Perceiving as Predicting

Perceiving as Predicting

Andy Clark,
School of Philosophy, Psychology, and Language Sciences,
University of Edinburgh, UK

Abstract

According to an emerging vision in computational cognitive neuroscience, perception (rich, full-blooded, world-presenting perception of the kind we humans enjoy) depends heavily on prediction. To visually perceive, if this schema is correct, is to meet incoming visual information with a set of matching top-down predictions that track the evolving visual signal across multiple spatial and temporal scales. In this chapter I first introduce this general explanatory schema. I then display some recent evidence for the schema, and discuss to what extent it marks a radical departure from previous ('feature-detection based') models of perception. I end by exploring some implications of the schema for questions concerning multi-modal and cross-modal effects in sensory processing, and for our understanding of the deep and fundamental relations between perception, imagination, and understanding.

1. Perceiving as Predicting

1.1 Predictive Coding

A familiar view depicts perception as essentially a process of 'bottom-up' feature detection. Thus, in the case of vision, detected colors, edges, and shapes might act as the building blocks for detected objects (cats, dogs) and states of affairs (dogs chasing a cat, perhaps). Scientific versions of the paradigm depict early perception as building towards a complex world model by a feedforward process of evidence accumulation. Visual cortex, to take the most-studied example, is thus "traditionally viewed as a hierarchy of neural feature detectors, with neural population responses being driven by bottom-up stimulus features" (Egner et al 2010 p. 16601). This is a view of the perceiving brain as passive and stimulus-driven, taking energetic inputs from the senses and turning them into a coherent percept by a kind of step-wise build-up moving from the simplest features to the more complex. From pixel intensities up to lines and edges and on to complex meaningful shapes (like teacups), accumulating structure and complexity along the way in a kind of Lego-block fashion. In

these models[i] then "sensory processing was considered to consist mainly of the sequential extraction and recombination of features, leading to the veridical reconstruction of object properties" (Engel, Fries, and Singer 2001 p.704).

"Predictive coding" – the main topic of the present treatment - works by a kind of reversal of such passive evidence accumulation schemes. In these models (see Rao and Ballard (1999), Lee and Mumford (2003), Friston (2005) (2010)) percepts emerge via a recurrent cascade of predictions that involve (mostly sub-personal) expectations, spanning multiple spatial and temporal scales, about the present nature and state of the world as presented via the driving sensory signal. That driving sensory signal is compared to the predictions, and mismatches send forward error signals that nuance or alter the prediction until a match is found and the sensory data is 'explained away'. This process runs concurrently and continuously (until it settles) across multiple levels of a processing hierarchy.

At first sight, this seems extremely implausible. How can perception, a process that surely puts us in contact with the world, be a matter of prediction? Doesn't this mistake perception for (something more like) imagination? And anyway, how can we issue a prediction unless we already know a good deal about what's out there?

To see how the predictive coding alternative works, it helps to start by noticing that the key predictions made by the brain concern not what is about to happen but what is already the case. Specifically, the predictions made by the brain concern the current states of some of its own neural populations. In perception, if these models are correct, each layer of neural processing is trying to predict the current input to the layer below (except for the bottom layer, such as the retina, which 'simply' transduces an energetic signal[ii]). Each layer does this while simultaneously responding to predictions from the layer above. The key task of the brain (or at any rate, the cortex) is thus to learn a stack of models that capture regularities in how the sensory signal is most likely to vary in time and space. By deploying the right models at the right time, the brain can then issue correct predictions (so it is minimizing its own prediction errors). This (as Hohwy (2007) also notes) induces a striking reversal in which the driving sensory signal is really providing corrective *feedback* on the top-down predictions. Friston expresses the point well:

> "In this view, cortical hierarchies are trying to generate sensory data from high-level causes. This means the causal structure of the world is embodied in the backward connections. Forward connections simply

> provide feedback by conveying prediction error to higher levels. In short, forward connections are the feedback connections. This is why we have been careful not to ascribe a functional label like feedback to backward connections." Friston (2005) p.825

It is the forward flow of error that must now carry any new information coming from the world, allowing new predictive models to be selected and deployed in the top-down cascade. The upshot is that:

> "In predictive coding schemes, sensory data are replaced by prediction error, because that is the only sensory information that has yet to be explained" Feldman and Friston (2010) p.2

Each layer in these systems thus displays two functionally distinct properties. It encodes how it takes the world to be, and it registers mismatches between those 'takings' and predictions coming from the layer above. Mismatches flow forward as error signals to the level above, while its best guesses about the state of the world flow downwards as predictions to the layer below. Perception occurs when, across multiple layers of such processing that capture regularities at many spatial and temporal scales, the hugely interanimated set of predictions match the evolving sensory inputs, explaining them away so that the forward flow of error ceases or settles.

Importantly, such models can be acquired by learning and that learning can *itself* be driven by the ongoing attempt to minimize errors in the multilayer prediction of inputs. This is because the brain's predictions improve when it uses a good model of the structured signal source. Good predictions thus increase the posterior probability of the model. In this way the attempt to predict can be used to drive the learning itself, generating the very models that are then used to predict. In this pleasingly boot-strappy way ('empirical Bayes') a multi-layer system can acquire its own priors (the expectations used in prediction) from the data, as it goes along.

Thus consider, as a simple early example, Rao and Ballard's (1999) model of predictive coding in the visual cortex. Rao and Ballard implemented a multilayer neural network whose input was samples (image patches) from pictures of natural scenes. These visual signals were processed via a hierarchical system in which each level tried (in the way just sketched) to predict the activity at the level below it using recurrent (feedback) connections. If the feedback

successfully predicted the lower level activity, no further action needed to ensue. Failures to predict enabled tuning and revision of the generative model (initially, just a random set of connection weights) being used to make the predictions, thus slowly delivering knowledge of the regularities governing the data. The forward connections between levels carried only the 'residual errors' (Rao and Ballard (1999) p.79) between top-down predictions and actual lower level activity, while the backward or recurrent connections carried the predictions themselves. Changing prediction thus corresponds to changing or tuning a hypothesis about the nature and temporal evolution of the lower level activity. This kind of prediction error calculation, operating within a hierarchical organization, allowed information pertaining to different spatial and temporal scales within the image to be played off one against the other such that:

> "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

After exposure to thousands of image patches, the system had learnt to use responses in the first level network to extract features such as oriented edges and bars, while the second level network captured combinations of such features corresponding to patterns involving larger spatial configurations. Using the predictive coding strategy, and given only the statistical properties of the signals derived from the natural images, the network was thus able to induce a simple multi-layered model of the structure of the data source (images of natural scenes).

Notice that as processing in this network unfolds, all that is passed forward from level 1 to level 2 is error (the deviations from the predictions being sent downwards from level 2), and all that is passed downward is prediction. When downward prediction fully accommodates ('cancels out') the incoming signal, no more error flows forward and we perceive the world. The simulation also neatly captured well-documented 'non-classical receptive field' effects such as 'end-stopping' (see also Rao and Sejnowski (2002)) where a neuron responds strongly to a short line falling within its classical receptive field but that response tails off as the line gets longer. The predictive coding explanation is that the response tails off as the line gets longer because longer lines and edges are the statistical norm in these natural scenes! So, after learning, longer lines are what is first predicted (and fed back, as a hypothesis) by the level two network. The strong firing of those level 1 'edge cells' when driven by shorter

lines thus reflects error or mismatch: the unusually short segment was not initially predicted by the higher-level network. This means that end-stopped cells may be learnt, and reflect the way the world is – they reflect the typical length of the lines and edges in natural scenes. In a different world, such cells would learn different responses.

1.2 Generative Models

An important feature of the internal models that power such 'predictive coding' approaches is that they are *generative* in nature. That is to say, the knowledge (model) encoded by the units and weights at an upper layer must be such that those units and weights are capable of predicting the response profile of the layer below. That means that the model at layer N+1 becomes capable of generating the sensory data (i.e. the input as it would there be represented) at layer N (the layer below) for itself. Since this story applies all the way down to layers that are attempting to predict the inputs at the lowest level (the level of sensory transduction) that means that such systems are fully capable of generating 'virtual' sensory data for themselves.

This is, in one sense, unsurprising. As Hinton (and for similar comments see Mumford (1992)) notes:

> "Vivid visual imagery, dreaming, and the disambiguating effect of context on the interpretation of local image regions suggests that the visual system can perform top-down generation" Hinton (2007b) p. 428

In another sense it is surprising. It means that perception (at least, as it occurs in creatures like us[iii]) is co-emergent with (something quite like) imagination. And it means too – or so I suggest – that no creature is truly able to perceive anything that, in principle, they cannot imagine. Otherwise put, any creature that can perceive some state of affairs X also has the resources endogenously to generate a kind of 'virtual' version of that percept via top-down means alone. Of course, this does not mean that they can, by some deliberate act of will, bring this about. Indeed, it seems very likely that for some creatures acts of deliberate imagining (which I suspect may require the use of self-cueing via language) are simply impossible. But if these models are correct, then any creature able to perceive some state of affairs X has the neural resources to

generate the very same sensory states (where that means the ones that would occur were X to be veridically detected) in the absence of X[iv]. Whether that generation induces a conscious experience of a quasi-sensory nature is a question I here leave open. But at the very least, there now emerges a deep duality between perception and the endogenous generation of 'virtual' sensory data.

The use of such 'generative models' for perception and recognition is increasingly dominant in both theoretical and applied work on machine learning. It is no accident that early explorations of these themes involved items with names such as the 'wake-sleep algorithm' (Hinton et al (1995)) and talk of the network generating patterns for itself 'in fantasy'. In all these models top-down connections are generative ones, capable of causing (generating) the very same kinds of patterns of lower-level activity that would ensue given apt sensory (bottom-up) input. A more recent example can be found in Hinton's 2007a treatment, aptly titled "To Recognize Shapes, First Learn to Generate Images". Here, instead of attempting to directly train a neural network to classify images, the network first learns to generate such images for itself. An important achievement of Hinton and colleagues (see Hinton et al (2006), and the review papers Hinton (2007b), (2010), Bengio (2009)) is to show how to learn, using unlabelled (i.e. not pre-classified) data a deep multi-layer version of such a generative model[v]. This was an important advance over previous 'connectionist' work (Rumelhart et al (1986)) that struggled to learn appropriate representations in a deep multi-layer context, and that required large bodies of pre-classified data to power learning using the back-propagation of error.

There are, however, important differences separating the recent work by Hinton and colleagues (using so-called Restricted Boltzmann Machines – see Hinton (2007)) and the predictive coding story (see the review by Huang (2011)). The differences mostly concern the kinds of message passing scheme that are, and are not, allowed, and the precise ways that top-down and bottom-up influences are used and combined during both learning and trained performance[vi]. What they share, though, is this emerging (indeed, state-of-the-art) emphasis on the use of generative models in learning and in recognition.

1.3 Analysis-by-Synthesis

Work that uses generative models to predict inputs implements the much older idea of 'analysis by synthesis' (Neisser (1967), Yuille and Kersten (2006)) where this names a processing strategy in which:

> "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).

In this paradigm the brain does not build its current model of worldly causes by accumulating, bottom-up, a mass of low-level cues. Instead, the brain –in learning, in perceiving, and (see Friston, Mattout, and Kilner (2011)) in acting - tries to predict the current suite of low-level cues from its best high-level models of possible causes (see Bar (2007), Hohwy (2007), Friston (2010)). In this way:

> "Predictive, or more generally, generative models turn the inverse problem [here, the problem of converting sensory 'measurements' into information about external objects and sates of affairs] on its head. Instead of trying to find functions of the inputs that predict their causes, they find functions of estimated causes that predict the inputs" Friston (2002) p. 233

1.4 Hierarchies of Hidden Causes

Finally, it is worth stressing that the top-down generation of the sensory patterns that characterizes these models or approaches always proceeds (after learning) via multiple layers of processing that involve intermediate levels of representation. Thus (to offer an admittedly simplistic example) a program capable of dealing with written text might learn layers that deal (respectively) in words, in letters, and in the various kinds of stroke that make up the letters. Each of these levels of structure has its own characteristic regularities. Certain strokes tend to go together, as they form distinct letters; certain letters tend to go together as they form real words; certain words tend to go together as they make grammatical sentences, and so on. Each level of the processing hierarchy thus deploys a probabilistic generative model whose target is the layer of the hierarchy immediately below[vii]. The internal processing hierarchy thus tracks nested causal structure in the source (sentences). In these models each layer embodies knowledge (taking the form of probability density distributions) about the hidden regularities (for example grammars, causes, or any so-called 'latent variables') that are structuring the data as it is registered at the level below. An interesting implication is thus that the layered structure of the internal model will attempt to recapitulate actual (but hidden) structure in the world. In this way:

> "The hierarchical structure of the real world literally comes to be 'reflected' by the hierarchical architectures trying to minimize prediction error, not just at the level of sensory input but at all levels of the hierarchy" Friston (2002) p. 238

To perceive the world, on these accounts, is to attempt to unearth layer upon layer of the actual causal structures that generated the sensory signals impinging on the organism.

2. The Case for Predictive Coding

2.1.Evidence

The predictive coding approach, by using a hierarchical generative model to do top-down sensory prediction in learning and recognition, makes good sense of - and very efficient use of - a complex neuro-anatomy in which recurrent connectivity is massive and apparently functionally asymmetric (see e.g. Friston (2005), Bubic et al (2010)). It also explains several superficially distinct phenomena via a single fundamental mechanism. These include priming, end-stopping (see section 1 above), repetition suppression, and confirmation bias. In the case of priming, recent results show that an expected percept becomes consciously available about 100 ms faster than an unexpected one (see Melloni et al (2011)). This, as the authors note, is easily explained if the process of stimulus recognition involves the activation of a top-down generative model that is attempting to match the incoming data stream with its own predictions. It is also well-known 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. This too is fluently explained by the predictive coding story: repetition normally reduces response because it reduces prediction error. Repetition-suppression may thus be a direct effect of predictive coding strategies at work in the brain, and would hence vary according to our local perceptual expectations.

More generally, there is an emerging body of supportive fMRI and EEG work dating back to a pioneering fMRI study by Murray et al (2002) that reveals just the kinds of relationships posited by the predictive coding story. Here, 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. Recent studies confirm this general profile (see eg Alink et al (2010)).

2.2 Questioning Predictive Coding

Early examples of the predictive coding approach (such as the seminal 1997 work by Rao and Ballard described in section 1 above) were, however, met with some puzzlement, since they seemed radically different from the more standard picture of an (admittedly attention-modulated) feedforward cascade of simple-to-complex feature detection. This puzzlement is well-captured by the comments from Koch and Poggio that accompanied the publication of this work. The passage is so perfectly expressive of some quite common worries that I hope the reader will forgive a long extract:

> "In predictive coding, the common-place view of sensory neurons as detecting certain 'trigger' or 'preferred' features is turned upside down in favor of a *representation of objects by the absence of firing activity*. This appears to be at odds with [data indicating that neurons] extending from V1 to inferior temporal cortex, respond with *vigorous activity to ever more complex objects*, including individual faces or paperclips twisted in just the right way and seen from a particular viewpoint"
>
> "In addition, what about all of the functional imaging data from humans revealing that particular cortical areas respond to specific image classes, such as faces or three-dimensional spatial layout? *Is it possible that this activity is dominated by the firing of… cells actively expressing an error signal, a discrepancy* between the input expected by this brain area and the actual image?"
>
> (Both quotes from Koch and Poggio (1999) p 10, my emphasis)

There are two main worries being expressed here. First, a worry that the accounts are abandoning representation in favour of silence, since well-predicted elements of the signal are 'explained away'. Second, a worry that the accounts thus seem in tension with strong evidence of increasingly complex representations tokened by activity in higher areas. Neither worry is justified however. To see why not, recall the architectural story outlined earlier. Each layer, that story insists, must somehow support two functionally distinct kinds

of processing. For simplicity, let's follow Friston (2005) and imagine this as each layer containing two functionally distinct kinds of cell or unit[viii]:

> - 'representation units', that encode that layer's current best hypothesis (pitched at its preferred level of description) and that feed that hypothesis down as prediction to the layer below.
>
> - 'error units', that pass activation forward when local within-layer activity is not adequately accounted for by incoming top-down prediction from the layer above

That means that more and more complex representations are indeed formed, and used in processing, as one moves up the hierarchy. It is just that the *flow* of representational information (the predictions), at least in the purest versions, is all downwards. It is in this sense that, as we saw earlier, the role of feedback is inverted in these models. Moreover, the upward flow of prediction error is itself a sensitive instrument, bearing fine-grained information about very specific failures of match. That's why it is capable of inducing, in higher areas, complex hypotheses (consistent sets of representations) that can then be tested against the lower-level states[ix]. As a result, neither of the two worries raised by Koch and Poggio gets a grip. There are representational populations all the way up, and their activity is determined by the flow of error signals and the hypotheses that they select.

2.3 An Example: The Fusiform Face Area

Consider again the standard model of a stream of increasingly complex feature-detection, such that responses (at the highest levels) reflect the presence of such items as faces, houses, etc. What the predictive coding story suggests is not that we abandon that model but that we enrich it, by adding within each layer cells specialized for the encoding and transmission of prediction error. Some cells at each level, if this is correct, are encoding features while others are registering errors relative to predictions about those features coming from the level above.

The right evidence here is only just appearing, but it actually seems to fit best with this more complex 'predictive coding' profile. Thus consider the well-established finding (Kanwisher et al (1997)) of increased activity in fusiform face area FFA when shown a face rather than (say) a house. Surely, a critic

might say, this is best explained by simply supposing that neurons in FFA have learnt to be active complex feature detectors for faces? It is immediately apparent that this is no longer straightforward, however, given that the predictive-coding story allows that FFA may indeed harbor units that specialize in the representation of faces, as well as ones that specialize in the detection of errors (mismatches between top-down predictions reaching FFA and the bottom-up signal). Thus, the difference is that if the predictive coding story is correct, FFA should *also* harbor error units that encode mismatches with predicted (face) activity. This provided a nice opportunity for some telling empirical tests.

Egner et al (2010) compared simple feature detection (with and without attention) and predictive coding models of recorded responses in FFA. The simple Feature Detection Model predicts, just as Koch and Poggio suggested, that FFA response should simply scale with the presence of faces in the presented image. The Predictive Coding Model, however, predicts something rather more complex. It predicts that FFA response should "reflect a summation of activity related to prediction ("face expectation") and prediction error ("face surprise")" (op cit p 1601). That is to say, it predicts that the (temporally rather blunt) fMRI signal recorded from the fusiform face area should reflect the activity of both putative kinds of cell: those specializing in prediction ("face-expectation") and those specializing in detecting errors in prediction ("face-surprise"). This was then tested by collecting fMRI data from area FFA while independently varying both the presented features (face vs. house) and – by means of a simple (though not explicitly revealed to the participants) preceding cue manipulating subject's unconscious degree of face expectation (low, medium, high) and hence their proper degree of 'face surprise'. To do this, the experimenters probabilistically paired presentations of face/house with a 250 ms preceding color frame cue giving 25% (low), 50% (medium) or 75% (high) chance of the next image being a face.

The results were clear. FFA activity showed a strong interaction between stimulus and face-expectation. FFA response was maximally differentiated only under conditions of low face expectation. Indeed, and quite surprisingly, FFA activity given *either* stimulus (face OR house) was indistinguishable under conditions of high face expectation! There is a very real sense then, in which FFA might (were it first investigated using these new paradigms) have been dubbed a 'face-expectation area'. The authors conclude that, contrary to any simple feature-detection model:

> "[FFA] responses appear to be determined by feature expectation and surprise rather than by stimulus features per se" Egner et al (2010) p16601

The authors also controlled (by further model comparisons) for the possible role of attentional effects. But these could not, in any case, have made much contribution since it was face surprise, not face expectation, that accounted for the larger part of the BOLD (fMRI)[x] signal. In fact, the best-fit predictive coding model used a weighting in which face-surprise (error) units contributed about twice as much[xi] to the BOLD signal as did face-expectation (representation) units, suggesting that much of the activity normally recorded using fMRI may be signaling prediction-error rather than detected features!

This is an important result. In the authors' own words:

> "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 (with or without attention)" Egner et al (2010) p. 16607

3. A New Look at Sensory Processing

Among the guiding themes of this volume, we find the notion of the senses as "integrated information pickup systems" and various puzzles involving multimodal and crossmodal effects, plasticity, and the individuation of the senses. Hierarchical predictive processing offers insights into all these phenomena, rendering unsurprising much that was previously puzzling, and also rendering a little more puzzling some things that we might otherwise take for granted. In this final section I offer a few (tentative and preliminary) reflections on this altered landscape.

3.1 Causes and Operators

Reich et al (2011) report some interesting new fMRI findings regarding the so-called Visual Word Form Area (VWFA). This is an area within the ventral stream that responds to proper letter strings: the kind that might reasonably form a word in a given language. Response in this area was already known to be independent of surface details such as case, font, and spatial location. The recent study shows that it is actually tracking something even more abstract

than visual word form. It appears to be tracking wordform, regardless of the modality of the transducing stream. Thus the very same area is activated in congenitally blind subjects during Braille reading. The fact that the early input here is tactile rather than visual makes no difference to the recruitment of VWFA. This supports the idea (Pascual-Leone and Hamilton (2001)) of such brain areas as 'metamodal operators' that are "defined by a given computation that is applied regardless of the sensory input received".

All this fits neatly, as Reich et al (2011 p.365) themselves note, with the predictive coding account in which higher levels of the cortical hierarchy learn to track the 'hidden causes' that account for, and hence predict, the sensory consequences of distal states of affairs. In a deliberate echo of the Egner et al work on the fusiform face area, Reich et al speculate that much activity in VWFA might reflect modality-transcending predictions about the sensory consequences of words. Just as (as we saw in 2.3 above) much of the activity in FFA is related to top-down face-prediction rather than the bottom-up detection of faces, so the VWFA might be generating top-down predictions using modality-transcending models of wordhood. The metamodality of VWFA would then "explain its ability to apply top-down predictions to both visual and tactile stimuli" (Reich et al (2011) p.365.

3.2 Cross-Modal and Multi-Modal Effects

The widespread existence of cross- and multi-modal context effects on early 'unimodal' sensory processing constitutes one of the major findings of the last decade of sensory neuroscience (see eg Hupe et al (1998), Murray et al (2002), Smith and Muckli (2010)). Thus Murrray et al (2002) display the influence of high-level shape information on the responses of cells in early visual area V1, while Smith and Muckli (2010) show similar effects (using as input partially occluded natural scenes) even on wholly non-stimulated (that is to say, not directly stimulated via the driving sensory signal) visual areas. In addition, 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 will commonly reflect a variety of multimodal associations. Strikingly, 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 improve 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 smörgåsbord of context effects flows very naturally from the hierarchical predictive coding model. If so-called visual, tactile, or auditory sensory cortex is actually operating using a cascade of feedback from higher levels to actively predict the unfolding sensory signals (the ones originally transduced using the various dedicated receptor banks of vision, sound, touch, etc) then we should not be in the least surprised to find extensive multi-modal and cross-modal effects (including these kinds of 'filling-in') even on 'early' sensory response[xii]. One reason this will be so is that the notion of 'early' sensory response is in one sense now misleading, for expectation-induced context effects will simply propagate all the way down the system, priming, generating, and altering 'early' responses as far down as V1. Any statistically valid correlations, registered within the 'metamodal' (or at least, increasingly information-integrating) areas towards the top of the processing hierarchy, can inform the predictions that then cascade down, through what were previously thought of as much more unimodal areas, all the way to the areas closer to the sensory peripheries. Such effects are 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 consistent with (indeed, mandated by) models that depict V1 activity as constantly negotiated on the basis of a flexible combination of top-down predictions and driving sensory signal[xiii].

3.3 Expectations and Conscious Perception

All this has implications for the study of (the neural correlates of) sensory awareness. The key observation (Melloni et al (2011)), and one that will surely add new layers of complexity to many familiar experimental paradigms, is that expectation speeds up conscious awareness.

We can creep up on this with some mundane reflections. It is intuitively obvious that, for example, a familiar song played using a poor radio receiver will sound much clearer than an unfamiliar one. But whereas we might have thought of this, within a simple feed-forward feature-detection framework, as some kind of memory effect, it now seems just as reasonable to think of it as a genuinely perceptual one. The clear-sounding percept, after all, is constructed in just the same way as the fuzzy-sounding percept, albeit using a better set of top-down predictions (priors, in the Bayesian translation of the story). That is to say – or so I would suggest - the familiar song *really does* sound clearer. It is

not that memory *later* does some filling-in that affects, in some backward-looking way, how we judge the song to have sounded. Rather, the top-down effects bite in the very earliest stages of processing, leaving us little[xiv] conceptual space (or so it seems to me) to depict the effects as anything other than enhanced-but-genuine perception.

We can illustrate this with a little thought experiment. Imagine we discover a creature whose auditory apparatus is highly tuned to the detection of some biologically relevant sound. Imagine too that that tuning consists largely in a strong set of priors for that sound, such that the creature can detect it despite considerable noise in the ambient signal (a kind of cocktail party effect). Surely we would simply describe this as a case of acute perception? Then we must say the same, it seems to me, of the music-lover hearing a familiar song from a low-quality radio.

In exactly this vein, Melloni et al (2011) show that the onset time required to form a reportable conscious percept varies according to our expectations. Following a fairly complex series of experiments (due to the need to carefully control for effects that would be best attributed to non-perceptual 'advance guessing' rather than to genuine enhanced visibility for the better-predicted stimulus), the authors conclude that 'expectations alter the threshold of visibility' (op cit p.1393). They explain this result by explicit appeal to a hierarchical predictive coding framework in which "conscious perception is the result of a hypothesis test that iterates until information is consistent across higher and lower areas" (op cit p. 1394). Using electroencephalographic (EEG) signatures, it was calculated that conscious perception could occur as rapidly as 100ms faster for a well-predicted stimulus, and hence that:

> "the signatures of visibility are not bound to processes with a strict latency but depend on the presence of expectations" (Melloni et al (2011) p. 1395

In addition, Muckli (2010) reports that predicted stimuli, although able to drive better and faster behavioral responses, showed reduced fMRI activation in V1. This is further evidence for the predictive coding story since:

> "Finding reduced activity related to increased performance fits well with the framework of predictive coding…but is difficult to explain otherwise" Muckli (2010) p. 135

3.4 Sensorimotor Contingency Theory

The notion that sensory experience is in some way bound up with predictions and expectations resonates to some degree with recent important and influential work in 'sensorimotor contingency theory' (O'Regan and Noë (2001), Noë (2004), Noë (2009)). There are, however, some notable differences. First, sensorimotor contingency theory (SMC) staunchly champions the view that the predictions that matter will mostly concern the ways the sensory signal will vary with bodily movement. That is to say, they are both prospective (they concern future variation in the incoming signal) and their contents are sensorimotor profiles. Such prospective sensorimotor predictions, though often extremely important, constitute merely one dimension of the very large space of features and properties that can figure in the downward-cascades posited by hierarchical predictive coding accounts. Moreover, the SMC model (at least as explicated and defended by Noë (2004) (2009)) looks committed to an implausibly shallow processing account (see Clark (2008) ch. 8) that omits any essential appeal to those multiple, stacked layers of internal representation - the crucial hierarchical generative model - that translate top-down predictions, via many intervening stages, into predictions concerning the actual ebb and flow of the driving sensory inputs. SMC models, though absolutely correct to highlight prediction and expectation in their account of perception, are thus neglecting the critical machinery (of prediction-induced hierarchical generative models) that enable brains like ours to infer complex hidden causal structures in the world.

3.5 Distinguishing and Extending the Senses

The predictive coding framework offers, we saw, a powerful way of accommodating all manner of cross- and multi-modal effects on perception. It depicts the senses as working together to provide feedback (recall the explanatory inversion highlighted in section 1.3 above) to a linked set of prediction devices that are attempting to track unfolding states of the world across multiple spatial and temporal scales. This delivers a very natural account of efficient multi-modal cue integration (see Ernst and Banks (2002)), and allows top-down effects to penetrate even the lowest (earliest) elements of sensory processing. It also induces (1.2 above) a potent duality between sensing and top-down generation, so that to perceive some state of affairs requires the system to be capable (though not necessarily of its own volition) of endogenously generating the relevant sensory signature. Certain things that might otherwise be taken for granted then stand in need of explanation.

One is the existence of distinct modalities in experience. Why, given that the senses work together to provide ongoing feedback on 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 states of affairs without the sense of distinct modalities? I would speculate (and no more than speculate) that the answer may involve the different kinds of uncertainty that are associated with different sensory channels. In a thick fog, for example, vision is unreliable (delivering information with high uncertainty) while audition is less affected. The brain will need to mark these differences so as to weight and integrate the available cues (from multiple sense organs) in different ways on different occasions. Perhaps we experience sensory modalities as different from one another just to the extent that they are prone (in many contexts) to deliver information with very different degrees of uncertainty? Where the uncertainties more nearly match, we experience one modality (eg vision) with multiple sense organs (two eyes).

This would mean that the character of sensory experience has (at least) two components. One is the stacked set of generative models that capture the regularities in the world. The other is the signature forms of uncertainty associated with the sensory channel itself.

3.6 Perceiving and Imagining

Another question the new framework raises is, Why is imagination not just like conscious perception? Given the role of generative models, and the deep duality between perception and the top-down generation of the sensory signal, one might expect imagination to share the full (rich, vivid) experiential signature of ordinary perception[xv]. Yet (bracketing cases of hallucination etc) this does not seem to be the case. Here too, I will venture one last (mere) speculation.

Sensory processing, on these models, involves predicting (across a hierarchy of processing regions) the driving signal transduced from the world. That means that the flow of information from the environment really matters, as it delivers the evolving data stream that the top-down model has to try (using the linked stack of generative models) to match. An important feature of this process is that the weight that is given to the driving sensory signal (hence the value of prediction errors concerning that signal) can be varied according to its degree of certainty or uncertainty. This is achieved by altering the gain (the 'volume' to use the standard auditory analogy) on the error-units accordingly. The effect of

this is to allow the brain to vary the balance between sensory inputs and prior expectations at different levels (see Friston (2009) p. 299). This means that the weighting of sensory prediction errors (hence the relative influence of sensory inputs and prior expectations) at any level of processing within the whole hierarchical cascade may itself be flexibly modulated. This is sometimes described as optimizing "the relative precision of empirical (top-down) priors and (bottom-up) sensory evidence" (Friston (2009) p. 299). All this is suggestive, it seems to me, of a possible explanation for the experiential asymmetry between perception and (ordinary non-vivid) imagination.

Thus suppose you are looking for an object on a crowded surface. You expect to see it somewhere, but you are not sure where. Your brain must temporarily increase the weighting on the fine spatial information carried by the driving signal. That way, you don't simply mistakenly see it *there* (at such and such a location) just because you are expecting to see it *somewhere*. To match the driving signal with a top-down prediction here demands accounting for the sensory signal in great detail, all the way down (as it were).

Non-vivid mental imagery, by contrast, may be calling only upon higher levels of the generative model. Thus compare the case where you are asked to imagine your walk to work. Here, early (closer to retinotopic) stages of the processing hierarchy can be allowed substantial leeway, and need not be forced to settle into one interpretation or another. It seems plausible (though this is, to repeat, currently no more than speculation) that under such conditions one might experience the self-generated imagery as fuzzier and less distinct than online perception, even though perception and imagination are simply different ways of deploying the very same circuits and fundamental capacities[xvi].

4. Conclusions: Perceiving, Imagining….Knowing?

The main purpose of this chapter has been to introduce the notion of sensory perception as a form of probabilistic prediction involving a hierarchy of generative models[xvii]. This broad vision brings together frontline research in machine learning and a growing body of neuroscientific conjecture and evidence. It provides a simple and elegant account of multi-modal and cross-modal effects in perception, and has implications for the study of (the neural correlates of) conscious experience. It also suggests, or so I have argued, a deep unity between perceiving and imagining. For to perceive the world (at least as we do) is to deploy internal resources capable of endogenously generating those same sensory effects: capable, that is, of generating those same activation

patterns via a top-down sweep involving multiple intermediate layers of processing. That suggests a fundamental linkage between 'passive perception' and active imagining, with each capacity being continuously bootstrapped by the other. Perceiving and imagining (if these models are on the right track) are simultaneous effects of a single underlying neural strategy.

In closing, I cannot resist sharing one further thought, even though it goes far beyond the conclusions warranted by the present treatment. It is that this unity of perceiving and imagining in turn suggests a deep continuity between perceiving (thus construed) and understanding[xviii]. For such systems are able to predict the way the sensory signal will evolve, on multiple temporal and spatial scales, and to generate those transformations in advance using endogenous resources alone. When such systems perceive the world, they know how the world is structured by hidden causes and they know how it is likely to evolve over time. This, surely, is to make deep inroads not just into the explanation of effective perception but also into the origins of meaning and semantics: the elusive realm of 'aboutness' itself.

* This paper owes much to discussions and exchanges at *The Senses Research Workshop* (Institut Jean-Nicod, Paris, 2009), at the workshop on *Predictive Coding and the Senses* (Institute of Philosophy, London, 2011), and during my stay as a Visiting Fellow at the SAGE Institute at UC Santa Barbara. Thanks to Dustin Stokes, Mohan Matthen, and Barry Smith for many useful discussions and for making the workshops possible. Thanks to Mike Gazzaniga and the audiences at UCSB, and especially Michael Rescorla. Thanks too to Karl Friston, Jakob Hohwy, and Chris Williams for helping me begin to negotiate the dizzying maze of work on predictive coding, active inference, and generative models. They are not to blame for any errors or speculative excesses. The paper also benefitted greatly from helpful suggestions from Jon Bird, and from the editors of the present volume. This work was supported in part by an AHRC Speculative Research Grant (PI Y. Rogers, Open University) 'Extending the Senses and Self Through Novel Technologies'

References

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

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

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

Bengio, Y and Le Cun, Y (2007) Scaling Learning Algorithms towards AI, in*: Large Scale Kernel Machines*, MIT Press

Biederman, I. (1987) Recognition-by-components: a theory of human image understanding. *Psychological Review* 94, 115–147

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

Byrne, A. and Logue, H. (eds.) (2009) *Disjunctivism: Contemporary Readings* (Cambridge MA: The MIT Press).

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

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

Dennett, D. (1988) Quining Qualia, in A. Marcel and E. Bisiach, eds, *Consciousness in Modern Science*, Oxford University Press. Reprinted in W. Lycan, ed., *Mind and Cognition: A Reader*, MIT Press, 1990, A. Goldman, ed. *Readings in Philosophy and Cognitive Science*, MIT Press, 1993

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.

Engel, A., Fries, p., and Singer, W. (2001) "Dynamic Predictions: Oscillations and Synchrony in Top-Down Processing" *Nature Reviews: Neuroscience*: 2: 704-716

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

Feldman, H., and Friston, K. (2010) "Attention, Uncertainty, and Free Energy" *Frontiers in Human Neuroscience* 4: 215: 1-23

Friston, K. (2002). "Beyond phrenology: What Can Neuroimaging Tell Us About Distributed Circuitry?" *Annual Review of Neuroscience* **25**(1): 221-250.

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? *Nature Reviews: Neuroscience* 11(2):127-38.

Friston K, Mattout J, Kilner J. Action understanding and active inference. *Biological Cybernetics*. 2011 104:137–160

Gregory, R (1998) Brainy mind. *British Medical Journal* 317:1693—5

Haddock, A., and Macpherson, F. (eds.) (2008) *Disjunctivism: Perception, Action, and Knowledge* (Oxford: Oxford University Press).

Heeger, D. & Ress, D. (2002) What Does fMRI Tell Us About Neuronal Activity? *Nature Reviews/Neuroscience* 3:142-151.

Heess, N., Williams, C., and Hinton, G. (2009) Learning generative texture models with extended Fields-of-Experts *Proceedings BMVC*

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 (2007a) 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.

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

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

Hinton, G., Dayan, P., Frey, B. J., and Neal, R. M. (1995). The wake-sleep algorithm for unsupervised neural networks. *Science,* 268:1158-1161

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

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

Huang, Y. and Rao, R. P. N. (2011), Predictive coding. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2:

Hubel, D. H. & Wiesel, T. N. (1965) Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *Journal of Neurophysiology* **28**, 229–289

Hupé, JM, James, AC, Payne, BR, Lomber, SG, Girard, P, and Bullier, J. (1998) Cortical feedback improves discrimination between figure and background by V1, V2 and V3 neurons. *Nature* 394: 784-787

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

Kanwisher NG, McDermott J, Chun MM (1997) The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience* 17:4302--4311.

Koch and Poggio (1999) "Predicting the Visual World: Silence is Golden" *Nature Neuroscience* 2:1:

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

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

Marr, D. (1982) *Vision* (Freeman, San Francisco).

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

Muckli, L (2010) What Are We Missing Here? Brain Imaging Evidence for Higher Cognitive Functions in Primary Visual Cortex V1 *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 loop. *Biological Cybernetics*, 66:241-251

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

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

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

Noë, A (2009) *Out of Our Heads: Why You Are Not Your Brain, and Other Lessons from the Biology of Consciousness* Farrar, Straus and Giroux: NY

O'Regan, J. K. and Noë, A. (2001) A sensorimotor approach to vision and visual consciousness. *Behavioral and Brain Sciences* 24/5: 883-975.

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

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).

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

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986) Learning internal representations by error propagation. In Rumelhart, D. E., McClelland, J. L., and the PDP Research Group, editors, *Paralled Distributed Processing. Explorations in the Microstructure of Cognition. Volume 1: Foundations*, pages 318-362. The MIT Press, Cambridge, MA.

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)

Spratling, M and Johnson, M (2006) A feedback model of perceptual learning and categorization. *Visual Cognition* 13:2: 129-165

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

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

NOTES

[i] Examples include Hubel and Wiesel (1965), Marr (1982), Biederman (1987).

[ii] Actually, a variant form of predictive coding also characterizes the work of the retina (see Hosoya et al (2005). But this nicety need not concern us here.

[iii] By 'perceive' I here mean 'perceive in a rich, full-blooded manner'. Obviously, a simple robot that locomotes to a light source need not, and probably should not, deploy a stack of generative models to do so. The need for generative models emerges most clearly when systems must deal with noise, ambiguity, and uncertainty.

[iv] The (theoretically mandated) duality of perception and generation means that a percept and a hallucination could (in principle at least) involve identical neural states. This may put pressure, it seems to me, on some (but not all) formulations of disjunctivism - the idea, roughly, that veridical percepts and hallucinations, illusions etc share no common kind. Much turns, of course, on how the somewhat obscure disjunctivist claim is to be unpacked. For a pretty comprehensive sampling of possible formulations, see essays in A. Haddock and F. Macpherson (eds.) (2008), and in Byrne, A. and Logue, H. (eds.) (2009).

[v] The crucial innovation was to learn one layer of representation at a time using what Hinton calls Restricted Boltzmann Machines, with a further tweak to fine-tune the resulting overall model – for an accessible summary, see Hinton (2007).

[vi] Compare, for example, the kinds of model described by Hinton (2007) with that of Jehee and Ballard (2009).

[vii] It will not in general, however, be easy to determine what individual units/neurons within a layer represent – see Hinton (2007) box 2, p.433.

[viii] Possible alternative implementations are discussed in Spratling and Johnson (2006), and in Engel et al (2001)):

[ix] It is also worth noting that models that fit the rather more general profile described in 1.2 above (viz, using hierarchical generative models to predict the sensory signals) are not compelled to endorse the full 'explaining away' procedure. Instead, at each level, the full 'silencing' of representation units by (good) downward prediction is actually only one – neat, easy-to-grasp, but potentially quite extreme - option among many for how best to combine top-down predictions with bottom-up inputs (for some glimpses of the much larger computational spaces hereabouts, see Feng et al (2002), Hinton (2007), Bengio and Lecun (2007), Heess et al (2009), Hinton (2010))

[x] This is a measure of relative neural activity ('brain activation') as indexed by changes in blood flow and blood oxygen level. The assumption is that neural activity incurs a metabolic cost that this signal reflects. It is thus widely acknowledged (see e.g. Heeger and Ross (2002)) to be a rather indirect, assumption-laden, and 'blunt' measure compared to, say, single cell recording. Nonetheless, new forms of multivariate pattern analysis are able to overcome some of the limitations of earlier work using this technique.

[xi] This could, the authors note, be due to some fundamental metabolic difference in processing cost between representing and error-detection, or it may be that for other reasons the BOLD signal tracks top- down inputs to a region more than bottom-up ones (see Egner at al (2010) p. 16607).

[xii] This may also have implications for other familiar questions, such as whether context *really* alters the nature of a perceptual experience. Does the wine really taste better when tasted within sight of the sea? Or do we merely then judge it to taste better (do we, that is, merely judge the *same taste differently*, due to some contextual effect)? Or is this simply a non-question (recall Dennett's (1988) treatment of the coffee tasters Chase and Sanborn)? The proposed framework allows us to at least frame the issue better, though a proper resolution remains elusive. Thus suppose selection of some top-level amodal feature complex ('fresh, healthy') is caused by the need to account for (predict) visual sea-features impinging on the eyes. If that in turn affects which higher–level generative models are selected and applied to predict the unfolding taste and flavor of the wine, that might favour encodings of e.g. 'young and vital' over close rivals like 'acidic and immature'. Since this kind of multi-modal give and take is just the norm for *any* perceptual unfolding on these models, I'd like to say that means the wine really *tastes* different. But unfortunately it is not yet clear that this is mandated. Someone could say (pursuing a more conservative option) that the wine really tastes different only if such contextual nuancing alters the suite of predictions that are being successfully applied far down the gustatory processing hierarchy, at levels that are intuitively encoding information about the low-level driving chemical signals themselves. If so, then were there no alteration in *those* predictions that would mean no alteration in the target experience. Alternatively, experience could be much more holistically determined so that alterations anywhere up the hierarchy would impact the percept, even if they make the very same predictions lower down. This is my own preferred (but admittedly unargued) option: conscious experience reflects the settling of the whole hierarchy into some temporarily stable state.

[xiii] Reflecting on this new vision of 'early' sensory processing, Lars Muckli writes that "It is conceivable that V1 is, first of all, the target region for cortical feedback and then, in a second instance, a region that compares cortical feedback to incoming information. 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

[xiv] There is doubtless some kind of slippery slope here, as we progressively degrade the driving signal and upregulate the expectations. Negotiating this complex terrain is, however, a task for another day.

[xv] Thanks to Mark Sprevak (personal communication) for raising this issue

[xvi] Reddy et al (2010) neatly demonstrate that imagery and perception are not simply activating overlapping neural areas but are actually deploying the very same fine-grained internal representations when they do so. In the cases they investigate, that overlap is restricted, as the present speculation suggests, to somewhat higher levels of the visual processing hierarchy. The authors suggest, though, that were the task to have demanded it, lower level areas such as V1 might have been re-activated in the same top-down manner.

[xvii] Technically, there is then a single (but hierarchical) generative model. Nothing in the present treatment turns on this detail.

[xviii] For a rather more conceptual route to what seems to me to be a similar conclusion, see Timothy Williamson's *New York Times* 'Opinionator' blog entry 'Reclaiming the Imagination' available at:
http://opinionator.blogs.nytimes.com/2010/08/15/reclaiming-the-imagination/