stephan (2007) p. 435-436). Instead, different organisms amount (Friston (2011b)) to different 'embodied models' of their specific needs and environmental niches, and their expectations and predictions are formed, encoded, weighted, and computed against such backdrops. This is both good news, and bad news. It's good, because it means the stories on offer can indeed accommodate all the forms of behaviour (exploration, thrill-seeking etc) we see. But it's bad (or at least, limiting) because it means that the accounts don't in themselves tell us much at all about these key features: features which nonetheless condition and constrain an organism's responses in a variety of quite fundamental ways.

In one way, of course, this is clearly unproblematic. The briefest glance at the staggering variety of biological (even mammalian) life forms tells us that whatever fundamental principles are sculpting life and mind, they are indeed compatible with an amazing swathe of morphological, neurological, and ethological outcomes. But in another way it can still seem disappointing. If what we want to understand is the specific functional architecture of the human mind, the distance between these very general principles of prediction-error minimization and the specific solutions to adaptive needs that we humans have embraced remains daunting. As a simple example, notice that the predictive processing account leaves wide open a variety of deep and important questions concerning the nature and format of human neural representation. The representations on offer are, we saw, constrained to be probabilistic (and generative model based) through and through. But that is compatible with the use of the probabilistic-generative mode to encode information using a wide variety of different schemes and surface forms. Consider the well-documented differences in the way the dorsal and ventral visual streams code for attributes of the visual scene. The dorsal stream (Milner and Goodale (2006)) looks to deploy modes of representation and processing that are *at some level of interest* quite distinct from those coded and computed in the ventral stream. And this will be true even if there is indeed, at some more fundamental level, a common computational strategy at work throughout the visual and the motor cortex.

Discovering the nature of various inner representational formats is thus representative of the larger project of uncovering the full shape of the human cognitive architecture. It seems likely, as argued by Eliasmith (2007) that this larger project will demand a complex combination of insights, some coming 'top-down' from theoretical (mathematical, statistical, and computational) models, and others coming 'bottom up' from neuroscientific work that uncovers the brain's actual resources as sculpted by our unique evolutionary (and – as we'll next see- sociocultural) trajectory.

3.3 Neats versus Scruffies (21st Century Replay).

Back in the late nineteen seventies and early nineteen eighties (the heyday of classical Artificial Intelligence) there was a widely held view that two personality types were reflected in theorizing about the human mind. These types were dubbed[24], by Roger Schank and Robert Abelson, the 'neats' versus the 'scruffies'. Neats believed in a few very general, truth-conducive principles underlying intelligence. Scruffies saw intelligence as arising from a varied bag of tricks: a rickety tower of rough and ready solutions to problems, often assembled using various quick patches and local ploys, and greedily scavenging the scraps and remnants of solutions to other, historically prior, problems and needs. Famously, this can lead to scruffy, unreliable, or sometimes merely unnecessarily complex solutions to ecologically novel problems such as planning economies, building railway networks, and maintaining the internet. Such historically path-dependent solutions were sometimes called 'kluges' –see e.g. Clark (1987) Marcus (2008). Neats favoured logic and provably correct solutions, while scruffies favoured whatever worked reasonably well, fast enough, in the usual ecological setting, for some given problem. The same kind of division emerged in early debates between connectionist and classical AI (see e.g. Sloman (1990)) with connectionists often accused of developing systems whose operating principles (after training on some complex set of input-output pairs) was opaque and 'messy'. The conflict reappears in more recent debates (Griffiths et al (2010), McClelland et al (2010)) between those favouring 'structured probabilistic approaches' and those favouring 'emergentist' approaches (where these are essentially connectionist approaches of the parallel distributed processing variety)[25].

My own sympathies (Clark (1989), (1997)) have always lain more on the side of the scruffies. Evolved intelligence, it seemed to me (Clark (1987)) was bound to involve a kind of unruly motley of tricks and ploys, with significant path-dependence, no premium set on internal consistency, and fast effective situated

response usually favoured at the expense of slower, more effortful, even if more truth-conducive modes of thought and reasoning. Seen through this lens, the 'Bayesian brain' seems, at first glance, to offer an unlikely model for evolved biological intelligence. Implemented by hierarchical predictive processing, it posits a single fundamental kind of learning algorithm (based on generative models, predictive coding, and prediction error minimization) that approximates the rational ideal of Bayesian belief update. Suppose such a model proves correct. Would this amount to the final triumph of the neats over the scruffies? I suspect it would not, and for reasons that shed additional light upon the questions about scope and limits raised in the previous section.

Favoring the 'neats', we have encountered a growing body of evidence (2.2 and 2.3 above) showing that for many basic problems involving perception and motor control, human agents (as well as other animals) do indeed manage to approximate the responses and choices of optimal Bayesian observers and actors. Nonetheless, a considerable distance still separates such models from the details of their implementation in humans or other animals. It is here that the apparent triumph of the neats over the scruffies may be called into question. For the Bayesian brain story tells us, at most, what the brain (or better, the brain in action) manages to compute. It also suggests a good deal about the forms of representation and computation that the brain must deploy: for example, it suggests (2.2 above) that it must deploy a probabilistic representation of sensory information, that it must take into account uncertainty in its own sensory signals, estimate the 'volatility' (frequency of change) of the environment itself (Yu (2007)), and so on. But that still leaves plenty of room for debate and discovery as regards the precise shape of the large-scale cognitive architecture within which all this occurs.

The hierarchical predictive processing account takes us a few important steps further. It offers a computationally tractable approximation to true Bayesian inference. It says something about the basic shape of the cortical micro-circuitry. And, at least in the formulations I have been considering, it predicts the presence of distinct neural encodings for representation and error. But even taken together, the mathematical model (the Bayesian brain) and the hierarchical, action-oriented, predictive processing implementation fail to specify the overall form of a cognitive architecture. They fail to specify, for example, how the brain (or better, the brain in the context of embodied action) divides its cognitive labours between multiple cortical and subcortical areas, what aspects of the actual world get sensorially coded in the first place, or how best to navigate the exploit-explore continuum (the grain of truth in the 'darkened room' worry discussed in 3.2 above)). It also leaves unanswered a

wide range of genuine questions concerning the representational formats used by different brain areas or for different kinds of problems. This problem is only compounded once we reflect (Anderson (2007), 3.4 following) that the brain may well tackle many problems arising later in its evolutionary trajectory by cannily redeploying resources that were once used for other purposes.

In the most general terms, then, important questions remain concerning the amount of work (where the goal is that of understanding the full human cognitive architecture) that will be done by direct appeal to action-oriented predictive processing and the amount that will still need to be done by uncovering evolutionary and developmental trajectory-reflecting tricks and ploys: the scruffy kluges that gradually enabled brains like ours to tackle the complex problems of the modern world.

3.4 Situated Agents

We may also ask what, if anything, the hierarchical predictive processing perspective suggests concerning situated, world-exploiting agency (Thelen and Smith (1994), Hutchins (1995), Wilson (1994) (2004), Haugeland (1998), Hurley (1998), Clark (1997) (2008), Clark and Chalmers (1998), Rowlands (1999) (2006), Noë (2004), (2009), Wheeler (2005), Menary (2007)). At least on the face of it, the predictive processing story seem to pursue a rather narrowly neurocentric focus, albeit one that reveals (1.5 above) some truly intimate links between perception and action. But dig a little deeper and what we discover is a model of key aspects of neural functioning that makes structuring our worlds genuinely continuous with structuring our brains and sculpting our actions. Cashing out all the implications of this larger picture is a future project, but a brief sketch may help set the scene.

Recall (1.5 and 1.6 above) that these models display perception and action working in productive tandem to reduce 'surprisal' (where this measures the implausibility of some sensory state given a model of the world). Perception reduces surprisal by assimilating inputs to prior expectations. Action reduces surprisal by altering the world (including moving the body) so as to conform with expectations. Working together, perception and action serve to selectively sample and actively sculpt the stimulus array. These direct links to active sculpting and selective sampling suggest deep synergies between the hierarchical predictive processing framework and work in embodied and situated cognition. For example, work in mobile robotics already demonstrates a variety of concrete ways in which perception and behaviour productively interact via loops through action and the environment: loops that may now be

considered as affording extra-neural opportunities for the minimization of prediction error. In precisely this vein, Verschure et al (2003), in work combining robotics and statistical learning, note that "behavioural feedback modifies stimulus sampling and so provides an additional extra-neuronal path for the reduction of prediction errors". (Verschure et al (2003) p.623)

More generally, consider recent work on the 'self-structuring of information flows'. This work, as the name suggests, stresses the importance of our own action-based structuring of sensory input (for example, the linked unfolding across multiple sensory modalities that occurs when we see, touch, and hear an object that we are actively manipulating). Such information self-structuring has been shown to promote learning and inference – see e.g. Pfeifer, et al (2007) and discussion in Clark (2008). Zahedi et al (in press) translate these themes directly into the present framework using robotic simulations in which the learning of complex co-ordination dynamics is achieved by maximizing the amount of predictive information present in sensorimotor loops.

Extensions into the realm of social action and multi-agent co-ordination are then close to hand. For a key proximal goal of information self-structuring, considered from the action-oriented predictive processing perspective, is the reduction of *mutual prediction error* as we collectively negotiate new and challenging domains (see, for example, recent work on synchronization and shared musical experience (Overy & Molnar-Szakacs (2009)) and the 'culture as patterned practices' approach suggested by Roepstorff, Niewöhner, and Beck (20100)). Such a perspective, by highlighting situated practice, very naturally encompasses various forms of longer-term material and social environmental structuring. Using a variety of tricks, tools, notations, practices, and media, we structure our physical and social worlds so as to make them friendlier for brains like ours. We colour-code consumer products, we drive on the right (or left), paint white lines on roads, and post prices in supermarkets. At multiple time-scales, and using a wide variety of means (including words, equations, graphs, other agents, pictures, and all the tools of modern consumer electronics) we thus stack the dice so that we can more easily minimize costly prediction errors in an endlessly empowering cascade of contexts from shopping and socializing, to astronomy, philosophy, and logic.

Consider, from this perspective, our many symbol-mediated loops into material culture via notebooks, sketchpads, iphones, and (Pickering and Garrod (2007)) conversations with other agents (and for some intriguing speculations concerning the initial emergence of all those discrete symbols in predictive, probabilistic contexts, see König, and Krüger (2006)). Such loops are

effectively enabling new forms of reentrant processing. They take a highly processed cognitive product (such as an idea about the world) clothe it in public symbols, and launch it out into the world so that it can re-enter our own system as a concrete perceptible (Clark (2006) (2008)), and one now bearing highly informative statistical relations to other such linguaform perceptibles[26]. It is courtesy of all that concrete public vehicling in spoken words, written text, diagrams, and pictures that *our* best models of reality (unlike those of other creatures) are stable, re-inspectable objects apt for public critique and refinement. Our best models of the world are thus the basis for cumulative communally distributed reasoning, rather than just the means by which individual thoughts occur. The same potent processing regimes, now targeting these brand new types of statistically-pregnant 'designer inputs', are then enabled to discover and refine new generative models, latching onto (and at times actively creating) ever more abstract structure in the world. Action and perception thus work together to reduce prediction error against the more slowly evolving backdrop of a culturally distributed process that spawns a succession of designer environments whose impact on the development (e.g. Smith and Gasser (2005)) and unfolding of human thought and reason can hardly be over-estimated.

Such culturally mediated processes may incur costs (3.3 above) in the form of various kinds of path-dependence (Arthur (1994)) in which later solutions build on earlier ones. In the case at hand, path-based idiosyncrasies may become locked in as material artifacts, institutions, notations, measuring tools, and cultural practices. But it is that very same trajectory-sensitive process that delivers the vast cognitive profits that flow from the slow, multi-generational development of stacked, complex 'designer environments' for thinking such as mathematics, reading[27], writing, structured discussion, and schooling, in a process that Sterelny (2003) nicely describes as 'incremental downstream epistemic engineering'. The upshot is that the human-built environment becomes a potent source of new heritable structure that surrounds our biological brains (see e.g. Oyama (1999), Griffiths & Gray (2001), Sterelny (2007), Stotz (2010)).

What are the potential effects of such stacked and heritable designer environments upon prediction-driven learning in cortical hierarchies? Such learning routines make human minds permeable, at multiple spatial and temporal scales, to the statistical structure of the world as reflected in the training signals. But those training signals are now delivered as part of a complex developmental web that gradually comes to include all the complex regularities embodied in the web of statistical relations among the symbols and

other forms of socio-cultural scaffolding in which we are immersed. We thus self-construct a kind of rolling 'cognitive niche' able to induce the acquisition of generative models whose reach and depth far exceeds their apparent base in simple forms of sensory contact with the world. The combination of 'iterated cognitive niche construction' and profound neural permeability by the statistical structures of the training environment is both potent and self-fueling. When these two forces interact, repeatedly reconfigured agents are enabled to operate in repeatedly reconfigured worlds, and the human mind becomes a constantly moving target. The full potential of the prediction-error minimization model of how cortical processing *fundamentally* operates will emerge only (I submit) when that model is paired with an appreciation of what immersion in all those socio-cultural designer environments can do (for some early steps in this direction, see Roepstorff, Niewöhner, and Beck (2010)). Such a combined approach would implement a version of so-called 'neuroconstructivism' (Mareschal et al (2007)) which asserts that:

> "The architecture of the brain.. and the statistics of the environment, [are] not fixed. Rather, brain-connectivity is subject to a broad spectrum of input-, experience-, and activity-dependent processes which shape and structure its patterning and strengths… These changes, in turn, result in altered interactions with the environment, exerting causal influences on what is experienced and sensed in the future" (Sporns (2007) p.179)

All this suggests a possible twist upon the worries (3.2 and 3.3 above) concerning the ability of the predictive processing framework to specify a full-blown cognitive architecture. Perhaps that lack is not a vice but a kind of virtue? For what is really on offer, or so it seems to me, is best seen as a framework whose primary virtue is to display some deep unifying principles covering perception, action, and learning. That framework in turn reveals us as highly responsive to the statistical structures of our environments, including the cascade of self-engineered 'designer environments'. It thus offers a standing invitation to evolutionary, situated, embodied, and distributed approaches to help 'fill in the explanatory gaps' while delivering a schematic but fundamental account of the complex and complementary roles of perception, action, attention, and environmental structuring.

4. Content and Consciousness

How, finally, do the accounts on offer relate to a human mental life? This, of course, is the hardest – though potentially the most important - question of all.

I cannot hope to adequately address it in the present treatment, but a few preliminary remarks may help to structure a space for subsequent discussion.

4.1 Agency and Experience

To what extent, if any, do these stories capture or explain facts about what we might think of as *personal* (or agent-level) cognition- the flow of thoughts, reasons, and ideas that characterize daily conscious thought and reason? A first (but fortunately merely superficial) impression is that they fall far short of illuminating personal level experience. For example, there seems to be a large disconnect between 'surprisal' (the implausibility of some sensory state given a model of the world – see 1.6 above) and agent-level surprise. This is evident from the simple fact that the percept that, overall, best minimizes surprisal (hence minimizes prediction errors) 'for' the brain may well be, for me the agent, some highly surprising and unexpected state of affairs – for example, the sight of a large pink rabbit (to borrow a striking image from John Haugeland) dancing in the middle of the room.

The two perspectives are, however, easily reconciled. The dancing pink rabbit is best understood as improbable but not (at least not in the relevant sense - recall 3.2 above) surprising. Instead, the 'pink rabbit percept' is the one that best respects what the system knows and expects about the world, given the current combination of driving inputs and assigned precision (reflecting the brain's degree of confidence in the signal). Given the right driving signal and a high enough assignment of precision, top-level theories of an initially agent-unexpected kind can still win out so as to explain away that highly-weighted tide of incoming sensory evidence. The sight of the dancing pink rabbit may then emerge as the least surprising (least 'surprisal-ing'!) percept available, given the inputs, the priors, and the current weighting on sensory prediction error

The broadly Bayesian framework can also seem at odds with the facts about conscious perceptual experience for a different reason. The world, it might be said, does not *look* as if it is encoded as an intertwined set of probability density distributions! It looks unitary and, on a clear day, unambiguous. But this phenomenology again poses no real challenge. What is on offer, after all, is a story about the brain's way of encoding information about the world. It is not directly a story about how things seem to agents deploying that means of encoding information. There is clearly no inconsistency in thinking that the brain's pervasive use of probabilistic encoding might yield conscious experiences that depict a single, unified and quite unambiguous scene. Moreover, in the context of an active world-engaging system, such an outcome

makes adaptive sense. For the only point of all that probabilistic betting is to drive action and decision, and action and decision lack the luxury of being able to keep all options indefinitely alive. It would do the evolved creature no good at all to keep experiencing the scene as to some degree uncertain if the current task requires a firm decision, and if its neural processing has already settled on a good, strongly supported bet as to what's (most probably) out there.

One way to begin to cash that out is to recall that biological systems will be informed by a variety of learnt or innate 'hyperpriors' concerning the general nature of the world. One such hyperprior, as remarked during the discussion of binocular rivalry in section 1.4, might be that there is only one object (one cause of sensory input) in one place, at a given scale, at a given moment[28]. Another – more germane to the present discussion - might be that the world is usually in one determinate state or another. To implement this, the brain might[29] simply use a form of probabilistic representation in which each distribution has a single peak (meaning that each overall sensory state has a single best explanation). This would rule out true perceptual ambiguity while leaving plenty of room for the kind of percept-switching seen in the binocular rivalry cases. The use of such a representational form would amount formally to the deployment of an implicit hyperprior (a 'prior about priors') to the effect that our uncertainty can be described using such a uni-modal probability distribution. Such a hyperprior makes adaptive sense, given the kinds of brute fact about action mentioned above (e.g. we can only perform one action at a time, choosing the left turn or the right but never both at once).

Such appeals to powerful (and often quite abstract) hyperpriors will clearly form an essential part of any larger, broadly Bayesian, story about the shape of human experience. Despite this, no special story needs to be told about either the very *presence* or the *mode of action* of such hyperpriors. Instead, they arise quite naturally within hierarchical models of the kind we have been considering where they may be innate (giving them an almost Kantian feel) or acquired in the manner of empirical (hierarchical) Bayes[30]. Nonetheless, the sheer potency of these highly abstract forms of 'systemic expectation' again raises questions about the eventual spread of explanatory weight: this time, between the framework on offer and whatever additional considerations and modes of investigation may be required to fix and reveal the contents of the hyperpriors themselves.[31]

4.2 Illuminating Experience: The Case of Delusions

It might be suggested that merely *accommodating* the range of human personal-level experiences is one thing, while truly *illuminating* them is another. Such positive impact is, however, at least on the horizon. We glimpse the potential in an impressive body of recent work conducted within the predictive processing (hierarchical predictive coding) framework addressing delusions and hallucination in schizophrenia (Fletcher and Frith (2009), Corlett, Frith, and Fletcher (2009)).

Recall the unexpected sighting of the 'dancing pink rabbit' discussed in the previous section. Here, the system already commanded an apt model able to 'explain away' the particular combination of driving inputs, expectations, and precision (weighting on prediction error) that specified the florid dancing bunny. But such is not always the case. Sometimes, dealing with ongoing highly-weighted sensory prediction error may require brand new generative models gradually to be formed (just as in normal learning). This might hold the key, as Fletcher and Frith (2009) suggest, to a better understanding of the origins of hallucinations and delusion (the two so-called 'positive symptoms') in schizophrenia. These two symptoms are often thought to involve two mechanisms and hence two breakdowns, one in 'perception' (leading to the hallucinations) and one in 'belief' (allowing these abnormal perceptions to impact top-level belief). It seems correct (see e.g. Coltheart (2007)) to stress that perceptual anomolies alone will not typically lead to the strange and exotic belief complexes found in delusional subjects. But must we therefore think of the perceptual and doxastic components as effectively independent?

A possible link emerges if perception and belief-formation, as the present story suggests, both involve the attempt to match unfolding sensory signals with top-down predictions. Importantly, the impact of such attempted matching is precision-mediated in that the systemic effects of residual prediction error vary according to the brain's confidence in the signal (2.3 above). With this in mind, Fletcher and Frith (2009) canvass the possible consequences of disturbances to a hierarchical Bayesian system such that prediction error signals are falsely generated and - more importantly - highly weighted (hence accorded undue salience for driving learning). There are a number of potential mechanisms whose complex interactions, once treated within the overarching framework of prediction error minimization, might conspire to produce such disturbances. Prominent contenders include the action of slow neuromodulators such as dopamine, serotonin, and acetylcholine (Corlett, Frith, and Fletcher (2009), Corlett, Taylor et al (2010)). In addition, Friston (2010, p.132) speculates that

fast synchronized activity between neural areas may also play a role in increasing the gain on prediction error within the synchronized populations[32]. The key idea, however implemented, is that understanding the positive symptoms of schizophrenia requires understanding disturbances in the generation and weighting of prediction error. The suggestion (Fletcher and Frith (2009), Corlett, Frith et al (2009), Corlett, Krystal et al (2009)) is that malfunctions within that complex economy (perhaps fundamentally rooted in abnormal dopaminergic functioning) yield wave upon wave of persistent and highly weighted 'false errors' that then propagate all the way up the hierarchy forcing, in severe cases (via the ensuing waves of neural plasticity) extremely deep revisions in our model of the world. The improbable (telepathy, conspiracy, persecution etc) then becomes the least surprising, and - since perception is itself conditioned by the top-down flow of prior expectations - the cascade of misinformation reaches back down, allowing false perceptions and bizarre beliefs to solidify into a coherent and mutually supportive cycle. The process is thus self-entrenching. As new generative models take hold, their influence flows back down so that incoming data is sculpted by the new (but now badly misinformed) priors so as to "conform to expectancies" (Fletcher and Frith (2009) p.348). False perceptions and bizarre beliefs thus form an epistemically insulated self-confirming cycle[33]. This, then, is the dark side of the seamless story (section 2 above) about perception and cognition. The predictive processing model merges - usually productively - perception, belief, learning, and affect into a single overarching economy: one within which dopamine and other neurotransmitters control the 'precision' (the weighting, hence the impact on inference and on learning) of prediction error itself. But when things go wrong, false inferences spiral and feed back upon themselves. Delusion and hallucination then become entrenched, being both co-determined and co-determining.

The same broadly Bayesian framework can be used (Corlett, Frith, et al (2009)) to help make sense of the way in which different drugs, when given to healthy volunteers, can temporarily mimic various forms of psychosis. Here too the key feature is the ability of the predictive coding framework to account for complex alterations in both learning and experience contingent upon the (pharmacologically modifiable) way driving sensory signals are meshed, courtesy of precision-weighted prediction errors, with prior expectancies and (hence) ongoing prediction. The psychotomimetic effects of ketamine, for example, are said to be explicable in terms of a disturbance to the prediction error signal (perhaps caused by AMPA upregulation) and the flow of prediction (perhaps via NMDA interference). This leads to a persistent prediction error and – crucially - an inflated sense of the importance or salience of the

associated events, which in turn drives the formation of short-lived delusion-like beliefs (see Corlett, Frith, et al (2009) p 6-7): also, discussion in Gerrans (2007)). The authors go on to offer accounts of the varying psychotomimetic effects of other drugs (such as LSD and other serotonergic hallucinogens, cannabis, and dopamine agonists such as amphetamine) as reflecting other possible varieties of disturbance within a hierarchical predictive processing framework[34].

This fluid spanning of levels constitutes, it seems to me, one of the key attractions of the present framework. We here move from considerations of normal and altered states of human experience, via computational models (highlighting prediction error based processing and the top-down deployment of generative models), to the implementing networks of synaptic currents, neural synchronies, and chemical balances in the brain. The hope is that by thus offering a new, multi-level, account of the complex, systematic interactions between inference, expectation, learning, and experience, these models may one day deliver a better understanding even of our own agent-level experience than that afforded by the basic framework of 'folk psychology'. Such an outcome would constitute a vindication of the claim (PM Churchland (1989) (In Press)) that adopting a 'neurocomputational perspective' might one day lead us to a deeper understanding of our own lived experience.

4.3 Perception, Imagery, and the Senses

Another area in which these models are suggestive of deep facts about the nature and construction of human experience concerns the character of perception and the relations between perception and imagery/visual imagination. Prediction-driven processing schemes, operating within hierarchical regimes of the kind described above, learn probabilistic generative models (within each neural population) targeting the activity patterns displayed in the neural population below. What is crucial here – what makes such models *generative* as we saw in 1.1 – is that they can be used 'top-down' to predict activation patterns in the level below. The practical upshot is that such systems, simply as part and parcel of learning to perceive, develop the ability to self-generate[35] perception-like states from the top down, by driving the lower populations into the predicted patterns.

There thus emerges a rather deep connection between perception and the potential for self-generated forms of mental imagery (Kossyln et al (1995), Reddy et al (2010)). Probabilistic generative model-based systems that can learn to visually perceive a cat (say) are, ipso facto, systems that can deploy a top-

down cascade to bring about many of the activity patterns that would ensue in the visual presence of an actual cat. Such systems thus display (for more discussion of this issue, see Clark (forthcoming)) a deep duality of perception and imagination[36]. The same duality is highlighted by Grush (2004) in the 'emulator theory of representation', a rich and detailed treatment that shares a number of key features with the predictive processing story.[37]

Hierarchical predictive processing also provides a mechanism that explains a variety of important phenomena that characterize sensory perception, such as cross- and multi-modal context effects on early sensory processing. Murrray et al (2002) displayed (as noted in 3.1 above) the influence of high-level shape information on the responses of cells in early visual area V1. Smith and Muckli (2010) show similar effects (using as input partially occluded natural scenes) even on wholly non-stimulated (i.e. not directly stimulated via the driving sensory signal) visual areas. Murray et al (2006) showed that activation in V1 is influenced by a top-down size illusion, while Muckli et al (2005) and Muckli (2010) report activity relating to an apparent motion illusion in V1. Even apparently 'unimodal' early responses are influenced (Kriegstein and Giraud (2006)) by information derived from other modalities, and hence commonly reflect a variety of multimodal associations. Even the expectation that a relevant input will turn out to be in one modality (e.g. auditory) rather than another (e.g. visual) turns out to impact performance, presumably by enhancing "the weight of bottom-up input for perceptual inference on a given sensory channel" (Langner et al (2011) p.10).

This whole avalanche of context effects emerges naturally given the hierarchical predictive processing model. If so-called visual, tactile, or auditory sensory cortex is actually exploiting a cascade of downward influence from higher levels whose goal is actively to predict the unfolding sensory signals (the ones originally transduced using the various dedicated receptor banks of vision, sound, touch, etc) extensive downward-reaching multi-modal and cross-modal effects (including various kinds of 'filling-in') will follow. For any statistically valid correlations, registered within the increasingly information-integrating (or 'metamodal' - Pascual-Leone and Hamilton (2001), Reich et al (2011)) areas towards the top of the processing hierarchy, can inform the predictions that cascade down, through what were previously thought of as much more unimodal areas, all the way to areas closer to the sensory peripheries. Such effects appear inconsistent with the idea of V1 as a site for simple, stimulus-driven, bottom-up feature-detection using cells with fixed (context-inflexible) receptive fields. But they are fully accommodated by models that depict V1

activity as constantly negotiated on the basis of a flexible combination of top-down predictions and driving sensory signal.

But then why, given this unifying model in which the senses work together to provide ongoing 'feedback' on top-down predictions that aim to track causal structure in the world, do we experience sight as different from sound, touch as different from smell, and so on? Why, that is, do we not simply experience the *overall best-estimated external states of affairs* without any sense of the structure of distinct modalities in operation as we do so?

This is a surprisingly difficult question, and any answer must remain tentative in advance of a mature scientific story about conscious experience itself. A place to start, though, is by noticing that despite the use of a single general processing strategy (the use of top-down predictions to attempt to explain away sensory prediction error), there remain important differences between what is being 'explained away' within the different modalities. This is probably best appreciated from the overarching perspective of Bayesian perceptual inference. Thus vision, haptics, taste, and audition each trade in sensory signals captured by distinct transducers and routed via distinct early processing pathways. The different sensory systems then combine priors and driving signals in ways that may yield *differing* estimates even of the very same distal state. It is true that the overall job of the perceptual system is to combine these multiple estimates into a single unified model of the distal scene. But different sensory systems specialize (unless one is pressed into unusual service, as in the interesting case of sensory substitution technologies[38]) in estimating different environmental features, and even where they estimate the same feature, their estimates, and the reliability (in context) of those estimates will vary. In a thick fog, for example, vision is unreliable (delivering shape information with high uncertainty) while touch is less affected, while when wearing thick gloves the reverse may be true. That means that even where two senses are reporting on the very same environmental state (e.g. shape by sight, and shape by touch) they may deliver different 'guesses' about what is out there: guesses that reflect inferences made on the basis of distinct priors, different sensory signals, and the differing uncertainties associated with those signals. These differences, it seems to me, should be enough to ground the obvious experiential differences between the various modalities, while the operation of a common underlying processing strategy (Bayesian inference, here implemented using hierarchical predictive coding) accounts for the ease with which multiple conflicting estimates are usually reconciled into a unified percept. In this way the framework on offer provides a powerful set of 'fundamental cognitive particles' (generative models and precision-weighted prediction-error driven processing)

whose varying manifestations may yet capture both the variety and the hidden common structure of our mental lives.

Difficult questions also remain concerning the best way to connect an understanding of such 'fundamental particles' and the gross structure of our daily (and by now massively culturally underwritten) conception of our own mental lives. In this daily or 'folk' conception, we rather firmly distinguish between perceptions, thoughts, emotions, and reasons, populating our minds with distinct constructs such as memories, beliefs, hopes, fears, and (agent-level) expectations. We thus depict minds and selves in ways that are likely to make at best indirect contact (see e.g. Dennett (1978) (1987), Clark (1989), Barrett (2009)) with the emerging scientific vision. Yet bridging between these visions (the manifest and the scientific image – Sellars (1962)) remains essential if we are to gain maximal benefits from a better understanding of the inner (and outer) machinery itself. It is essential if, for example, we aspire to deploy our new understandings to improve social relations and education, to increase human happiness, or to inform our responses to social problems. To bridge this gap will plausibly require effort and compromise from both sides (Humphrey (2000)), as the folk conception alters under the influence of a scientific understanding that must itself recognize the causal potency of the daily constructs: constructs which we encounter and model just as surely as we encounter and model marriage, divorce, and taxes.

4.4 Sensing and World

What, then, of the mind-world relation itself? Hohwy (2007) suggests that:

> One important and, probably, unfashionable thing that this theory tells us about the mind is that perception is indirect…[…]…what we perceive is the brain's best hypothesis, as embodied in a high-level generative model, about the causes in the outer world. (Hohwy (2007) p.322)

There is something right about this. The bulk of our daily perceptual contact with the world, if these models are on the mark, is determined as much by our expectations concerning the sensed scene as by the driving signals themselves. Even more strikingly, the forward flow of sensory information consists only in the propagation of error signals, while richly contentful predictions flow downwards, interacting in complex non-linear fashions via the web of reciprocal connections. One result of this pattern of influence is a greater efficiency in the use of neural encodings, since:

"an expected event does not need to be explicitly represented or communicated to higher cortical areas which have processed all of its relevant features prior to its occurrence" (Bubic at al (2010) p. 10)

If this is indeed the case, then the role of perceptual contact with the world is only to check and when necessary correct the brain's best guessing as to what is out there. This is a challenging vision, since it suggests that our expectations are in some important sense the primary source of all the contents of our perceptions, even though such contents are constantly being checked, nuanced, and selected by the prediction error signals consequent upon the driving sensory input[39]. Perhaps surprisingly, the immediate role of the impinging world is thus most marked when error signals, in a well-functioning brain, drive the kinds of plasticity that results in perceptual learning, rather than in the cases where we are simply successfully engaging a well-understood domain.

Nonetheless, we may still reject the bald claim that "what we perceive is the brain's best hypothesis". For even if our own prediction is indeed (at least in familiar, highly learnt contexts) doing much of the heavy lifting, it remains correct to say that *what* we perceive is not some internal representation or hypothesis but (precisely) the world. We do so courtesy of the brain's ability to latch on to how the world is by means of a complex flow of sub-personal processes. That flow, if these stories are on track, fully warrants the 'Helmholtzian' description of perception as inference. But it is precisely by such means that biological beings are able to establish a truly tight mind-world linkage. Brains like these are statistical sponges structured (1.2 above) by individual learning and evolutionary inheritance so as to reflect and register relevant aspects of the causal structure of the world itself[40].

One place where this becomes especially evident is in the treatment (2.2 above) of visual illusions as Bayes-optimal percepts. The idea, recall, is that the percept - even in the case of various effects and illusions - is an accurate estimation of the most likely real-world source or property given noisy sensory evidence and the statistical distribution, within some relevant sample, of real-world causes. This is an important finding that has now been repeated in many domains, including the sound-induced flash illusion (Shams and Beierholm (2005)), ventriloquism effects (Alais and Burr (2004)) and the impact of figure-ground convexity cues in depth perception (Burge et al (2010)). Additionally, Weiss et al's (2002) Bayes-optimal account of a class of static (fixation-dependent) motion illusions has now been extended to account for a much wider set of motion illusions generated in the presence of active eye movements during smooth pursuit (see Freeman et al (2010), and discussion in Ernst (2010)).

Perceptual experience, even in these illusory cases, thus looks to be veridically tracking statistical relations between the sensory data and its most probable real-world sources. The intervening mechanisms thus introduce no some worrisome barrier between mind and world. Rather, it is only *because* of such sub-personal complexities that agents like us can be perceptually open to the world itself [41].

5. Taking Stock

5.1 Comparison with Standard Computationalism

Just how radical is the story we have been asked to consider? Is it best seen as an alternative to mainstream computational accounts that posit a cascade of increasingly complex feature detection (perhaps with some top-down biasing), or is it merely a supplement to them: one whose main virtue lies in its ability to highlight the crucial role of prediction error in driving learning and response? I do not think we are yet in a position to answer this question with any authority. But the picture I have painted suggests an intermediate verdict, at least with respect to the central issues concerning representation and processing.

Concerning representation, the stories on offer are potentially radical in at least two respects. First, they suggest that probabilistic generative models underlie both sensory classification and motor response. And second, they suggest that the forward flow of sensory data is replaced by the forward flow of prediction error. This latter aspect can, however, make the models seem even more radical than they actually are. For recall that the forward flow of prediction error is here combined with a downward flow of predictions, and at every stage of processing the models posit (as we saw in some detail in 2.1) functionally distinct 'error units' and 'representation units'. The representation units that communicate predictions downwards do indeed encode increasingly complex and more abstract features (capturing context and regularities at ever-larger spatial and temporal scales) in the layers and processing furthest removed from the raw sensory input. In a very real sense then, much of the standard architecture of increasingly complex feature detection is here retained. What differs is the shape of the flow of information, and (relatedly) the pivotal role assigned to the computation and propagation of prediction error.

A related issue concerns the extent to which the 'new' framework reproduces traditional insights concerning the specialization of different cortical areas. This is a large question whose full resolution remains beyond the scope of the present discussion. But in general, the hierarchical form of these models

suggests a delicate combination of specialization (different layers learn and deploy different generative models aimed at the layer below) and integration (the system settles in a way largely determined by the flow and weighting of prediction error, where this flow is itself varied according to current context and the reliability and relevance of different types of information).[42]

A second source of potential radicalism lies with the suggestion (section 1.5 above) that, in extending the models to include action ('action-oriented predictive processing') we might simultaneously do away with the need to appeal to goals and rewards, replacing them with the more austere construct of predictions. In this vein we read that:

> "Crucially, active inference does not invoke any 'desired consequences'. It rests only on experience-dependent learning and inference: experience induces prior expectations, which guide perceptual inference and action" Friston, Mattout, and Kilner (2011) p. 157

In this desert landscape vision, there are neither goals nor reward signals as such. Instead, there are only (both learnt and species-specific) expectations, across many spatial and temporal scales, which directly enslave both perception and action. Cost functions, in other words, are replaced expectations concerning actions and their sensory (especially proprioceptive) consequences. Here, I remain unconvinced. For even if such an austere description is indeed possible (and for some critical concerns, see Gershman and Daw (In Press)), that would not immediately justify us in claiming that it thereby constitutes the better tool for understanding the rich organization of the cognitive economy. To see this, we need only reflect that it's all 'just' atoms, molecules and the laws of physics too, but that doesn't mean those provide the best constructs and components for the systemic descriptions attempted by cognitive science. The desert landscape theorist thus needs to do more, it seems to me, to demonstrate the explanatory advantages of abandoning more traditional appeals to value, reward, and cost (or perhaps to show that those appeals make unrealistic demands on processing or implementation – see Friston (2011a))

What may well be right about the desert landscape story, it seems to me, is the suggestion that utility (or more generally, personal and hedonic value) is not simply a kind of add-on, implemented by what Gershman and Daw (2011) describe as a 'segregated representation of probability and utility in the brain'. Instead, it seems likely that we represent the very events over which probabilities become defined in ways that ultimately fold in their personal, affective, and hedonic significance. This folding in is probably especially

marked in frontolimbic cortex (Merker (2004)). But the potent web of backward connections ensures that such folding in, once it has occurred, is able (as noted by Barrett and Bar (2009) – see also section 2.2 above) to impact processing and representation at every lower stage of the complex processing hierarchy. If this proves correct, then it is prediction error calculated relative to these affectively rich and personal-history-laden expectations that drives learning and response.

Thus construed, an action-oriented predictive processing framework is not so much revolutionary as reassuringly integrative. Its greatest value lies in suggesting a set of deep unifying principles for understanding multiple aspects of neural function and organization. It does this by describing an architecture capable of combining high-level knowledge and low-level (sensory) information in ways that systematically deal with uncertainty, ambiguity, and noise. In so doing it reveals perception, action, learning, and attention as different but complementary means to the reduction of (potentially affect-laden and goal-reflecting) prediction error in our exchanges with the world. It also, and simultaneously, displays human learning as sensitively responsive to the deep statistical structures present in both our natural and human-built environments. Thus understood, action-oriented predictive processing leaves much *unspecified*, including (1) the initial variety of neural and bodily structures (and perhaps internal representational forms) mandated by our unique evolutionary trajectory and (2) the acquired variety of 'virtual' neural structures and representational forms installed by our massive immersion in 'designer environments' during learning and development.

To fill in these details requires, or so I have argued, a deep (but satisfyingly natural) engagement with evolutionary, embodied, and situated approaches. Within that context, seeing how perception, action, learning, and attention might all be constructed out of the same base materials (prediction and prediction error minimization) is powerful and illuminating. It is there that Friston's ambitious synthesis is at its most suggestive, and it is there that we locate the most substantial empirical commitments of the account. Those commitments are to the computation (by dedicated error units or some functionally equivalent means) and widespread use by the nervous system of precision-weighted prediction error, and its use as proxy for the forward flow of sensory information. The more widespread this is, the greater the empirical bite of the story. If it doesn't occur, or occurs only in a few special circumstances, the story fails as a distinctive empirical account[43].

5.2 Conclusions: Towards A Grand Unified Theory of the Mind?

Action-oriented predictive processing models come tantalizing close to overcoming some of the major obstacles blocking previous attempts to ground a unified science of mind, brain, and action. They take familiar elements from existing, well-understood, computational approaches (such as unsupervised and self-supervised forms of learning using recurrent neural network architectures, and the use of probabilistic generative models for perception and action) and relate them on the one hand to a priori constraints on rational response (the Bayesian dimension) and, on the other hand, to plausible and (increasingly) testable accounts of neural implementation. It is this potent positioning between the rational, the computational, and the neural that is their most attractive feature. In some ways, they provide the germ of an answer to Marr's dream: a systematic approach that addresses the levels of (in the vocabulary of Marr (1982)) the computation, the algorithm, and the implementation.

The sheer breadth of application is striking. Essentially the same models here account for a variety of superficially disparate effects spanning perception, action, and attention. Indeed, one way to think about the primary 'added value' of these models is that they bring perception, action, and attention into a single unifying framework. They thus constitute the perfect explanatory partner, I have argued, for recent approaches that stress the embodied, environmentally embedded, dimensions of mind and reason[44]. Perception, action, and attention, if these views are correct, are all in the same family business: that of reducing sensory prediction error resulting from our exchanges with the environment. Once this basic family business is revealed, longer-term environmental structuring (both material and socio-cultural) falls neatly into place. We structure our worlds and actions so that most of our sensory predictions come true.

But this neatness hides important complexity. For another effect of all that material and socio-cultural scaffolding is to induce substantial path-dependence as we confront new problems using pre-existing material tools and inherited social structures. The upshot, or so I have argued, is that a full account of human cognition cannot hope to 'jump' directly from the basic organizing principles of action-oriented predictive processing to an account of the full (and in some ways idiosyncratic) shape of human thought and reason. What emerges instead is a kind of natural alliance. The basic organizing principles highlighted by action-oriented predictive processing make us superbly sensitive to the structure and statistics of the training environment. But our human training environments are now so thoroughly artificial, and our explicit forms

of reasoning so deeply infected by various forms of external symbolic scaffolding, that understanding distinctively human cognition demands a multiply hybrid approach. Such an approach would combine the deep computational insights coming from probabilistic generative approaches (among which figure action-oriented predictive processing) with solid neuroscientific conjecture *and* with a full appreciation of the way our many self-structured environments alter and transform the problem spaces of human reason. The most pressing practical questions thus concern what might be thought of as the "distribution of explanatory weight" between the accounts on offer, and approaches that explore or uncover these more idiosyncratic or evolutionary path-dependent features of the human mind, and the complex transformative effects of the socio-cultural cocoon in which it develops.

Questions also remain concerning the proper scope of the basic predictive processing account itself. Can that account really illuminate reason, imagination, and action-selection in all its diversity? What do the local approximations to Bayesian reasoning look like as we depart further and further from the safe shores of basic perception and motor control? What new forms of representation are then required, and how do they behave in the context of the hierarchical predictive coding regime? How confident are we of the basic Bayesian gloss on our actual processing? (Do we, for example, have a firm enough grip on when a system is computing its outputs using a 'genuine approximation' to a true Bayesian scheme, rather than merely behaving 'as if' it did so?).

The challenges (empirical, conceptual, and methodological) are many and profound. But the potential payoff is huge. What is on offer is a multi-level account of some of the deepest natural principles underlying learning and inference, and one that may be capable of bringing perception, action, and attention under a single umbrella. The ensuing exchanges between neuroscience, computational theorizing, psychology, philosophy, rational decision theory, and embodied cognitive science promise to be among the major intellectual events of the early 21st Century.

Thanks and Acknowledgements.

References

Adams, F. and Aizawa, K. (2001). The Bounds of Cognition. *Philosophical Psychology* 14:1: 43-64.

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

Alais, D, and Burr, D. (2004) The ventriloquist effect results from near-optimal bimodal integration. *Current Biology* 14:257–262.

Alink, A., Schwiedrzik, C.M., Kohler,A., Singer, W.,and Muckli, L. (2010) Stimulus Predictability Reduces Responses in Primary Visual Cortex *J. Neurosci.* 30: 2960-2966

Anderson, M (2007) The massive redeployment hypothesis and the functional topography of the brain. *Philosophical Psychology* 20:2:143-174

Arthur, B (1994) *Increasing Returns and Path Dependence in the Economy*. University of Michigan Press, Ann Arbor.

Bar, M. (2007). The Proactive Brain: Using analogies and associations to generate predictions. *Trends in Cognitive Sciences*, 11(7), 280-289.

Barrett, L. F. (2009). The future of psychology: Connecting mind to brain. *Perspectives in Psychological Science*, 4, 326-339.

Barrett, L. F., & Bar, M. (2009). See it with feeling: Affective predictions in the human brain. *Royal Society Phil Trans* B, 364, 1325-1334

Bengio, Y (2009) Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–127

Berniker M and Körding KP. (2008) Estimating the sources of motor errors for adaptation and generalization *Nature Neuroscience*. 11, 1454 – 1461

Born, R. T., Tsui, J. M. and Pack, C. C. (2009) Temporal dynamics of motion integration, in *Dynamics of Visual Motion Processing*. Ilg, U. and Masson, G. eds. (New York: Springer)

Brainard, D. (2009) "Bayesian Approaches to Color Vision." In *The Visual Neurosciences*, 4th ed., ed. M. Gazzaniga. Cambridge, MA: MIT Press.

Brown H, Friston K, and Bestamnn S. (2011) Active inference, attention and motor preparation. *Frontiers in Psychology* 2: 218 (doi: 10.3389/fpsyg.2011.00218)

Bubic A, von Cramon DY and Schubotz RI (2010) Prediction, cognition and the brain. *Front. Hum. Neurosci.* 4:25: 1-15

Burge,J., Fowlkes, C.,and Banks, M. (2010) Natural-Scene Statistics Predict How the Figure–Ground Cue of Convexity Affects Human Depth Perception *Journal of Neuroscience*, 30(21):7269-7280

Chater, N., and Manning, C. (2006) Probabalistic models of language processing and acquisition *Trends in Cognitive Sciences* 10:7: 335-344

Churchland, P. M. (1989). *The Neurocomputational Perspective.* Cambridge: MIT/Bradford Books.

Churchland, P.M (In Press) *Plato's Camera: Howe the Physical Brain Captures a Landscape of Abstract Universals* (MIT Press, Cambridge, MA)

Clark, A (1987) The Kludge in the Machine *Mind And Language* 2: 4: 277-300

Clark, A (1989) *Microcognition: Philosophy, Cognitive Science and Parallel Distributed Processing,* MIT Press/Bradford Books, Cambridge: MA)

Clark, A (1997) *Being There: Putting Brain, Body and World Together Again* (MIT Press, Camb. MA).

Clark, A (2006) Language, Embodiment and the Cognitive Niche *Trends in Cognitive Sciences* 10:8:370-374

Clark, A (2008) *Supersizing the Mind: Action, Embodiment, and Cognitive Extension* (Oxford University Press, NY)

Clark, A (forthcoming) Perceiving as Predicting, in M. Mohan, S. Biggs, and D. Stokes (eds) *Perception and Its Modalities* (Oxford University Press, NY)

Clark, A and Chalmers, D. (1998). The Extended Mind. *Analysis* 58:1:7-19

Clifford, C.W.G., Webster M.A., Stanley, G.B., Stocker, A.A., Kohn, A., Sharpee, T.O. & Schwartz, O. (2007). Visual adaptation: neural, psychological and computational aspects, *Vision Research*, 47, 3125-3131.

Coltheart, M. (2007). "Cognitive neuropsychiatry and delusional belief" (The 33rd Sir Frederick Bartlett Lecture), *The Quarterly Journal of Experimental Psychology*, 60 (8): 1041–1062.

Corlett PR, Frith CD, and Fletcher PC (2009) From drugs to deprivation: a Bayesian framework for understanding models of psychosis. *Psychopharmacology (Berl)* 206:4: p.515-30

Corlett P. R., Krystal J. K., Taylor J. R., Fletcher P. C. (2009). Why do delusions persist? *Front. Hum. Neurosci.* 3, 12. doi: 10.3389/neuro.09.012.2009

Corlett PR, Taylor JR, Wang XJ, Fletcher PC, and Krystal JH (2010) Toward a neurobiology of delusions. *Progress In Neurobiology*. 92: 3 p.345-369

Dayan, P (1997) Recognition in hierarchical models. In F Cucker & M Shub, editors, *Foundations of Computational Mathematics*. Berlin, Germany: Springer.

Dayan, P and Hinton, G (1996) Varieties of Helmholtz machine. *Neural Networks*, 9, 1385-1403

Dayan, P., Hinton, G. E., & Neal, R. M. (1995). The Helmholtz machine. *Neural Computation*, 7, 889–904.

Deheane, S (2009) *Reading in the brain*. New York: Penguin

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society series B*, 39:1-38.

Deneve, S. (2008) Bayesian spiking neurons I: inference. *Neural Comput*. **20:** 91–117.

de-Wit,L., Machilsen,B., and Putzeys, T. (2010) Predictive Coding and the Neural Response to Predictable Stimuli *J. Neurosci*. 30: 8702-8703

den Ouden, H.E.M.,Daunizeau, J., Roiser, J., Friston, K., and Stephan, K (2010) Striatal Prediction Error Modulates Cortical Coupling *J. Neurosci*. 30: 3210-3219.

Dennett, D. (1978). *Brainstorms: Philosophical Essays on Mind and Psychology*. Cambridge: Bradford Books/MIT Press

Dennett, D. (1987). *The Intentional Stance*. Cambridge, MA, MIT Press.

Dennett, D (1991) *Consciousness Explained* (Little Brown, Boston)

Desimone R, Duncan J (1995), "Neural mechanisms of selective visual attention." *Annu Rev Neurosci* 18:193-222

Di Paolo, E. (2009). Extended life. *Topoi*, 28(1), 9-21

Doya K, Ishii S, Pouget A, Rao R. (eds) (2007) *Bayesian Brain: Probabilistic*

Approaches to Neural Coding. Cambridge, MA: MIT Press

Dumoulin SO, and Hess RF (2006) Modulation of V1 activity by shape: image-statistics or shape-based perception? *Journal of Neurophysiology*. 95: 3654–3664.

Egner, T., Monti, J. M., Summerfield, C. (2010). Expectation and surprise determine neural population responses in the ventral visual stream. *Journal of Neuroscience,* 30(49): 16601-16608.

Eliasmith, C. (2007). How to build a brain: From function to implementation. *Synthese.* 153(3): 373-388

Engel, A. K., Fries, P., & Singer, W. (2001). Dynamic predictions: Oscillations and synchrony in top-down processing. *Nature Reviews*, 2, 704-716.

Ernst, M. O. (2010) Eye Movements: Illusions in Slow Motion. *Current Biology* 20(8), R357-R359

Ernst, M. O. and M. S. Banks (2002) Humans Integrate Visual and Haptic Information in a Statistically Optimal Fashion. *Nature* 415: 429-433

Fabre-Thorpe M (2011) The characteristics and limits of rapid visual categorization. *Frontiers in Psychology* 2:243. doi: 10.3389/fpsyg.2011.00243

Feldman H and Friston K (2010) Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience* 2: 4 article 215 (doi: 10.3389/fnhum.2010.00215)

Feldman, J (2010) Cognitive Science should be unified: comment on Griffiths et al. and McClelland et al. *Trends Cogn. Sci.* 14:8: 341

Fletcher, P and Frith, C (2009) Perceiving is believing: a Bayesian appraoch to explaining the positive symptoms of schizophrenia. *Nature Reviews: Neuroscience* 10: 48-58

Freeman, T.C.A., Champion, R.A., & Warren, P.A. (2010). A Bayesian model of perceived head-centred velocity during smooth pursuit eye movement. *Current Biology*, 20, 757-762

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

Friston K. (2002) Beyond phrenology: what can neuroimaging tell us about distributed circuitry? *Annual Rev Neurosci.* 25:p. 221-50

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

Friston K. (2005). A theory of cortical responses. *Philos Trans R Soc Lond B Biol Sci.*29;360(1456):815-36.

Friston K. (2009). The free-energy principle: a rough guide to the brain? *Trends Cogn Sci.* 13: 293–301

Friston K. (2010) The free-energy principle: a unified brain theory? *Nat Rev Neurosci.* 11(2):127-38.

Friston, K (2011a) What is optimal about motor control? *Neuron* 72: 488-498

Friston K (2011b) Embodied Inference: or "I think therefore I am, if I am what I think". In Wolfgang Tschacher and Claudia Bergomi (eds) *The Implications of Embodiment (Cognition and Communication)* Imprint Academic 89-125

Friston KJ, Daunizeau J, and Kiebel SJ. (2009) Reinforcement learning or active inference? *PLoS (Public Library of Science) One.* 2009 Jul 29;4(7): e6421

Friston KJ, Daunizeau J, Kilner J, Kiebel SJ. (2011). Action and behavior: a free-energy formulation. *Biol Cybern.* Feb 2011. [Epub ahead of print]

Friston. K and Kiebel, S (2009) Cortical circuits for perceptual inference. *Neural Networks*, 22:1093-1104

Friston K, Mattout J, and Kilner J. (2011) Action understanding and active inference. *Biol Cybern.* 104:137–160

Friston, K., & Stephan, K. (2007). Free energy and the brain. *Synthese*, 159(3), 417–458.

Gerrans, P (2007) Mechanisms of Madness. Evolutionary Psychiatry Without Evolutionary Psychology. *Biology and Philosophy*. 22, 35-56. 2007.

Gershman, S.J. & Daw, N.D. (in press). Perception, action and utility: the

tangled skein. In M. Rabinovich, M., K. Friston & P. Varona, Eds, *Principles of Brain Dynamics: Global State Interactions*. MIT Press.

Gold, J.N. and Shadlen, M.N. (2001) Neural computations that underlie decisions about sensory stimuli. *Trends Cogn. Sci.* 5, 10:16 238 – 255

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

Griffiths, T., Chater, N., Kemp, C., Perfors, A., and Tenenbaum, J.B. (2010). Probabilistic models of cognition: exploring representations and inductive biases. *Trends Cogn. Sci.* 14: 8: 357–364

Griffiths. P.E & Gray, R.D (2001) Darwinism and Developmental Systems in Oyama, S, Griffiths. P.E & Gray, R.D (eds) *Cycles of Contingency: Developmental systems and evolution* Cambridge, Mass., MIT Press: 195-218.

Grill-Spector K, Henson R, Martin A. (2006) Repetition and the brain: Neural models of stimulus-specific effects. *Trends in Cognitive Sci.* 10(1): 14 – 23

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

Harnad, S. (1990) The Symbol Grounding Problem. *Physica D* 42: 335-346.

Haugeland, J (1998) Mind Embodied and Embedded. in J. Haugeland *Having Thought: Essays in the Metaphysics of Mind* (Harvard University Press, Cambridge, MA) 207-240

Hawkins, J and Blakeslee, S (2004) *On Intelligence* (Owl Books, NY)

Helbig, H., and Ernst, M. (2007). Optimal Integration of Shape Information from Vision and Touch. *Experimental Brain Research* 179: pp. 595-605.

Helmholtz, H. (1860/1962). *Handbuch der physiologischen optik* (Southall, J. P. C. (Ed.), English trans.),Vol. 3. New York: Dover.

Hesselmann, G., Sadaghiani, S., Friston, K. J., Kleinschmidt, A. (2010) Predictive coding or evidence accumulation? False inference and neuronal fluctuations. *PloS One,* 5(3), e9926.

Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8), 1711-1800.

Hinton, G. E. (2007a) Learning Multiple Layers of Representation. *Trends in Cognitive Sciences*, 11, 428-434.

Hinton, G. E. (2007b) To recognize shapes, first learn to generate images. In P. Cisek, T. Drew and J. Kalaska (Eds.) *Computational Neuroscience: Theoretical Insights into Brain Function*. Elsevier, Holland

Hinton, G. E. (2010) Learning to represent visual input. *Philosophical Transactions of the Royal Society, B.* Vol 365, pp 177-184.

Hinton, GE, Dayan, P, Frey, BJ & Neal, RM (1995). The wake-sleep algorithm for unsupervised neural networks. *Science*, 268, 1158-1160.

Hinton, G.E. and Ghahramani, Z. (1997) Generative Models for Discovering Sparse Distributed Representations *Philosophical Transactions Royal Society B*, 352:1177-1190

Hinton, G. E., Osindero, S. and Teh, Y. (2006) A fast learning algorithm for deep belief nets. *Neural Computation*, 18, pp 1527-1554

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the Dimensionality of Data with Neural Networks. *Science*, 313(5786), 504-507

Hinton GE, and von Camp D (1993) Keeping neural networks simple by minimizing the description length of weights. In: *Proceedings of COLT-93*, pp 5–13

Hinton GE & Zemel RS (1994). Autoencoders, minimum description length and Helmholtz free energy. In J. Cowan, G. Tesauro, and J. Alspector (Eds.), *Advances in Neural Information Processing Systems* 6, San Mateo, CA: Morgan Kaufmann.

Hochstein, S., and Ahissar, M. (2002) View from the top: hierarchies and reverse hierarchies in the visual system. *Neuron* 36(5):791-804.

Hohwy, J. (2007). Functional Integration and the mind *Synthese* 159:3: 315-328

Hohwy, J., Roepstorff, A., Friston, K. (2008). Predictive coding explains

binocular rivalry: an epistemological review. *Cognition* 108 (3): 687-701.

Hollerman, J. R. & Schultz, W. Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neurosci.* **1**, 304–309 (1998).

Hosoya, T., Baccus, S.A., and Meister, M. (2005) Dynamic predictive coding by the retina. *Nature* 436:7: 71-77

Howe CQ, Lotto RB, Purves D (2006) Comparison of bayesian and empirical ranking approaches to visual perception. *J Theor Biol* 241: 866–875

Huang, Y and Rao, R (2011) Predictive Coding *Wiley Interdisciplinary Reviews: Cognitive Science*, 2: 580–593

Humphrey, N (2000) How to solve the mind-body problem *Journal of Consciousness Studies*, 7, 5-20

Hurley, S. (1998) *Consciousness in Action.* (Cambridge, MA: Harvard)

Hutchins, E (1995) *Cognition In The Wild* (MIT Press: Camb. MA)

Jehee JFM, Ballard DH. (2009). Predictive Feedback Can Account for Biphasic Responses in the Lateral Geniculate Nucleus. *PLoS Comput Biol* 5(5): e1000373.

Kawato, M, Hayakama, H & Inui, T (1993). A forward-inverse optics model of reciprocal connections between visual cortical areas. *Network*, 4, 415-422.

Knill, D., and Pouget, A. (2004) The Bayesian brain: the role of uncertainty in neural coding and computation *Trends in Neuroscience.* 27(12):712-9

Kohonen, T. (1989) *Self-organization and Associative Memory* (Springer-Verlag, Berlin)

König, P und Krüger, N (2006) Perspectives: Symbols as self-emergent entities in an optimization process of feature extraction and predictions. *Biological Cybernetics* 94(4): 325-334.

Körding, KP., Tenenbaum JB., and Shadmehr R. (2007) The dynamics of memory as a consequence of optimal adaptation to a changing body. *Nature Neuroscience*, 10:779-786

Kosslyn, S.M., Thompson, W.L., Kim, I.J., and Alpert, N.M. (1995) Topographical representations of mental images in primary visual cortex. *Nature* 378, 496–498.

Kriegstein, K., and Giraud, A. (2006) Implicit Multisensory Associations Influence Voice Recognition *PLoS Biology* 4:10: e326

Kveraga, K., Ghuman, A., and Bar, M. (2007) Top-down predictions in the cognitive brain. *Brain and Cognition* 65: 145-168

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211-240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998) Introduction to Latent Semantic Analysis. *Discourse Processes* 25: 259-284.

Langner, R., Kellermann, T., Boers, F., Sturm,W., Willmes, K., and Eickhoff, S.B. (2011) Modality-Specific Perceptual Expectations Selectively Modulate Baseline Activity in Auditory, Somatosensory, and Visual Cortices *Cerebral Cortex* (advance access e-publication doi:10.1093/cercor/bhr083)

Lee, M (2010) Emergent and structured cognition in Bayesian models: comment on Griffiths et al. and McClelland et al. *Trends Cogn. Sci.* 14: 8: 345-6

Lee, S.H., Blake,R., & Heeger,D.J.(2005).Traveling waves of activity in primary visual cortex during binocular rivalry. *Nature Neuroscience* 8, 2.

Lee, T.S., and Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of Optical Society of America, A.* . 20(7): 1434-1448.

Lehnert, W (2007) Cognition, Computers, and Car Bombs: How Yale Prepared Me for the 90's in Schank, R and Langer, E. *Beliefs, Reasoning, and Decision Making: Psycho-logic in Honor of Bob Abelson* (Lawrence Erlbaum Associates, Hillsdale, NJ. pp. 143-173).

Leopold,D.,& Logothetis,N.(1999) Multistable phenomena: Changing views in perception.*Trends in Cognitive Sciences* 3: 254–264.

Linsker, R. (1989). An application of the principle of maximum information preservation to linear systems. In D. S. Touretzky (Ed.). *Advances in neural information processing systems* (Vol. 1, pp. 86–194). Berlin: Springer.

MacKay DJC (1995) Free-energy minimization algorithm for decoding and cryptoanalysis. *Electron Lett* 31:445–447

MacKay, DM (1956). The epistemological problem for automata. In CE Shannon & J McCarthy, editors, *Automata Studies,* Princeton, NJ: Princeton University Press, 235-251.

Maher, B. (1988) Anomalous experience and delusional thinking: The logic of explanations. In T.F. Oltmanns and B.A. Maher (Eds.) *Delusional beliefs* Chichester:Wiley. p.15–33.

Maloney, L. T. & Mamassian, P. (2009), Bayesian decision theory as a model of visual perception: Testing Bayesian transfer. *Visual Neuroscience*, 26, 147-155

Maloney, L. T. & Zhang, H. (In Press), Decision-theoretic models of visual perception and action. *Vision Research*

Mamassian, P., Landy, M., and Maloney, L. (2002) Bayesian Modeling of Visual Perception. In R. Rao, B. Olshausen, and M. Lewicki. (eds) *Probabilistic Models of the Brain*, Cambridge, MA: MIT Press.

Marcus, G (2008) *Kluge: The Haphazard Construction of the Human Mind* (Houghton-Mifflin, Boston)

Mareschal D, Johnson M, Sirois S, Spratling M, Thomas M, Westermann G (2007). *Neuroconstructivism - I: How the Brain Constructs Cognition.* Oxford University Press

Marr, D. (1982), *Vision: A Computational Approach*, San Francisco, Freeman & Co.

McClelland, J., Botvinick, M., Noelle, D., Plaut, D., Rogers, T., Seidenberg, M., and Smith, L. (2010). Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends Cogn. Sci.* 14: 8: 348–356.

McClelland, J. & Rumelhart, D. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of Basic Findings.

Psychological Review, 88, 375-407

McClelland, J., Rumelhart, D., and the PDP Research Group (1986), *Parallel Distributed Processing*, Volume II, Cambridge, Mass.: MIT Press.

Melloni, L., Schwiedrzik, CM., Muller, N., Rodriguez, E., and Singer, W (2011) Expectations Change the Signatures and Timing of Electrophysiological Correlates of Perceptual Awareness *Journal Of Neuroscience* 31: 4: p1386-p1396

Menary, R (2007) *Cognitive Integration: Attacking The Bounds of Cognition.* Palgrave Macmillan

Meng, M. and Tong, F. (2004) Can attention selectively bias bistable perception? differences between binocular rivalry and ambiguous figures. *Journal of Vision*, 4. 539-551

Merker, B. (2004) Cortex, countercurrent context, and dimensional integration of lifetime memory. *Cortex, 40:* 559-576.

Milner, D. and Goodale, M. (2006) *The Visual Brain in Action: second edition* Oxford: Oxford University Press.

Muckli, L (2010) What Are We Missing Here? Brain Imaging Evidence for Higher Cognitive Functions in Primary Visual Cortex V1 *Int J Imaging Syst Techno (IJIST)* 20: 131-139

Muckli L, Kohler A, Kriegeskorte N, Singer W (2005) Primary visual cortex activity along the apparent-motion trace reflects illusory perception. *PLoSBio* l3:e265.

Mumford D (1992) On the computational architecture of the neocortex. II. The role of cortico-cortical loops. *Biol Cybern* 66:241–251.

Mumford, D (1994). Neuronal architectures for pattern-theoretic problems. In C. Koch and J. Davis, editors, *Large-Scale Theories of the Cortex.* Cambridge, MA: MIT Press, 125-152.

Murray, S.O., Boyaci, H., and Kersten, D. (2006) The representation of perceived angular size in human primary visual cortex. *Nature Reviews: Neuroscience* 9: 429–434.

Murray S.O., Kersten D., Olshausen B.A., Schrater P., Woods D.L.(2002) Shape perception reduces activity in human primary visual cortex. *Proc. Natl. Acad Sci. U.S.A.* 99(23):15164–15169

Murray, S. O., Schrater, P., & Kersten, D. (2004). Perceptual grouping and the interactions between visual cortical areas. *Neural Networks*, 17(5-6), 695-705

Musmann, H. (1979) Predictive Image Coding, in W. K. Pratt (Ed.), *Image Transmission Techniques,* Academic Press, New York

Neal, R.M. and Hinton, G. (1998) A view of the EM algorithm that justifies incremental, sparse, and other variants, In Jordan, M.I. (ed) *Learning in Graphical Models*, Dordrecht: Kluwer p. 355-368

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

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

Olshausen, BA & Field, DJ (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381, 607-609.

Overy K., & Molnar-Szakacs I. (2009). Being Together in Time: Musical Experience and the Mirror Neuron System. *Music Perception*, vol. 26 (5), 489-504.

Oyama, S. (1999) *Evolution's Eye: Biology, Culture and Developmental Systems*, Duke University Press, Durham, North Carolina.

Pack CC and Born RT. (2001) Temporal dynamics of a neural solution to the aperture problem in visual area MT of macaque brain. *Nature* 409: 1040 – 1042

Pascual-Leone, A., and Hamilton, R. (2001). The metamodal organization of the brain. *Progress in Brain Research.* 134, 427–445.

Pfeifer, R., Lungarella, M., Sporns, O., Kuniyoshi, Y.(2007) On the information theoretic implications of embodiment - principles and methods. *Lecture Notes in Computer Science (LNCS)* vol. 4850 (Springer, Heidelberg).

Pickering M.J. & Garrod S. (2007) Do people use language production to make

predictions during comprehension? *Trends in Cognitive Sciences* (11) pp 105-110

Pouget, A., Dayan, P., and Zemel, R. (2003) Inference and computation with population codes. *Annual Review of Neuroscience* 26, 381-410.

Prinz, J.J. (2005). A neurofunctional theory of consciousness. In Andy Brook and Kathleen Akins (eds) *Cognition and the brain: Philosophy and neuroscience movement* (Cambridge: Cambridge University Press) p. 381-396.

Purves D, Lotto RB (2003) *Why we see what we do: An empirical theory of vision.* (Sunderland, MA: Sinauer Associates)

Rao, R and Ballard, D. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects, *Nature Neuroscience* 2, 1, 79

Rao, R. and Sejnowski, T. (2002) Predictive Coding, Cortical Feedback, and Spike-Timing Dependent Cortical Plasticity in Rao, Olshausen, and Lewicki (eds) *Probabilistic Models of the Brain* (MIT Press, Camb. MA).

Rauss, Schwartz, and Pourtois (in press) Top-down effects on early visual processing in humans: a predictive coding framework *Neuroscience and Biobehavioral Reviews*

Reddy, L., Tsuchiya, N. & Serre, T., (2010). Reading the mind's eye: decoding category information during mental imagery. *NeuroImage*, 50(2), p.818-825

Reich, L., Szwed, M., Cohen, L., and Amedi, A. (2011) A ventral stream reading center independent of visual experience *Current Biology* 21, 363-368

Rescorla, M (In Press) Bayesian Perceptual Psychology to appear in the *Oxford Handbook of the Philosophy of Perception* (ed, M. Matthen) CHECK

Rieke, F (1999) *Spikes: Exploring the Neural Code* (MIT Press, Camburdge, MA)

Robbins, H. (1956) An Empirical Bayes Approach to Statistics. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics* p. 157–163.

Roepstorff, Niewöhner, and Beck (2010) Enculturating brains through patterned practices *Neural Networks* 23: 1051-1059

Rowlands, M (1999) *The Body in Mind: Understanding Cognitive Processes* (Cambridge University Press, Cambridge, UK)

Rumelhart, D., McClelland, J., and the PDP Research Group (1986) *Parallel Distributed Processing*, vol. I, Cambridge, Mass.: MIT Press.

Schwartz, O, Hsu, A & Dayan, P (2007) Space and time in visual context *Nature Reviews Neuroscience* 8 522-535.

Sellars, W (1962) Philosophy and the Scientific Image of Man. In *Frontiers of Science and Philosophy*, (ed) Robert Colodny (Pittsburgh: University of Pittsburgh Press) p. 35-78

Shams, L, Ma, WJ, & Beierholm, U. (2005). Sound-induced flash illusion as an optimal percept *Neuroreport*, 16(10), 1107-10

Shi, Yun Q and Sun, H. (1999) *Image and Video Compression for Multimedia Engineering: Fundamentals, Algorithms, and Standards.* CRC Press

Sloman, A (1990) Must Intelligent Systems be Scruffy? In J.E.Tiles, G.T.McKee, G.C.Dean (eds) *Evolving Knowledge in Natural Science and Artificial Intelligence* (London: Pitman, 1990)

Smith, F., and Muckli, L. (2010) Nonstimulated early visual areas carry information about surrounding context *Proceedings of the National Academy of Science (PNAS)* early edition (in advance of print)

Smith, L and Gasser, M (2005) The Development of Embodied Cognition: Six Lesson from Babies *Artificial Life* 11:1:13-30

Smith, P.L., & Ratcliff, R. (2004). Psychology and neurobiology of simple decisions. *Trends in Neuroscience*, 27, 161-168.

Sporns, O. (2007) What neuro-robotic models can teach us about neural and cognitive development. In Mareschal D, Sirois S, Westermann G, and Johnson M.H (eds) *Neuroconstructivism: perspectives and prospects Vol. 2* 179–204. Oxford, UK: Oxford University Press

Spratling, M. (2008) Predictive coding as a model of biased competition in visual attention. *Vision Research*, 48(12):1391-408

Spratling, M. and Johnson, M. (2006) A feedback model of perceptual learning and categorisation. *Visual Cognition*, 13(2):129-65

Srinivasan M.V., Laughlin S.B. and Dubs A. (1982) Predictive coding: a fresh view of inhibition in the retina. *Proc. R. Soc. Lond. B* 216, 427-459

Sterelny K (2003) *Thought in a Hostile World: The Evolution of Human Cognition* (Blackwell, Oxford)

Sterelny, K (2007) Social Intelligence, Human Intelligence and Niche Construction *Proceedings of the Royal Society, London, (Series B)*, 362, 1480, pp 719-730

Stotz, K. (2010). Human Nature and Cognitive–Developmental Niche Construction. *Phenomenology and the Cognitive Sciences* 9 (4) 483-501

Summerfield, C. and Egner, T. (2009) Expectation (and attention) in visual cognition. *Trends Cogn. Sci.* 13, 403–409

Summerfield C, Trittschuh EH, Monti JM, Mesulam MM, Egner T. (2008). Neural repetition suppression reflects fulfilled perceptual expectations. *Nature Neuroscience.* 11:9:1004-1006.

Thelen, E and Smith, L (1994) *A Dynamic Systems Approach To The Development Of Cognition And Action* (MIT Press, Camb. MA)

Thompson, E (2007) *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*, Harvard University Press, Harvard, USA

Todorov, E., & Jordan, M. I. (2002). Optimal feedback control as a theory of motor coordination. *Nature Neuroscience*, 5, 1226–1235

Todorov, E (In Press) Parallels between sensory and motor information processing. *The Cognitive Neurosciences*, 4th ed, Gazzaniga (ed), MIT Press, Camb. MA

Toussaint, M "probabilistic inference as a model of planned behavior" *Künstliche Intelligenz* (In Press)

Tribus, M. (1961) *Thermodynamics and Thermostatics: An Introduction to Energy, Information and States of Matter, with Engineering Applications.* D. Van Nostrand Company Inc., New York

Varela, F., Thompson, E., and Rosch.,E (1991) *The Embodied Mind* (MIT Press, Camb. MA)

Velleman, J. D. (1989) *Practical Reflection*, Princeton: Princeton University Press.

Verschure, P., Voegtlin, T., and Douglas, R. (2003) "Environmentally mediated synergy between perception and behaviour in mobile robots" *Nature* 425, 620–624

Vilares I, and Körding K. (2011) Bayesian models: the structure of the world, uncertainty, behavior, and the brain. *Annals of the New York Acad Sci.* 1224: 22-39

Waelti, P., Dickinson, A. & Schultz, W. (2001) Dopamine responses comply with basic assumptions of formal learning theory. *Nature* **412**, 43–48.

Weiss Y., Simoncelli E.P. and Adelson E.H. (2002) Motion Illusions as Optimal Percepts *Nature Neuroscience* 5: 6: 598 – 604

Wheeler, M (2005) *Reconstructing the Cognitive World* (MIT Press, Camb. MA)

Wilson, R. A. (1994) Wide Computationalism *Mind* 103:351-372

Wilson, R. A. (2004) *Boundaries of the Mind: The Individual in the Fragile Sciences--Cognition* (Cambridge University Press, Cambridge, UK

Yu, A J. (2007) Adaptive behavior: Humans act as Bayesian learners. *Current Biology* 17: R977-R980.

Yuille and Kersten (2006) Vision as Bayesian inference: analysis by synthesis? *Trends in Cognitive Science*: 10: 7: 301-308

Zahedi, K., Ay, N., and Der, R (In Press) Higher coordination with less control - a result of information maximization in the sensori-motor loop. *Adaptive Behavior*

Notes

[1] This remark is simply described as a 'scribbled, undated, aphorism' in the online digital archive of the scientist's journal: see http://www.rossashby.info/index.html

[2] I am greatly indebted to an anonymous BBS referee for encouraging me to bring these key developments into clearer (both historical and conceptual) focus.

[3] The obvious problem was that this generative model itself needed to be learnt: something that would in turn be possible if a good recognition model was already in place, since that could provide the right targets for learning the generative model. The solution (Hinton et al (1995)) was to use each to gradually bootstrap the other, using the so-called 'wake-sleep algorithm' - a computationally tractable approximation to 'maximum likelihood learning' as seen in the expectation-maximization (EM) algorithm of Dempster et al (1977). Despite this, the Helmholtz Machine remained slow and unwieldy when confronted with complex problems requiring multiple layers of processing. But it represents an important early version of an unsupervised multi-layer learning device or 'deep architecture' (Hinton (2002), Hinton, G. E., & Salakhutdinov, R. R. (2006), Hinton et al (2006), Hinton (2007b), Hinton (2010) – for reviews, see Hinton (2007a) and Bengio (2009)).

[4] This names the probability of an event (here, a worldly cause) given some set of prior beliefs and the evidence (here, the current pattern of sensory stimulation). For our purposes, it thus names the probability of a worldly (or bodily) cause, conditioned on the sensory consequences.

[5] In speaking of 'predictive processing' rather than resting with the more common usage 'predictive coding', I mean to highlight the fact that what distinguishes the target approaches is not simply the use of the data compression stategy known as predictive coding. Rather, it is the use of that strategy in the special context of hierarchical systems deploying probablisitic generative models. Such systems exhibit powerful forms of learning and are able flexibly to combine top-down and bottom-up flows of information within a multi-layer cascade.

[6] Since these proposals involves the deployment of top-down probabilistic generative models within a multi-layer architecture, it is the organizational structure of the neocortex that most plausibly provides the requisite implementation. This is not to rule out related modes of processing using other structures, for example in non-human animals, but simply to isolate the 'best fit'. Nor is it to rule out the possibility that, moment-to-moment, details of the large-scale routing of information flow within the brain might depend on gating effects that, although cortically mediated, implicate additional structures and areas. For some work on such gating effects among cortical structures themselves, see den Ouden et al (2010).

[7] I have adopted the neuroanatomist practice of labeling connections simply as 'backward' and 'forward' so as to avoid the functional implications of the labels 'feedback' and 'feedforward'. This is important in the context of predictive processing models, since it is now the forward connections that are really providing (by conveying prediction error) feedback on the downwards-flowing predictions - see Friston (2005), Hohwy (2007), and discussion in section 2.5 following). Thanks to one of the BBS reviewers for this helpful terminological suggestion.

[8] Hosoya et al here build on earlier work by Srinivasan et al (1982) – see also information-theoretic treatments of mutual information, such as Linsker (1989). For a larger perspective, see Clifford et al (2007)

[9] What about more common forms of perceptual alternation, such as those induced by ambiguous figures like the Necker cube or the duck-rabbit? In these instances, the gross driving sensory input is exactly the same for the two percepts, so switching cannot be induced simply by the ongoing influence of the unexplained portions of bottom-up input. Instead, such cases are best explained by a similar process involving attentional modulations (which may, but need not, be deliberate). Attention (see section 2.3 following) serves to increase the gain on select error units. By altering the gain on some error units and not others, the impact of the driving sensory signal is effectively altered so that the best interpretation flips. Attention thus engages the same (broadly Bayesian) mechanism, but via a different (and potentially less automatic) route. This also explains, within the present framework, why we have much more control over the alternation rate in the case of ambiguous figures (as demonstrated by Meng and Tong (2004)).

[10] This is also known (see e.g. Friston, Daunizeau, and Kiebel (2009)) as 'active inference'. I coin 'action-oriented predictive processing' as it makes clear that this is an action-encompassing generalization of the (hierarchical) predictive coding story about perception. It also suggests (rightly) that action becomes conceptually primary in these accounts, since it provides the only way (once a good world model is in place and aptly activated) to actually alter the sensory signal so as to reduce sensory prediction error – see Friston (2009) pp.295. In addition Friston's most recent work on active inference looks to involve a strong commitment (see especially Friston (2011a)) to the wholesale replacement of value functions, considered as determinants of action, with expectations ('prior beliefs', though note that 'belief' here is very broadly construed) about action. This is an interesting and challenging suggestion that goes beyond claims concerning formal equivalence and even beyond the observations concerning deep conceptual relations linking action and perception. 'Action-oriented predictive processing' as I shall use the term, remains deliberately agnostic on this important matter (see also 5.1 following).

[11] I note in passing that this radical view resonates with some influential philosophical work concerning high level (reflective) intentions and actions: specifically, Velleman's (1989) account of practical reasoning in which intentions to act are depicted as self-fulfilling expectations about one's own actions – see e.g. Velleman (1989) p. 98.

[12] The most fundamental aspect of the appeal to free energy, Friston claims, is that it provides an organismically-computable window on surprise itself, since "...surprise cannot be quantified by an agent, whereas free energy can." (Friston (2010) p.55). I read this as meaning, in the present context, that prediction error is organismically-computable, since it represents (as we saw in 1.2 above) an internally-calculable quantity. This, however, is not a feature that I will attempt to explore in the present treatment.

[13] For an interesting critique of the most ambitious version of the free energy story, see Gershman and Daw (In Press) section 5.1.

[14] De-Wit (personal correspondence) notes that his usage follows that of e.g. Murray et al (2004) and Dumoulin and Hess (2005), both of whom contrast 'predictive coding' with 'efficient coding', where the former uses top-down influence to subtract out predicted elements of lower-level activity, and the latter uses top-down influence to enhance or sharpen it. This can certainly make it look as if the two stories (subtraction and sharpening) offer competing accounts of, for example, fMRI data such as Murray et al (2002) showing a dampening of response in early visual areas as higher areas settled into an interpretation of a shape stimulus. The accounts would be alternatives since the dampening might then reflect *either* the subtraction of well- predicted parts of the early response ('predictive coding') or the quashing of the rest of the early signal and the attendant sharpening of the consistent elements. The models I am considering, however, accommodate both subtraction *and* sharpening (see main text for details). This is therefore an instance (see 5.1 following) in which more radical elements of the target proposals (here, the subtracting away of predicted signal elements) turn out , on closer examination, to be consistent with more familiar effects (such as top-down enhancement).

[15] The consistency between selective sharpening and the dampening effects of 'explaining away' also makes it harder – though not impossible - to tease apart the empirical implications of predictive coding and 'evidence accumulation' accounts such as Gold and Shadlen (2001) – for a review, see Smith and Ratcliff (2004). For an attempt to do so, see Hesselmann et al (2010).

[16] In this (2008) treatment Spratling further argues that the forms of hierarchical predictive coding account we have been considering are mathematically equivalent to some forms of 'biased competition' model, but that they nonetheless suggest different claims concerning neural implementation. I take no position on these interesting claims here.

[17] Thanks to one of the BBS reviewers for suggesting this important nuance to the temporal story

[18] This means that we need to be very careful when generalizing from ecologically strange laboratory conditions that effectively deprive us of such ongoing context. For some recent discussion, see Kveriga et al (2007), Bar (2007), Barrett and Bar (2009), Fabre-Thorpe (2011).

[19] An interesting alternative to the inference-rich Bayesian account is suggested by Purves and Lotto (2003) who offer a more direct account in terms of the bare statistics of image-

source relationships – For a comparison with Bayesian approaches, see Howe, Lotto, and Purves (2006).

[20] I here adapt, merely for brevity of exposition, a similar example from Friston (2002) p. 237

[21] The clear lineage here is with work in connectionism and recurrent artificial neural networks (see for example Rumelhart, McClelland et al (1986) and early discussions such as Churchland (1989), Clark (1989)). What is most exciting about the new proposals, it seems to me, is that they retain many of the insights from this lineage (which goes on to embrace work on Helmholz machines and ongoing work on 'deep architectures' – see 1.1 above) while making explicit contact with both Bayesian theorizing and contemporary neuroscientific research and conjecture.

[22] For an excellent discussion of this recent work, see de-Wit, Machilsen, and Putzeys (2010)

[23] Lee de-Wit (personal communication) raises the intriguing possibility that the distinction between encoding error and encoding representational content might be realized in alternate dynamics of the very same neuronal substrate, with early responses encoding error and later ones settling into a representation of something like 'agreed content'. In a related vein, Engel et al (2001) discuss the potential role of neural synchrony as a means of implementing top-down influence on early processing.

[24] These terms, according to a memoir by Wendy Lehnert (2007), were introduced by Bob Abelson as part of a keynote address to the *3rd Annual Meeting of the Cognitive Science Society* in 1981.

[25] The hierarchical predictive coding family of models that (along with their extensions to action) form the main focus of the present treatment are not, in my view, happily assimilated to either of these camps. They clearly share Bayesian foundations with the 'pure' structured probabilistic approaches highlighted by Griffiths et al, but their computational roots lie (as we saw in 1.1) in work on machine learning using artificial neural networks. Importantly, however, hierarchical predictive processing models now bring 'bottom-up' insights from cognitive neuroscience into increasingly productive contact with those powerful computational mechanisms of learning and inference, in a unifying framework able (as Griffiths et al correctly stress) to accommodate a very wide variety of surface representational forms. Moreover, such approaches are computationally tractable since local (prediction error minimizing) routines are being used to approximate Bayesian inference. For some excellent antidotes to the appearance of deep and irreconcilable conflict hereabouts, see Feldman (2010), Lee (2010).

[26] We glimpse the power of the complex internal statistical relationships enshrined in human languages in Landauer and colleagues ((Landauer et al (1998), Landauer and Dumais (1997)) fascinating work on 'latent semantic analysis'. This work reveals the vast amount of information now embodied in statistical (but deep, not first order) relations between words and the larger contexts (sentences and texts) in which they occur. The symbolic world we

humans now immerse ourselves in is demonstrably chock-full full of information about meaning-relations in itself, even before we (or our brains) attempt to hook any of it to practical actions and the sensory world.

[27] E.g. Stanislas Dehaene's (2009) 'neural re-cycling' account of the complex interplay between neural precursors, cultural developments, and neural effects within the key cognitive domains of reading and writing.

[28] Such hyperpriors could, for example, be 'built-in' by 'winner-takes-all' forms of lateral (within layer) cortical inhibition - see Hohwy et al (2008) p.691.

[29] As helpfully pointed out by one of the BBS referees.

[30] The introduction of hyperpriors into these accounts is just a convenient way of gesturing at the increasing levels of abstraction at which prior expectations may be pitched. Some expectations, for example, may concern the reliability or shape of the space of expectations itself. In that sense, hyperpriors, although they can sound quite exotic, are in no way ad hoc additions to the account. Rather, they are just priors in good standing (but maintaining the distinction makes it a bit easier to express and compute some things). Like all priors, they then impact system dynamics in various ways, according to their specific contents.

[31] This worry (concerning the appeal to hyperpriors) was first drawn to my attention by Mark Sprevak (personal communication).

[32] A much better understanding of such multiple interacting mechanisms (various slow neuromodulators perhaps acting in complex concert with neural synchronization) is now needed, along with a thorough examination of the various ways and levels at which the flow of prediction and the modulating effects of the weighting of prediction error (precision) may be manifest (for some early forays, see Corlett, Taylor, et al (2010) – see also Friston and Kiebel (2009)). Understanding more about the ways and levels at which the flow and impact of prediction error may be manipulated is vitally important if we are to achieve a better understanding of the multiple ways in which 'attention' (here understood – see 2.3 above - as various ways of modifying the gain on prediction error) may operate so as to bias processing by flexibly controlling the balance between top-down and bottom-up influence.

[33] There are probably milder versions of this everywhere, both in science (Maher (1988)) and everyday life. We tend to see what we expect, and we use that to confirm the model that is both generating our expectations and sculpting and filtering our observations.

[34] Intriguingly, the authors are also able to apply the model to one non-pharmacological intervention: sensory deprivation.

[35] This need not imply an ability deliberately to engage in such a process of self-generation. Such rich, deliberate forms of imagining may well require additional resources, such as the language-driven forms of cognitive 'self-stimulation' described in Dennett (1991) chapter 8.

[36] It is perhaps worth remarking that, deep duality notwithstanding, nothing in the present view requires that the system, when engaged in imagery-based processing, will typically support the very same kinds of stability and richness of experienced detail that daily sensory engagements offer. In the absence of the driving sensory signal, no stable ongoing information about low-level perceptual details is there to constrain the processing. As a result, there is no obvious pressure to *maintain* or perhaps even to generate (see Reddy et al (2010)) a stable hypothesis at the lower levels: there is simply whatever task-determined downwards pressure the active higher-level encoding exerts.

[37] Common features include the appeal to forward models and the provision of mechanisms (such as Kalman filtering – see Grush (2004), Friston (2002), Rao and Ballard (1999)) for estimating uncertainty and (thus) flexibly balancing the influence of prior expectations and driving sensory inputs. Indeed, Grush (op cit p. 393) cites the seminal predictive coding work by Rao and Ballard (1999) as an account of visual processing compatible with the broader emulator framework. In addition, Grush's account of perception as 'environmental emulation' (Grush (2004) section 5.2) looks highly congruent with the depiction (Friston (2003) and elsewhere) of perception as reconstructing the hidden causes structuring the sensory signal. Where the accounts seem to differ is in the emphasis placed on prediction error as (essentially) a replacement for the sensory signal itself, the prominance of a strong Bayesian interpretation (using the resources of 'empirical Bayes' applied across a hierarchy of processing stages), and the attempted replacement of motor commands by top-down proprioceptive predictions alone (for a nice treatment of this rather challenging speculation, see Friston (2011a)). It would be interesting (although beyond the scope of the present treatment) to attempt a more detailed comparison.

[38] An account of such transformed uses might be possible within the action-oriented predictive coding framework. The key to such an account would, I conjecture, be to consider the potential of the substituting technlogies to deliver patterns of sensory stimulation that turn out to be best predicted by the use of the very same intermediate-level generative models that characterize the substituted modality. See also Prinz (2005).

[39] Thanks to Susanna Siegel for useful discussion of this point.

[40] For some further discussion, see Friston (2005) p.822.

[41] This way of describing things was suggested by my colleague Matt Nudds (personal communication).

[42] For the general story about combining specialization and integration, see Friston (2002) and discussion in Hohwy (2007). For a more recent account, including some experimental evidence concerning the possible role of prediction error in modulating inter-area coupling, see den Ouden et al (2010).

The empirical bet is thus, as Egner et al recently put it, that "the encoding of predictions (based on internal forward models) and prediction errors may be a ubiquitous feature of cognition in the brain…rather than a curiosity of reward learning…or motor planning" (Egner et al (2010 p. 16607)).

[44] When brought under the even-more-encompassing umbrella of the 'free energy principle' (1.6 above), the combined ambition is formidable. If these accounts were indeed to mesh in the way Friston (2010) suggests, that would reveal the very deepest of links between life and mind, confirming and extending the perspective known as 'enactivist' cognitive science – see e.g. Varela, Thompson, and Rosch (1991), Thompson (2007), Di Paolo (2009)).