BEHAVIORAL AND BRAIN SCIENCES (2013) **36**, 181–253
doi:10.1017/S0140525X12000477

EDITOR'S NOTE

A exceptionally large number of excellent commentary proposals inspired a special research topic for further discussion of this target article's subject matter, edited by Axel Cleeremans and Shimon Edelman in Frontiers in Theoretical and Philosophical Psychology. *This discussion has a preface by Cleeremans and Edelman and 25 commentaries and includes a separate rejoinder from Andy Clark. See:*

http://www.frontiersin.org/Theoretical_and_Philosophical_Psychology/researchtopics/Forethought_as_an_evolutionary/1031

Whatever next? Predictive brains, situated agents, and the future of cognitive science

Andy Clark
School of Philosophy, Psychology, and Language Sciences, University of Edinburgh, EH8 9AD Scotland, United Kingdom
andy.clark@ed.ac.uk
http://www.philosophy.ed.ac.uk/people/full-academic/andy-clark.html

Abstract: Brains, it has recently been argued, are essentially prediction machines. They are bundles of cells that support perception and action by constantly attempting to match incoming sensory inputs with top-down expectations or predictions. This is achieved using a hierarchical generative model that aims to minimize prediction error within a bidirectional cascade of cortical processing. Such accounts offer a unifying model of perception and action, illuminate the functional role of attention, and may neatly capture the special contribution of cortical processing to adaptive success. This target article critically examines this "hierarchical prediction machine" approach, concluding that it offers the best clue yet to the shape of a unified science of mind and action. Sections 1 and 2 lay out the key elements and implications of the approach. Section 3 explores a variety of pitfalls and challenges, spanning the evidential, the methodological, and the more properly conceptual. The paper ends (sections 4 and 5) by asking how such approaches might impact our more general vision of mind, experience, and agency.

Keywords: action; attention; Bayesian brain; expectation; generative model; hierarchy; perception; precision; predictive coding; prediction; prediction error; top-down processing

1. Introduction: 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 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

 0140-525X/13 $40.00

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. Such a process, operating over multiple linked higher-level models, yields 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's 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 & 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 and so forth. 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 & Manning 2006, p. 340, their emphasis)

Helmholz's 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 & Hinton 1996; see also Hinton & Zemel 1994).[2] The Helmholtz Machine sought to learn new representations in a multilevel 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 multilevel (hierarchical and bidirectional) system come to encode a probabilistic model of the activities of units and groups of units within lower levels, 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, for example, Kawato et al. (1993), Hinton and Zemel (1994), Mumford (1994), Hinton et al. (1995), Dayan et al. (1995), Olshausen and Field (1996), Dayan (1997), and 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 multilevel cascade – that lies at the heart of "hierarchical predictive coding" approaches to perception; for example, Rao and Ballard (1999), Lee and Mumford (2003), Friston (2005). Such approaches, along with their recent extensions to action – as exemplified in Friston and Stephan (2007), Friston 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 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 – see Rao and Ballard (1999), Lee and Mumford (2003), Friston (2005), Hohwy et al. (2008), Jehee and Ballard (2009), Friston (2010), Brown et al. (2011); and, 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 & 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 therefore just the difference (a.k.a. the "prediction error") between the actual current signal and the predicted one. This affords major savings on bandwidth, an economy that was the driving force behind the development of the techniques by James Flanagan and others at Bell Labs during the 1950s (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,

ANDY CLARK is Professor of Logic and Metaphysics in the School of Philosophy, Psychology, and Language Sciences at the University of Edinburgh in Scotland. He is the author of six monographs, including *Being There: Putting Brain, Body and World Together Again* (MIT Press, 1997), *Mindware* (Oxford University Press, 2001), *Natural-Born Cyborgs: Minds, Technologies and the Future of Human Intelligence* (Oxford University Press, 2003), and *Supersizing the Mind: Embodiment, Action, and Cognitive Extension* (Oxford University Press, 2008). In 2006 he was elected Fellow of the Royal Society of Edinburgh.

and in motion-compressed coding for video. The information that needs to be communicated "upward" 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 & Friston 2010) for sensory information itself. Later, when we consider predictive processing in the larger setting of information theory and entropy, we will see that prediction error reports the "surprise" induced by a mismatch between the sensory signals encountered and those predicted. More formally – and to distinguish it from surprise in the normal, experientially loaded sense – this is known as *surprisal* (Tribus 1961).

Hierarchical predictive processing combines the use, within a multilevel bidirectional 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 sect. 1.5) 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 (in sects. 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", 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 is 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 bidirectional 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 et al. 1977; Neal & Hinton 1998) that allows priors and models to co-evolve across multiple linked layers of processing so as to account for the sensory data. The presence of bidirectional hierarchical structure thus induces "empirical priors"[6] in the form of the constraints that one level in the hierarchy places on the level below, and these constraints are progressively tuned by the sensory input itself. 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 (Friston 2005; Lee & Mumford 2003).[7]

A classic early example, combining this kind of hierarchical learning with the basic predictive coding strategy described in section 1.1, 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 multilevel cascade in which each level attempts to predict the activity at the level below it via backward[8] 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. This automatically adjusts probabilistic representations at the higher level so that top-down predictions cancel prediction errors at the lower level (yielding rapid perceptual inference). At the same time, prediction error is used to adjust the structure of the model so as to reduce any discrepancy next time around (yielding slower timescale perceptual learning). Forward connections between levels thus carry the "residual errors" (Rao & 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 bidirectional hierarchy of cortical areas allows information pertaining to regularities at different spatial and temporal scales to settle into a mutually consistent whole in which each "hypothesis" is used to help 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 & Ballard 1999, p. 80)

In the 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[9] indicating residual (unpredicted) activity.

To test these ideas, Rao and Ballard implemented a simple bidirectional 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 sect. 3.1) 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 learned 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 center 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 center-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.[10]

These computations of predicted salience might be made solely on the basis of average image statistics. Such an approach would, however, lead to trouble in many ecologically realistic situations. To take some of the more dramatic examples, consider an animal that frequently 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. This is true in less dramatic cases too, such as when we move from inside a building to a garden or lake. Hosoya et al. thus predicted that, in the interests of efficient, adaptively potent, encoding, the behavior 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 behaviors 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 et al.'s (2008) hierarchical predictive coding model of binocular rivalry. Binocular rivalry (see, e.g., essays in Alais & Blake 2005, and the review article by Leopold & 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 et 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 et al. 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 hierarchical generative model. 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 multilevel setting, a visual percept is determined by a process of prediction operating across many levels of a (bidirectional) 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 (multiscale) 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 et al. 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 – for example, 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 the 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; it 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 that of Lee et al. 2005) is that whereas they posit a kind of direct, attention-mediated but essentially feedforward, competition between the inputs, the predictive processing account posits "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.[11]

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 that, for example, houses and faces are not present in the same place, at the same scale, at the same time. This kind of general knowledge may itself be treated as a systemic prior, albeit one pitched at a relatively high degree of abstraction (such priors are sometimes referred to as "hyperpriors"). In the case at hand, what is captured is the fact that "the prior probability of both a house and face being co-localized in time and space is extremely small" (Hohwy et al. 2008, p. 691). 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" (Hohwy et al. 2008, p. 691). (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: for example, seeing the painting as valuable, as a Rembrandt, as an image of a cow, etc.)

1.5. Action-oriented predictive processing

Recent work by Friston (2003; 2010; and with colleagues: Brown et al. 2011; Friston et al. 2009) generalizes this basic "hierarchical predictive processing" model to include action. According to what I shall now dub "action-oriented predictive processing,"[12] 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 & Blakeslee 2004, p. 158)

A closely related body of work in so-called optimal feedback control theory (e.g., Todorov 2009; Todorov & Jordan 2002) displays the motor control problem as mathematically equivalent to Bayesian inference. Very roughly – see Todorov (2009) 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 2009). 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 (2009, p. 29) comments, "does not distinguish between the problems of sensor processing, motor control, or planning." Toussaint's bold claim is modified, however, by the important caveat (op. cit., p. 29) that we must, in practice, deploy approximations and representations that are specialized for different tasks. But 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 inference, where the agent moves its sensors in ways that amount to actively seeking or generating the sensory consequences that they (or rather, their brains) expect (see Friston 2009; Friston et al. 2010). 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 et al. 2009, p. 12)

In some (I'll call them the "desert landscape") versions of this story (see especially Friston 2011b; Friston et al. 2010) 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.[13] I return briefly to these "desert landscape" scenarios in section 5.1 further on.

1.6. The free energy formulation

That large-scale picture (of creatures enslaved to sense and to act in ways that make most of their sensory predictions come true) finds fullest expression in the so-called free-energy minimization framework (Friston 2003; 2009; 2010; Friston & Stephan 2007). 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 "surprisal" (where this names the sub-personally computed implausibility of some sensory state given a model of the world – see Tribus (1961) and sect. 4.1 in the present article). Entropy, in this information-theoretic rendition, is the long-term average of surprisal, and reducing information-theoretic free energy amounts to improving the world model so as to reduce prediction errors, hence reducing surprisal[14] (since better models make better predictions). The overarching rationale (Friston 2010) is that good models help us to maintain our structure and organization, hence (over extended but finite timescales) to appear to resist increases in 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 & 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 expressions, if this story is correct, of this single principle operating at varying time-scales.

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. Essentially, the hope is to illuminate the very possibility of self-organization in biological systems (see, e.g., Friston 2009, p. 293). A full assessment of the free energy principle is, however, far beyond the scope of the present treatment.[15] In the remainder of this article, 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 demands the success of some mutually supportive stack of states of a generative model (recall sect. 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 needs to be passed forward through the system is the error signal, which is what remains once predictions and driving signals have been matched.[16] In these models it is therefore the backward (recurrent) connectivity that carries the main information processing load. We should not, however, 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 et al. 2010, p. 8702)

It is potentially misleading because this stresses only one aspect of what is (at least in context of the rather specific models we have been considering[17]) 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 downward 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 that of Desimone & Duncan 1995) that posit top-down *enhancements* of selected aspects of the sensory signal, because:

> 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 [hence actually sharpens responses among the laterally competing representations]. (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.[18] Thus Spratling, in a recent formal treatment of this issue,[19] suggests that any apparent contrast here reflects:

> A misinterpretation of the model that may have resulted from the strong emphasis the predictive coding hypothesis places on the *error-detecting nodes* and the corresponding *underemphasis on the role of the prediction nodes in maintaining an active representation of the stimulus.* (Spratling 2008a, 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 backward 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 made possible, it should be stressed, only by positing the existence of "two functionally distinct subpopulations, 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 forward, while deep pyramidal cells play the role of representation units, passing predictions (made on the basis of a complex generative model) downward (see, e.g., Friston 2005; 2009; Mumford 1992). 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.

2.2. Encoding, inference, and the "Bayesian Brain"

Neural representations, should the hierarchical predictive processing account prove correct, encode probability density distributions[20] in the form of a probabilistic generative model, 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 that 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 et al. 2003), and relative timing effects (Deneve 2008) – for a very useful review, see Vilares & Körding 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," namely, 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" (Knill & Pouget 2004, p. 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. At first, the system may avoid committing itself to any single interpretation, while confronting 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 favors 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 & Bar 2009) followed by the details. This affords a kind of "forest first, trees second" approach (Friston 2005, p. 825; Hochstein & Ahissar 2002).

This does not mean, however, that context effects will always take time to emerge and propagate downward.[21] 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. This is especially important in the crucial range of cases where we, by our own actions, help to bring the new stimulus about. 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.[22] 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, colors, 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 & Pouget 2004, p. 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 sect. 3.1) evidence that biological systems approximate, in multiple domains, the Bayesian profile thus understood. To take just one example (for others, see sect. 3.1) Weiss et al. (2002) – in a paper revealingly titled "Motion illusions as optimal percepts" – used an optimal Bayesian estimator (the "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.[23] 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 & Pouget [2004] for a balanced review).[24] 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 & 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 (Berniker & Körding 2008; Körding et al. 2007; Yu 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 & Mamassian 2009) yield the same behavioral repertoire as a "Bayes' optimal" system. Nonetheless, the hierarchical and bidirectional 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 behavioral evidence (again, see sect. 3.1) that such a mechanism is somehow implemented in the brain. Experimental tests have also recently been proposed (Maloney & Mamassian 2009; Maloney & Zhang 2010) 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 thought and investigation.

Hierarchical predictive processing models also suggest something about the nature of the representation relation itself. To see this, recall (sect. 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 there is organism-detectable regularity in space or time; see Hosoya et al. 2005; Schwartz et al. 2007) allows a learner reliably to match its internal generative model to the statistical properties of the signal source (the world) yielding contents that are, I submit, as "grounded" (Harnad 1990) and "intrinsic" (Adams & 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 bidirectional 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 model providing the "top-down" predictions is 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 it 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 change 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., McClelland & Rumelhart 1981; Rumelhart 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*.[25] 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 100 msec) according to the presence or absence of apt expectations, and that the neural (here, EEG) signatures of conscious perception vary accordingly – a result these authors 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 et al. (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 et al. 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" (Egner et al. 2010, 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 as 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. (2011) suggest that on such accounts:

> 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. 2011, p. 1249)

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 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 & Frith [2009, pp. 53–54] who refer the reader to work on prediction error and the mesolimbic dopaminergic system such as Holleman & 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 et al. 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 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. Successful perception requires the brain to minimize surprisal. But the agent is able to see very (agent-) surprising things, at least in conditions where the brain assigns high reliability to the driving signal. Importantly, that requires that other high-level theories, though of an initially agent-unexpected kind, win out so as to reduce surprisal by explaining away the highly weighted sensory evidence. In extreme and persistent cases (more on this in sect. 4.2), this may require gradually altering the underlying generative model itself, 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 internal models that are making the predictions. Top-level (more "cognitive") models[26] 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, often 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 sect. 4.2) are mutually, reciprocally, entrenching.

2.4. 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 modeling and Bayesian theory. It thus offers what is arguably the first truly systematic bridge[27] linking 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 (sect. 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 it 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" (sect. 2.2) suggests. Good examples here include compelling bodies of work on cue integration (see also sects. 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 & Banks 2002; Knill & 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 that of 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 slower and smoother motions – see Weiss et al. (2002) and Ernst (2010). Another example is the Bayesian treatment of color perception (see Brainard 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 sect. 4.4) is impossible to doubt. It is thus a major virtue of the hierarchical predictive coding account that it effectively implements a computationally tractable version of the so-called Bayesian Brain Hypothesis (Doya et al. 2007; Knill & Pouget 2004; see also Friston 2003; 2005; and comments in sects. 1.2 and 2.2 above). But behavioral demonstrations of Bayesian 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 non-classical receptive field effects (Rao & Sejnowski 2002). 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 it 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 Sejnowski (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 & Ballard 1999, based on the simulation study sketched in sect. 1.2) 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 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 msec) 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 (sect. 1.1) is in full evidence because:

> 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 & 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.[28] 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 as well? 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.[29] Finally, the study by Egner et al. (2010; described in sect. 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 (see sect. 2.1 earlier) 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, p. 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 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 et al. 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, sect. 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., Born et al. 2009; Pack & Born 2001).

One extremely important and as yet not well-tested implication of the general architectural form of these models is (recall sect. 2.1) that each level of processing should contain two functionally distinct sub-populations of units. 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 backward connections are free to construct (in a potentially much more complex, and highly nonlinear 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 & 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 (surprisal). Crucially, there also exists (sect. 2.1) 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.[30]

3.2. Scope and limits

According to Mumford:

> 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 an intriguing footnote, Mumford then adds:

> 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, Note 5)

This remark highlights 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 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 2010; Friston et al. 2009) 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 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 (2011a, 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 & Stephan 2007, pp. 435–36). Instead, different organisms amount (Friston 2011a) 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 behavior (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 & 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 that, as argued by Eliasmith (2007), 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 (twenty-first century replay)

Back in the late 1970s and early 1980s (the heyday of classical Artificial Intelligence [AI]) there was a widely held view that two personality types were reflected in theorizing about the human mind. These types were dubbed, by Roger Schank and Robert Abelson, the "neats" versus the "scruffies."[31] 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, for example, Clark (1987) and Marcus (2008). Neats favored logic and provably correct solutions, while scruffies favored 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 favoring "structured probabilistic approaches" and those favoring "emergentist" approaches (where these are essentially connectionist approaches of the parallel distributed processing variety).[32]

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 favored 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 (sects. 2.2 and 2.3) 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 (sect. 2.2) that the brain 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 labors 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 sect. 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; also see sect. 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 (Clark 1997; 2008; Clark & Chalmers 1998; Haugeland 1998; Hurley 1998; Hutchins 1995; Menary 2007; Noë 2004; 2009; Rowlands 1999; 2006; Thelen & Smith 1994; Wheeler 2005; Wilson 1994; 2004). At least on the face of it, the predictive processing story seems to pursue a rather narrowly neurocentric focus, albeit one that reveals (sect. 1.5) 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 (sects. 1.5 and 1.6) 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 matching inputs with prior expectations. Action reduces surprisal by altering the world (including moving the body) so that inputs 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 behavior 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 (e.g., 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. (2010) translate these themes directly into the present framework using robotic simulations in which the learning of complex coordination dynamics is achieved by maximizing the amount of predictive information present in sensorimotor loops.

Extensions into the realm of social action and multi-agent coordination 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, e.g., recent work on synchronization and shared musical experience: Overy & Molnar-Szakacs 2009; and the "culture as patterned practices" approach suggested by Roepstorff et al. 2010). 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 color-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, smartphones, and, as Pickering & Garrod (2007) have observed, conversations with other agents. (For some intriguing speculations concerning the initial emergence of all those discrete symbols in predictive, probabilistic contexts, see König & 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 2006a; 2008), and one now bearing highly informative statistical relations to other such linguaform perceptibles.[33] 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 & Gasser 2005) and unfolding of human thought and reason can hardly be overestimated.

Such culturally mediated processes may incur costs (sect. 3.3) 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,[34] 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 intergenerationally transmissible structure that surrounds our biological brains (see, e.g., Griffiths & Gray 2001; Iriki & Taoka 2012; Oyama 1999; Sterelny 2007; Stotz 2010; Wheeler & Clark 2009).

What are the potential effects of such stacked and transmissible 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 et al. 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 (sects. 3.2 and 3.3) 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 sect. 1.6) 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 – imagine, for example, the sudden unveiling of a large and doleful elephant elegantly smuggled onto the stage by a professional magician.

The two perspectives are, however, easily reconciled. The large and doleful elephant is best understood as improbable but not (at least not in the relevant sense – recall sect. 3.2) surprising. Instead, that 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 sensory 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 doleful elephant 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. Nonetheless, systemic priors did not render that percept very likely in advance, hence (perhaps) the value to the agent of the feeling of surprise.

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 learned 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.[35] 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[36] 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 to the deployment of an implicit formal hyperprior (formal, because it concerns the form of the probabilistic representation itself) to the effect that our uncertainty can be described using such a unimodal probability distribution. Such a prior 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 bidirectional 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.[37] 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.[38]

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 (Corlett et al. 2009a; Fletcher & Frith 2009).

Recall the unexpected sighting of the elephant described 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 doleful, gray presence. 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 "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 (sect. 2.3). 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 important – 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 et al. 2009a; Corlett 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.[39] 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 (Corlett et al. 2009a; 2009b; Fletcher & Frith 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 – because 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.

Such a process is 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 & Frith 2009, p. 348). False perceptions and bizarre beliefs thus form an epistemically insulated self-confirming cycle.[40] This, then, is the dark side of the seamless story (sect. 2) 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 et al. 2009a) to help make sense of the ways 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 et al. 2009a, pp. 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.[41]

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, multilevel 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 (Churchland 1989; 2012) 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 in which each neural population targets the activity patterns displayed by the neural population below. What is crucial here – what makes such models *generative* as we saw in section 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[42] 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 (Kosslyn 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.[43] 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.[44]

Hierarchical predictive processing also provides a mechanism that explains a variety of important phenomena that characterize sensory perception, such as cross- and multimodal context effects on early sensory processing. Murray et al. (2002) displayed (as noted in sect. 3.1) 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 & 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 multimodal and cross-modal effects (including various kinds of "filling-in") will follow. For any statistically valid correlations, registered within the increasingly information-integrating (or "meta-modal" – Pascual-Leone & 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[45]) 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, whereas 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.

Such differences, it seems to me, should be enough to ground the obvious experiential differences between the various modalities. At the same time, 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., Barrett 2009; Clark 1989; Dennett 1978; 1987) 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 folk-psychological constructs: constructs which we encounter and model just as surely as we encounter and model other constructs such as 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 downward, 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 et 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 concerning what is out there. This is a challenging vision, as 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.[46] 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 result 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." 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 (sect. 1.2) by individual learning and evolutionary inheritance so as to reflect and register relevant aspects of the causal structure of the world itself.[47]

One place where this becomes especially evident is in the treatment (sect. 2.2) 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 et al. 2005), ventriloquism effects (Alais & 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 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.[48]

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: 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 sect. 2.1) functionally distinct "error units" and "representation units." The representation units that communicate

predictions downward do indeed encode increasingly complex and more abstract features (capturing context and regularities at ever-larger spatial and temporal scales) in the processing levels 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 and integration. Different levels learn and deploy different sets of predictions, corresponding to different bodies of knowledge, aimed at the level below (specialization) but the system settles in a way largely determined by the overall 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 (integration).[49]

A second source of potential radicalism lies with the suggestion (sect. 1.5) 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 et al. 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 by 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 & Daw 2012), that would not immediately justify our 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 2011b).

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, p. 296) 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 & Bar 2009; see also sect. 2.2) 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 it is 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.[50]

5.2. Conclusions: Towards a grand unified theory of the mind?

Action-oriented predictive processing models come tantalizingly 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.[51] 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 multilevel 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 twenty-first century.

ACKNOWLEDGMENTS

This target article has benefitted enormously from comments and reactions from a wide variety of readers and audiences. Special thanks are due to the BBS referees, who provided an especially rich and challenging set of comments and suggestions. The present incarnation of this article owes a great deal to their patient and extensive help and probing. Thanks also to Karl Friston, Jakob Hohwy, Tim Bayne, Andreas Roepstorff, Chris Thornton, Liz Irvine, Matteo Colombo, and all the participants at the *Predictive Coding Workshop* (School of Informatics, University of Edinburgh, January 2010); to Phil Gerrans, Nick Shea, Mark Sprevak, Aaron Sloman, and the participants at the first meeting of the *UK Mind Network* held at the Faculty of Philosophy, Oxford University, March 2010; to Markus Werning, and the organizers and participants of the 2010 meeting of the *European Society for Philosophy and Psychology*, held at Ruhr-Universität Bochum, August 2010; to Nihat Ay, Ray Guillery, Bruno Olshausen, Murray Sherman, Fritz Sommer, and the participants at the *Perception & Action Workshop*, Santa Fe Institute, New Mexico, September 2010; to Daniel Dennett, Rosa Cao, Justin Junge, and Amber Ross (captain and crew of the hurricane-Irene-blocked 2011 *Cognitive Cruise*); to Miguel Eckstein, Mike Gazzaniga, Michael Rescorla, and the faculty and students at the *Sage Center for the Study of Mind*, University of California, Santa Barbara, where, as a Visiting Fellow in September 2011, I was privileged to road-test much of this material; and to Peter König, Jon Bird, Lee de-Wit, Suzanna Siegel, Matt Nudds, Mike Anderson, Robert Rupert, Bill Phillips, and Rae Langton. A much earlier version of some of this material was prepared thanks to support from the AHRC, under the ESF Eurocores CONTACT (Consciousness in Interaction) project, AH/E511139/1.

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 multilayer learning device, or "deep architecture" (Hinton 2002; 2007b; 2010; Hinton & Salakhutdinov 2006; Hinton et al. 2006; for reviews, see Bengio 2009; Hinton 2007a).

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 strategy known as predictive coding. Rather, it is the use of that strategy in the special context of hierarchical systems deploying probabilistic 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 multilayer cascade.

6. In what follows, the notions of prior, empirical prior, and prior belief are used interchangeably, given the assumed context of a hierarchical model.

7. Because these proposals involve the deployment of top-down probabilistic generative models within a multilayer 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 nonhuman 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).

8. 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 downward-flowing predictions – see Friston (2005), Hohwy (2007), and discussion in section 2.5 of the present article. Thanks to one of the BBS reviewers for this helpful terminological suggestion.

9. Notice that an error signal thus construed is highly informative, and in this respect it differs from the kinds of error signal familiar from control theory and systems engineering. The latter are mostly simple signals that represent the amount of error/mismatch. The former ("prediction error signals") are much richer and carry information not just about the quantity of error but (in effect) about the mismatched content itself. It is in this sense that the residual errors are able, as it is sometimes said (Feldman & Friston 2010) to stand in for the forward flow of sensory information itself. Prediction errors are as structured and nuanced in their implications as the predictions relative to which they are computed. (Thanks to an anonymous BBS referee for suggesting this important clarification).

10. 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).

11. 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 sect. 2.3) 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 & Tong 2004).

12. This is also known (see, e.g., Friston et al. 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, p. 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 sect. 5.1).

13. 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).

14. The most fundamental aspect of the appeal to free energy, Friston claims, is that it provides an organismically computable window on surprise (i.e., surprisal) 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 sect. 1.2) an internally calculable quantity. This, however, is not a feature I will attempt to explore in the present treatment.

15. For an interesting critique of the most ambitious version of the free energy story, see section 5.1 in Gershman and Daw (2012).

16. This kind of efficiency, as one of the BBS referees nicely noted, is something of a double-edged sword. For, the obvious efficiencies in forward processing are here bought at the cost of the multilevel generative machinery itself: machinery whose implementation and operation requires a whole set of additional connections to realize the downward swoop of the bidirectional hierarchy. The case for predictive processing is thus not convincingly made on the basis of "communicative frugality" so much as upon the sheer power and scope of the systems that result.

17. In personal correspondence, Lee de-Wit notes that his usage follows that of, for example, Murray et al. (2004) and Dumoulin and Hess (2006), 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 sect. 5.1) 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).

18. 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's (2001) – for a review, see Smith and Ratcliff (2004). For an attempt to do so, see Hesselmann et al. (2010).

19. In this (2008a) 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.

20. For an early occurrence of this proposal in the literature of cognitive neuroscience, see Anderson and Van Essen (1994). That treatment also anticipates (although it does not attempt to model) the crucial role of top-down expectations and dynamic forms of Bayesian inference.

21. Thanks to one of the BBS reviewers for suggesting this important nuance to the temporal story.

22. 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 Kveraga et al. (2007), Bar (2007), Barrett and Bar (2009), and Fabre-Thorpe (2011).

23. 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 et al. (2006).

24. Some of the earliest work depicting perception and perceptual illusions as involving Bayesian inference is that of Hans-Georg Geissler, working in the 1970s in East Germany. This work, unfortunately, was not widely known outside the DDR (Deutsche Demokratische Republik) but see, for example, Geissler (1983; 1991).

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

26. Technically, there is always a single hierarchical generative model in play. In speaking here of multiple internal models, I mean only to flag that the hierarchical structure supports many levels of processing which distribute the cognitive labor by building distinct "knowledge structures" that specialize in dealing with different features and properties (so as to predict events and regularities obtaining at differing temporal and spatial scales).

27. The clear lineage here is with work in connectionism and recurrent artificial neural networks (see, e.g., Rumelhart 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 sect. 1.1) while making explicit contact with both Bayesian theorizing and contemporary neuroscientific research and conjecture.

28. Such effects have long been known in the literature, where they emerged in work on sensory habituation, and most prominently in Eugene Sokolov's pioneering studies of the orienting reflex. Sokolov concluded that the nervous system must learn and deploy a "neuronal model" that is constantly matched to the incoming stimulus, since even a *reduction* in the magnitude of some habituated stimulus could engage "dishabituation" and prompt a renewed response. See Sokolov (1960). See also Bindra (1959), Pribram (1980), and Sachs (1967). Here and elsewhere I am extremely grateful to one of the BBS referees, whose extensive knowledge of the history of these ideas has greatly enriched the present treatment.

29. For an excellent discussion of this recent work, see de-Wit et al. (2010).

30. 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.

31. 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.

32. 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 sect. 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 because 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) and Lee (2010).

33. We glimpse the power of the complex internal statistical relationships enshrined in human languages in Landauer and colleagues' fascinating work on "latent semantic analysis" (Landauer & Dumais 1997; Landauer et al. 1998). 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 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.

34. For example, Stanislas Dehaene's (2009) "neural recycling" account of the complex interplay between neural precursors, cultural developments, and neural effects within the key cognitive domains of reading and writing.

35. 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).

36. As helpfully pointed out by one of the BBS referees.

37. 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.

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

39. 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 et al. 2010; see also Friston & 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 sect. 2.3 – 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.

40. There are probably milder versions of this everywhere, both in science (Maher 1988) and in 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.

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

42. 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.

43. 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 downward pressure the active higher-level encoding exerts.

44. Common features include the appeal to forward models and the provision of mechanisms (such as Kalman filtering – see Friston 2002; Grush 2004; Rao & Ballard 1999) for estimating uncertainty and (thus) flexibly balancing the influence of prior expectations and driving sensory inputs. Indeed, Grush (2004, 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" (see section 5.2 of Grush 2004) 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 prominence 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.

45. 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 technologies 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).

46. Thanks to Susanna Siegel for useful discussion of this point.

47. For some further discussion, see Friston (2005, p. 822).

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

49. 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).

50. The empirical bet is thus, as Egner and colleagues 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).

51. When brought under the even-more-encompassing umbrella of the "free energy principle" (sect. 1.6), 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., Di Paolo 2009; Thompson 2007; Varela et al. 1991).

Open Peer Commentary

The problem with brain GUTs: Conflation of different senses of "prediction" threatens metaphysical disaster

doi:10.1017/S0140525X1200221X

Michael L. Anderson[a] and Tony Chemero[a,b]

[a]*Department of Psychology, Franklin & Marshall College, Lancaster, PA 17604-3003;* [b]*Departments of Philosophy and Psychology, University of Cincinnati, Cincinnati, OH 45221.*

michael.anderson@fandm.edu **http://www.agcognition.org**
tony.chemero@fandm.edu **http://edisk.fandm.edu/tony.chemero**

Abstract: Clark appears to be moving toward epistemic internalism, which he once rightly rejected. This results from a double over-interpretation of predictive coding's significance. First, Clark argues that predictive coding offers a Grand Unified Theory (GUT) of brain function. Second, he overreads its epistemic import, perhaps even conflating causal and epistemic mediators. We argue instead for a plurality of neurofunctional principles.

The predictive coding model of brain function is a deeply important development for neuroscience, and Andy Clark does the field a service with this careful, thorough, and accessible review. We are concerned, however, that Clark's account of the broad implications of model – and in particular his attempt to turn it into a Grand Unified Theory (GUT) of brain function – may be at least four dogmas of empiricism out-of-date (Anderson 2006; Chemero 2009; Davidson 1974; Quine 1951). Clark's adoption of a thoroughgoing inferential model of perception, his neo-neo-Kantian view of the relationship between mind and world, and his insistence that every sensory modality operates according to the same underlying causal-epistemic logic – all (individually and severally) threaten to return us to the bad old days of epistemic internalism (e.g., Rorty 1979) that the field, including the author of *Being There* (Clark 1997), rightly left behind.

Here we suggest that Clark (although not he alone) has made an error in conflating different senses of "prediction" that ought to be kept separate. The first sense of "prediction" (henceforth prediction$_1$) is closely allied with the notion of correlation, as when we commonly say that the value of one variable "predicts" another (height predicts weight; education predicts income; etc.). Prediction$_1$ is essentially model-free, and it comes down to simple relationships between numbers. The second sense of "prediction" (prediction$_2$), in contrast, is allied instead with abductive inference and hypothesis testing. Prediction$_2$ involves such cognitively sophisticated moves as inferring the (hidden) causes of our current observations, and using that hypothesis to predict future

observations, both as we passively monitor and actively intervene in the world. It is theory laden and model-rich.

We have no trouble believing that a fundamental part of our exquisite attunement to environmental contingencies involves sensitivity to (and the ability to make use of) inter- and cross-modal correlations in sensory signals. Sensitivity to temporal and spatial (e.g., across the retina) correlations could underwrite many functional advantages, including the ones Clark highlights, such as reducing sensory bandwidth and drawing attention to salient departures from expectations. In this sense we share Clark's belief that predictive$_1$ coding is likely to be a ubiquitous and fundamental principle of brain operation; neural nets are especially good at computing correlations.

However, we don't think that evidence for predictive$_1$ coding warrants a belief in predictive$_2$ coding. And it is only from predictive$_2$ coding that many of Clark's larger implications follow.

Clark makes the move from predictive$_1$ coding to predictive$_2$ coding largely by relying on an innovative account of binocular rivalry offered by Hohwy et al. (2008). In Clark's somewhat simplified version of their proposal, the experienced alternation between seeing the face stimulus presented to one eye and the house stimulus presented to the other is explained by a knowledge-driven alternation between rival hypotheses (face at location *x*, house at location *x*) neither of which can account for all of the observations. According to Clark, the reason the images don't fuse and lead to a visual steady-state is because we *know* that faces and houses can't coexist that way. If this knowledge-driven account is the correct way to understand something as perceptually basic as binocular rivalry, then predictive$_2$ coding can begin to look like a plausible, multilevel and unifying explanation of perception, action and cognition: perception is cognitive and inferential; inference perceptual; and all of it is active.

But while the predictive$_2$ coding model of binocular rivalry may be consistent with much of the data, it is far from the only possible explanation of the phenomenon. Here is an outline of a reasonable predictive$_1$ coding account: Given the generally high-level of cross-correlation in the inputs of our two eyes, the left eye signal would predict$_1$ greater correlation with the right eye than is currently in evidence; this would weaken the inputs associated with the left eye, unmasking the inputs associated with the right eye, which would predict$_1$ cross-correlated left eye signals . . . and so on. However far this particular proposal could be taken, the point is one can account for the phenomenon with low-level, knowledge-free, redundancy-reducing inhibitory interactions between the eyes (see, e.g., Tong et al. 2006). After all, binocular rivalry also occurs with orthogonal diffraction gratings, indicating that high-level knowledge of what is visually possible needn't be the driver of the visual oscillation; humans don't have high-level knowledge about the inconsistency of orthogonal gratings. In general, although not every pair of stimuli induce bistable perceptions, the distinction between those that do and those that don't appears to have little to do with knowledge (see Blake [2001] for a review). Adopting a predictive$_2$ coding account is a theoretical choice not necessitated by the evidence. It is hardly an inconsequential choice.

Using predictive$_2$ coding as a GUT of brain function, as Clark proposes, is problematic for several reasons. The first problem is with the very idea of a grand unified theory of brain function. There is every reason to think that there can be no grand unified theory of brain function because there is every reason to think that an organ as complex as the brain functions according to diverse principles. It is easy to imagine knowledge-rich predictive$_2$ coding processes employed in generating expectations that we will confront a jar of mustard upon opening the refrigerator door, while knowledge-free predictive$_1$ coding processes will be used to alleviate the redundancy of sensory information. We should be skeptical of *any* GUT of brain function. There is also a problem more specific to predictive$_2$ coding as a brain GUT. Taking all of our experience and cognition to be the result of high-level, knowledge-rich predictive$_2$ coding makes it seem as if the world that we experience and think about is a projection of our minds. Western philosophy has been down this lonely and unproductive road many times. It would be a shame if the spotlight that Clark helpfully shines on this innovative work in neuroscience were to lead us back there.

Attention and perceptual adaptation

doi:10.1017/S0140525X12002245

Ned Block[a] and Susanna Siegel[b]

[a]*Department of Philosophy, New York University, New York, NY 10003;* [b]*Department of Philosophy, Harvard University, Cambridge, MA 02138.*

ned.block@nyu.edu **ssiegel@fas.harvard.edu**

http://www.nyu.edu/gsas/dept/philo/faculty/block/

http://www.people.fas.harvard.edu/~ssiegel/

Abstract: Clark advertises the predictive coding (PC) framework as applying to a wide range of phenomena, including attention. We argue that for many attentional phenomena, the predictive coding picture either makes false predictions, or else it offers no distinctive explanation of those phenomena, thereby reducing its explanatory power.

According to the predictive coding view, at every level of the visual/cortical hierarchy, there are two kinds of units: error units and representation units. Representations propagate downward in the visual hierarchy whereas error signals propagate upward. Error in this sense might be better called "discrepancy," since it is the discrepancy between what the visual system predicts (at a given level) and what is represented at that level. Clark advertises the predictive coding (PC) framework as applying to a wide range of phenomena, including attention, which Clark says "is achieved by altering the gain (the 'volume,' to use a common analogy) on the error-units" (sect. 2.3, para. 6). We argue that for many attentional phenomena, the predictive coding picture either makes false predictions, or else it offers no distinctive explanation of those phenomena, thereby reducing its explanatory power.

Consider a basic result in this area (Carrasco et al. 2004), which is that attention increases perceived contrast by enhancing "the representation of a stimulus in a manner akin to boosting its physical contrast" (Ling & Carrasco 2006, p. 1243). A cross-modal study using auditory attention-attractors (Störmer et al. 2009) showed that the contrast-boosting effect correlated with increased activity in early stages of visual processing that are sensitive to differences in contrast among stimuli. The larger the cortical effect, the larger the effect on perceivers' judgments. Increasing the contrast of a stimulus has an effect on the magnitude of perceptual adaptation to that stimulus, causing greater threshold activation in the tilt after-effect and longer recovery time. Ling and Carrasco (2006) showed that attending to a stimulus while adapting to that stimulus has the same effect as increasing the contrast of the adapting stimulus. After attending to the adaptor (70% contrast), the contrast sensitivity of all observers was equivalent to the effect of adapting to a 81–84% contrast adaptor.

How do these results look from a PC perspective? Suppose that at time t_1, the perceiver is not attending to the left side of space but nonetheless sees a striped grid on the left with apparent contrast of 70%. Because there is no movement or other change, at time t_2, the visual system predicts that the patch will continue at 70%. But at t_2 the perceiver attends to the patch, raising the apparent contrast to, say, 82%. Now at t_2 there is an error, a discrepancy between what is predicted and what is "observed." Since the PC view says attention is turning up the volume on the error representations, it predicts that at t_3 the signal (the represented contrast) should rise even higher than 82%. But that does not happen.

There are two important lessons. First, the initial changes due to attending come before there is an error (at t_2 in the example),

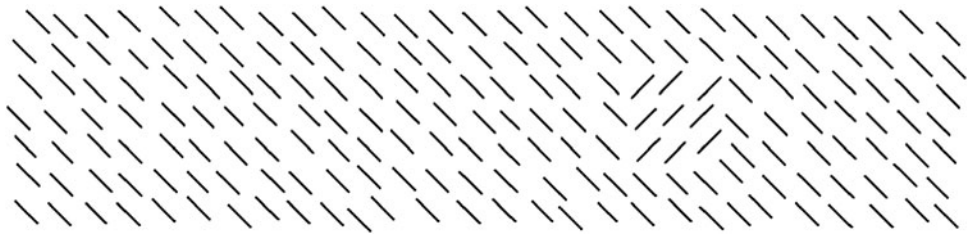

Figure 1 (Block & Siegel). A display of one of the textured figures (the square on the right) used by Yeshurun and Carrasco (1998). The square appeared at varying degrees of eccentricity. With low resolution in peripheral locations, attention improved detection of the square; but with high resolution in central locations, attention impaired detection.

so the PC viewpoint cannot explain them. Second, the PC view makes the false prediction that the changes due to attending will be magnified.

Sometimes PC theorists assume the error signal is equal to the input. Perhaps this identification makes some sense if the perceiver's visual system has no "expectations," say because the eyes have just opened. But once the eyes have opened and things in the environment are seen, it makes no sense to take the error signal to be the sensory input.

The PC picture also seems to lack a distinctive explanation of why attention increases spatial acuity. Yeshurun and Carrasco (1998) showed that increased attention can be detrimental to performance when resolution was already on the border of too high for the scale of the texture, increasing acuity to the point where the subject does not see the forest for the trees. Too little attention can also be detrimental, making it harder to see the trees. Yeshurun and Carrasco varied resolution of perception by presenting textured squares (such as the one in Fig. 1) at different eccentricities (the more foveal, the better the resolution). But they also varied resolution by manipulating the focus of spatial attention: With the eyes focused at the center, they attracted attention to the left or to the right. Combining contributions to resolution from eccentricity and attention, they found that there was an optimal level of resolution for detecting the square, with detection falling off on both ends. Single cell recordings in monkey visual cortex reveal shrinking receptive fields (the area of space that a neuron responds to) in mid-to-high level vision, specifically in V4, MT, and LIP, and this shrinkage in receptive fields is a contributor to explaining the increase in acuity (Carrasco 2011).

Does the PC framework have a distinctive explanation of attentional effects on spatial acuity, in terms of "gain in error-units"? If, due to the level of acuity, one does not see the square, then the prediction of no square will be confirmed, and there will be no discrepancy ("error") to be magnified. Since the gain in error units is the only distinctive resource of the PC view for explaining attentional phenomena, the view seems to have no distinctive explanation of this result either. Can the predictive coding point of view simply borrow Carrasco's explanation? That explanation is a matter of shrinkage in receptive fields of neurons in the *representation nodes*, not anything to do with prediction error, so the predictive coding point of view would have to concede that attention can act directly on representation nodes without a detour through error nodes.

Finally, attention to certain items – for example, random dot patterns – makes them appear larger. Anton-Erxleben et al. (2007) showed that the size of the effect is inversely related to the size of the stimulus, explaining the result in terms of receptive field shift (such shifts are also observed from single cell recordings in monkey visual areas; Womelsdorf et al. 2006). This explanation depends on the retinotopic and therefore roughly spatiotopic organization common to many visual areas – not on error units. Neurons whose receptive fields lie on the periphery of the pattern shift their receptive fields so as to include the pattern, moving the portion of the spatiotopically represented space to include the pattern, resulting in the representation of the pattern as occupying a larger area. Here too, predictive coding offers no distinctive explanation.

The facts of attention and adaptation do not fit well with the predictive coding view or any picture based on how "sensory neurons should behave" (Lochmann et al. 2012) rather than the facts of how they do behave. Without a distinctive explanation of these facts, the explanatory promises of predictive coding are overdrawn.

Attention is more than prediction precision

doi:10.1017/S0140525X12002324

Howard Bowman,[a] Marco Filetti,[a] Brad Wyble,[b] and Christian Olivers[c]

[a]*Centre for Cognitive Neuroscience and Cognitive Systems, and the School of Computing, University of Kent at Canterbury, Kent CT2 7NF, United Kingdom;* [b]*Department of Psychology, Syracuse University, Syracuse, NY 13244;* [c]*Department of Cognitive Psychology, Faculty of Psychology and Education, VU University Amsterdam, 1081 BT Amsterdam, The Netherlands.*

H.Bowman@kent.ac.uk M.Filetti@kent.ac.uk
bwyble@gmail.com c.n.l.olivers@vu.nl
http://www.cs.kent.ac.uk/people/staff/hb5/
http://www.cs.kent.ac.uk/people/rpg/mf266/
www.bradwyble.com http://olivers.cogpsy.nl

Abstract: A cornerstone of the target article is that, in a predictive coding framework, attention can be modelled by weighting prediction error with a measure of precision. We argue that this is not a complete explanation, especially in the light of ERP (event-related potentials) data showing large evoked responses for frequently presented target stimuli, which thus are predicted.

The target article by Andy Clark champions predictive coding as a theory of brain function. Perception is the domain in which many of the strongest claims for predictive coding have been made, and we focus on that faculty. It is important to note that there are other unifying explanations of perception, one being that the brain is a *salience detector*, with salience referring broadly to relevance to an organism's goals. These goals reflect a short-term task set (e.g., searching a crowd for a friend's face), or more ingrained, perhaps innate motivations (e.g., avoiding physical threat). A prominent perspective is, exactly, that one role of attention is to locate and direct perception towards, salient stimuli.

The target article emphasises the importance of evoked responses, particularly EEG event-related potentials (ERPs), in adjudicating between theories of perception. The core idea is that the larger the difference between an incoming stimulus and the prediction, the larger the prediction error and thus the larger the evoked response. There are indeed ERPs that are clearly modulated by prediction error, for example, the Mismatch Negativity (evoked by deviation from a repeating pattern of stimulus presentation), the N400 (evoked by semantic anomalies), and

P3 responses to oddball stimuli. In addition, stimuli that violate our expectations do often capture attention (Horstmann 2002), consistent with predictive coding. However, such surprise-driven orienting is just one aspect of attention, and we question whether prediction error provides an adequate explanation for attentional functioning as a whole.

A central aspect of attention, which makes perception highly adaptive, is that it can purposefully select and enhance *expected* stimuli. This arises when an arrow cues where a target will appear, or a verbal instruction indicates it will be red. However, in this context, ERPs are largest to the target stimuli (P1, N1, N2pc, P3; Luck 2006), in line with a saliency account. Such heightened responses to predicted stimuli do not seem to sit comfortably with predictive coding. As Clark highlights, resolution of this conundrum has, in analogy with statistical tests, focused on *precision* (Feldman & Friston 2010). The two-sample t-test, say, is a ratio of the difference of two means, and variability in the estimate of that difference. Precision-weighted prediction error is such a test: The difference between prediction and observation is weighted by the precision or confidence in that difference – that is, the inverse of variability, or, in other words, the signal fed back up the sensory pathway, the evoked response, is a precision-weighted prediction error. Importantly, attention is proposed to increase precision; that is, the brain has greater confidence in its estimate of disparity between predicted and observed when that observation is being spot-lit by attention, and, indeed, perception does seem more accurate in the presence of attention (Chennu et al. 2009). This then enables predictive coding to generate big bottom-up responses to expected, in the sense of attended stimuli, as simulated for spatial attention in (Feldman 2010).

Although predictive coding is an elegant and intriguing approach, obstacles remain to its being fully reconciled with the saliency perspective. First, precision-weighting has a multiplicative effect. Hence, there has to be a difference between *observed* and *predicted* in the first place for precision to work on. If *observed* is exactly as expected, however big precision might be, the precision-weighted prediction error will be zero. Yet classic EEG experiments show that attentional enhancement of ERP components (e.g., P1 and N1) is greatest when targets appear in the same location for many trials (Van Voorhis & Hillyard 1977). One could of course argue that there is always some error, and that the effects of attention on precision are extremely large relative to that error. However, depending upon the extent to which precision modulates the prediction error, one could obtain classically predictive or anti-predictive (i.e., salience sensitive) patterns, and both patterns are found experimentally. Thus, the theory really requires a computational explanation of how the modulatory effect of precision varies across experimental contexts, otherwise there is a risk that it becomes effectively unfalsifiable.

Second, prediction error is passed back up the sensory pathway so that parameters can be adjusted to improve predictions (i.e., learning), and the amount parameters change is a function of the size of the precision-weighted prediction error. This, however, raises a further problem with a big precision-weighted prediction error being generated through a large (attention-governed) precision, when *observed* and *predicted* are similar. Specifically, in this case, the parameters should not change and certainly not a lot, even though precision-weighted prediction error might mandate it.

Third, directing attention, and thus improving precision, at a pre-determined location is one thing. But what makes attention so adaptive is that it can guide towards an object at an unpredictable location – simply on the basis of features. For example, we could ask the reader to find the nearest word printed in bold. Attention will typically shift to one of the headers, and indeed momentarily increase precision there, improving reading. But this makes precision weighting a *consequence* of attending. At least as interesting is the mechanism *enabling* stimulus selection in the first place. The brain has to first deploy attention before a precision advantage can be realised for that deployment. Salience theory proposes that stimuli carrying a target feature become more salient and thus draw attention. But which predictive coding mechanism is sensitive to the match between a stimulus feature and the target description? In typical visual search experiments, observers are looking for, and finding, the same target in trial after trial. For example, in our rapid serial visual presentation experiments, each specific distractor appears very rarely (once or twice), while pre-described targets appear very frequently. We obtained effectively no evoked response for distractors but a large deflection for the target (see Fig. 1). It seems that predictive coding mandates little if any response for this scenario. If anything, should the distractors not have generated the greatest response, since they were (a) rare, and (b) not matching predictions?

Even if one could devise a predictive coding framework that allocated a higher precision to the target representation (which is a step beyond its spatial allocation in Feldman 2010), it is unclear how it could generate a massive precision-weighted prediction error

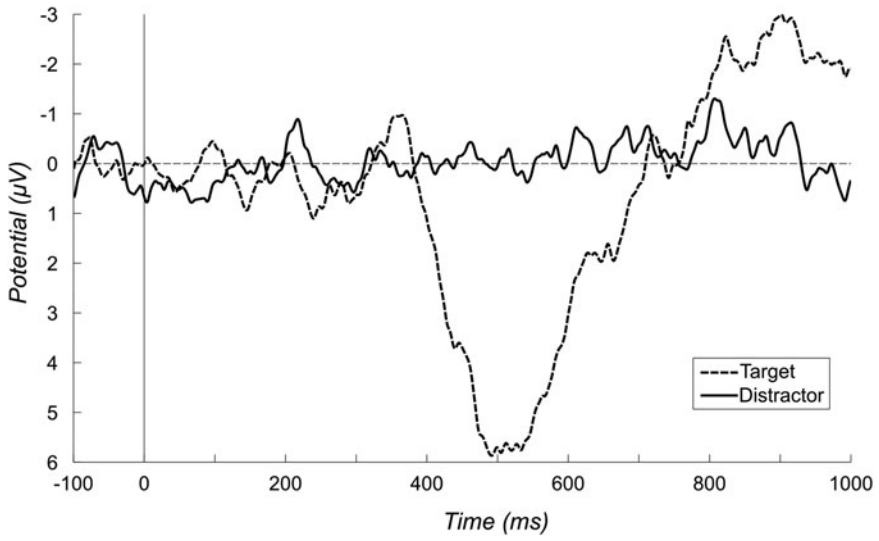

Figure 1 (Bowman et al.). An anti-predictive ERP pattern.

specifically for targets, where *predicted* and *observed* match exactly. It is also unclear why such an error is needed.

Applications of predictive control in neuroscience

doi:10.1017/S0140525X12002282

Bruce Bridgeman

Department of Psychology, University of California–Santa Cruz, Santa Cruz, CA 95064.

bruceb@ucsc.edu **http://people.ucsc.edu/~bruceb/**

Abstract: The sensory cortex has been interpreted as coding information rather than stimulus properties since Sokolov in 1960 showed increased response to an unexpected stimulus decrement. The motor cortex is also organized around expectation, coding the goal of an act rather than a set of muscle movements. Expectation drives not only immediate responses but also the very structure of the cortex, as demonstrated by development of receptive fields that mirror the structure of the visual world.

Prediction is a powerful principle in neuroscience, and it is not a new one. It has been central to interpretation of brain function since the influential work of E. N. Sokolov (1960) (see target article, Note 28). He found that cortical responses depend not on the amplitude of an incoming signal, but on its information value. An expected stimulus caused hardly a ripple, while an unexpected one triggered what Sokolov termed an orienting response. The key experiment was to repeat a stimulus until its cortical signal nearly disappeared (habituation of the orienting response, or Clark's "repetition suppression"). Then Sokolov decreased the stimulus amplitude or its duration. Sokolov reasoned that if the cortex were merely echoing stimulus properties the response should have decreased, but instead it increased. With a qualitative change, no amount of fussing with nonlinearities and thresholds could explain the result. The cortex was coding not stimulus properties but stimulus information, the difference between signal and expectation. In this context it is no wonder that we ignore and fail to remember most of the vast streams of signals emanating from our millions of sensory receptors. So Clark's prediction thesis has been the dominant interpretation of cortical sensory coding for more than a half-century.

Another insight that shaped neuroscience is that the brain is not about representing the stimulus; it is about organizing action. The evidence begins with an anatomical paradox that the precentral "motor" cortex is innervated by the dorsal thalamus, a region homologous to the dorsal spinal cord that processes sensory information (Pribram 1971, p. 241). Pribram asks why the motor cortex should be closely tied to an otherwise sensory structure. His answer is that the motor cortex is really a sensory cortex for an image of achievement, analogous to the images in sensory regions and organized similarly. Motor cortex codes environmental contingencies, not literal muscle movements, and continuously compares progress in execution of an act with its goal.

Similarly, it has long been known that receptive fields in sensory cortex are shaped not only by anatomy but also by experience, so that they encode best what is predicted to be present in the environment. I was privileged to witness the first evidence that sensory experience could tune the receptive field properties of the primary visual cortex (V1). Helmut Hirsch, then a Stanford graduate student, was studying kittens that he raised wearing masks that exposed one eye to vertical stripes and the other to horizontal stripes. Together with Nico Spinelli and Robert Phelps we began recording from single cells in V1 of the mask-reared kittens, using the first automated receptive-field mapping apparatus. We prepared our first kitten and dipped our microelectrode into its cortex.

The first cells we recorded had large, poorly defined receptive fields of the sort to expect in a visually deprived cat. Then around 10 p.m. we found a cell with a huge, vertically oriented receptive field. Perhaps it was an artifact, the bursting discharges of an injured cell as the mapping stimulus swept vertically across our screen. So we changed to a horizontal scan. The field remained, five times bigger than any oriented receptive field ever recorded from a cat. Our jaws dropped as we looked at each other, a moment of discovery – this wasn't a normal cortex, but something completely different. It was the magical moment in science when you know something about nature that no one else knows. We covered one eye, then the other; the receptive field disappeared and reappeared. Later that night we recorded several other similar fields, all vertical or horizontal, all monocular, and all huge. It turned out later that the receptive field orientations matched the mask orientations for the corresponding eye (Hirsch & Spinelli 1970). Plasticity in this cat's cortex extended beyond any mere selection of normal receptive fields, beyond anything that anyone had suspected. The cat had reorganized its cortex from visual experience alone. Clearly the cortex, by the structure of its receptive fields, was predicting future input.

This would be an interesting curiosity if not for its under-appreciated implication that the same process must be occurring in normal cats, and, by extension, in humans as well. Sensory receptive fields are tuned to the structure of the world that the animal encounters in its early experience. The receptive fields of normal animals have a 1/f statistical structure, as does the natural world.

It is even possible that the dominance of the foveal projection onto V1, a quarter of the entire surface in humans, is a consequence of the huge number of projections coming up from the periphery. The small size of V1 receptive fields representing the fovea might originate from the better optics and smaller convergence of the foveal anatomy. The distribution of receptive field orientations and spatial frequencies reflects the properties of the normal visual environment (Switkes et al. 1978); the cortex is predicting its own input by its very structure. This is precisely what Clark realizes when he concludes, "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" (sect. 3.4, para. 1). But the evidence has been there all along.

When the predictive brain gets it really wrong

doi:10.1017/S0140525X12002233

Gavin Buckingham and Melvyn A. Goodale

The Brain and Mind Institute, Natural Sciences Centre, The University of Western Ontario, London ON N6A 5B7, Canada.

gbucking@uwo.ca **mgoodale@uwo.ca**
http://publish.uwo.ca/~gbucking/
http://psychology.uwo.ca/faculty/goodale/

Abstract: Clark examines the notion of the "predictive brain" as a unifying model for cognitive neuroscience, from the level of basic neural processes to sensorimotor control. Although we are in general agreement with this notion, we feel that there are many details that still need to be fleshed out from the standpoint of perception and action.

In his target article, Clark paints a diverse picture of how prediction is a ubiquitous part of brain and behaviour interactions. Taking heavy cues from Friston's "free energy principle," his target article summarises ideas at the neural level, suggesting that the critical variable for sensory coding and motor control is the deviation from the expected signal, rather than the sensory or motor processing per se. In the field of sensorimotor control, this Bayesian approach is a popular one (e.g., Körding & Wolpert 2004). Many researchers have built their careers showing that, in a wide range of contexts, an individual's motor behaviour can be modeled as the approximately optimal combination of the "undiluted" sensory input and the prior probability

of that sensory event occurring, thus biasing the response one way or the other. Similarly, a wide range of psychophysical experiments have demonstrated that our conscious perception of events in the world represents not veridical sensory input, but the integration of multiple sources of evidence from our sensory system and our prior experience, rather than the veridical (and noisy) sensory input itself (Gregory 1998). An especially compelling case for this Bayesian standpoint can be made from the study of perceptual illusions, and several classic visual illusions can be explained with this optimal integration strategy (Geisler & Kersten 2002; Weiss et al. 2002). In these contexts, this integration is thought to overcome the noise in the system of our sensory organs, maximising the likelihood of perceptual or motor "success."

Despite the apparent descriptive power of optimally combining sensory prediction with sensory input, there are common situations where conscious perception is clearly not a product of Bayesian-style optimal integration. In fact, when we lift an object and experience its weight, our conscious perception of how heavy it feels is almost exactly the opposite of what might be expected if a perceiver integrates perpetual priors with sensory input. This incongruence is easily demonstrated with the famous size–weight illusion (SWI), first described in 1891 by Augustin Charpentier (translation by Murray et al. 1999). The SWI occurs when small and large objects, that otherwise look similar to one another, are adjusted to have identical weights. When individuals lift these objects, the small one feels substantially heavier than the (equally-weighted) larger one – an effect that is persistent and apparently cognitively impenetrable. The mechanism that underpins this illusion is still something of a mystery. It has long been contended (in a rather vague way) that the illusion is caused by the violation of an individual's expectations about how heavy each object will be – namely, the expectation that the large objects will outweigh the small objects (Ross 1969). It is not difficult to imagine how this prior is built up, given the consistency of the relationship between size and weight outside of the laboratory setting. It is repeatedly encountering this positive size/weight relationship throughout our entire lives that presumably serves to establish a very powerful prior for our perceptions of heaviness (Flanagan et al. 2008). Crucially, however, this prior is not integrated into the lifter's percept of how heavy the objects feel, as one might predict from a Bayesian optimal integration standpoint. Instead, the lifter's conscious perception of heaviness *contrasts* the prior expectation, leading some authors to label the effect as "anti-Bayesian" (Brayanov & Smith 2010). Variants of the SWI can even manifest in a single, unchanging, object, which can be made to feel different weights by simply manipulating an individual's expectations of what they are about to lift (Buckingham & Goodale 2010).

The functional significance of this contrastive effect has been the source of great (and largely unresolved) debate – why would our perceptual system be so stricken with errors? Extending the conclusions of a recent study by Baugh and colleagues (Baugh et al. 2012), it could be proposed that the SWI is a product of a perceptual system specialised for the detection and subsequent flagging of outliers in the statistics of the environment. Thus, conscious weight perception can be framed as an example of a task where it is important to emphasise the unexpected nature of the stimuli, in a system which presumably favours more efficient coding of information.

As lifting behaviour is a largely predictive process, our fingertip forces are driven by our expectations of how heavy something looks. And, in a more conventional Bayesian fashion, the weighting of these priors is rapidly adjusted (or rapidly ignored) by the presence of lifting errors. This provides the sensorimotor system with the best of both worlds – lifting behaviour that is flexible enough to rapidly adapt to constantly changing environments (e.g., a bottle of water which is being emptied by a thirsty drinker), but will automatically "snap back" to the (generally correct) lifting forces when the context of the lift is altered (so that the next time a fresh bottle of water is grasped, the sensorimotor prediction will have a good chance of being accurate). Thus, when lifting SWI-inducing cubes for the first time, lifters will apply excess force to the large cube and apply insufficient force to the small cube the first time they lift them, but will lift these two identically-weighted cubes with appropriately identical forces after only a few experiences with them (Flanagan & Beltzner 2000). Clearly, this adaptive behaviour is a consequence of a complex interaction between short-term and long-term priors (Flanagan et al. 2008) – a process that looks far more like the Bayesian processes outlined by Clark in his target article (Brayanov & Smith 2010). It is tempting to ascribe a causal relationship between the force errors and the perceptual ones. Remarkably, however, the two kinds of errors appear to be completely isolated from one another: The magnitude of the SWI remains constant from one trial to the next, even in the face of the rapid trial-to-trial adaptation of the gripping and lifting forces. This complicates the situation even further by suggesting that there must be independent sets of priors for motor control and perceptual/cognitive judgements, which ultimately serve quite different functions.

In conclusion, we have outlined how the deceptively simple SWI paradigm can uncover the operation of multiple priors operating simultaneously, with different weightings and different goals. It is worth noting, however, that while the predictive brain makes sense in a post-hoc way, providing a computationally plausible parameter for both the perceptual and lifting effects (Brayanov & Smith 2010), it is still very much a black-box explanation – and, to date, the term "prior" seems to serve only as a convenient placeholder in lieu of any tangible mechanism linking expectations to the perceptual or motor effects they appear to entail.

Expecting ourselves to expect: The Bayesian brain as a projector

doi:10.1017/S0140525X12002208

Daniel C. Dennett

Center for Cognitive Studies, Tufts University, Medford, MA 02155.
ddennett@tufts.edu
ase.tufts.edu/cogstud/incbios/dennettd/dennettd.htm

Abstract: Clark's essay lays the foundation for a Bayesian account of the "projection" of consciously perceived properties: The expectations that our brains test against inputs concern the particular affordances that evolution has designed us to care about, including especially expectations of our own expectations.

The "Bayesian" brain as a "hierarchical prediction machine" is an enticing new perspective on old problems, for all the reasons Clark articulates, ranging over fields as disparate as neuroanatomy, artificial intelligence, psychiatry, and philosophy; but he also catalogues some large questions that need good answers. While waiting for the details to come in, I want to suggest some other benefits that this perspective promises. If it turns out not to be sound, in spite of all the converging evidence Clark describes, we will have all the more reason for regret.

It is everybody's job – but particularly the philosophers' job – to negotiate the chasm between what Wilfrid Sellars (1962) called the *manifest image* and the *scientific image*. The manifest image is the everyday world of folk psychology, furnished with people and their experiences of all the middle-sized things that matter. The scientific image is the world of quarks, atoms, and molecules, but also (in this context particularly) sub-personal neural structures with particular roles to play in guiding a living body safely through life. The two images do not readily fall into registration, as everybody knows, leaving lots of room for confusion and compensatory adjustment (nicely exemplified by the surprise/surprisal pair).

Consider what I will call Hume's Strange Inversion (cf. Dennett 2009). One of the things in our world is causation, and we think we see causation because the causation in the world directly causes us to see it – the same way round things in daylight cause us to see round things, and tigers in moonlight cause us to see tigers. When we *see* the thrown ball causing the window to break, the causation itself is somehow perceptible "out there." Not so, says Hume. This is a special case of the mind's "great propensity to spread itself on external objects" (*Treatise of Human Nature*, Hume 1739/1888/1964, I, p. xiv). In fact, he insisted, what we do is misinterpret an inner "*feeling*," an anticipation, as an external property. The "customary transition" in our minds is the *source* of our sense of causation, a quality of "perceptions, not of objects," but we mis-attribute it to the objects, a sort of benign user-illusion, to speak anachronistically. As Hume notes, "the contrary notion is so riveted in the mind" (p. 167) that it is hard to dislodge. It survives to this day in the typically unexamined assumption that all perceptual representations must be flowing inbound from outside.

Here are a few other folk convictions that need Strange Inversions: sweetness is an "intrinsic" property of sugar and honey, which causes us to like them; observed intrinsic sexiness is what causes our lust; it was the funniness out there in the joke that caused us to laugh (Hurley et al. 2011). There is no more familiar and appealing verb than "project" to describe this effect, but of course everybody knows it is only metaphorical; colors aren't *literally* projected (as if from a slide projector) out onto the front surfaces of (colorless) objects, any more than the idea of causation is somehow beamed out onto the point of impact between the billiard balls. If we use the shorthand term "projection" to try to talk, metaphorically, about the mismatch between manifest and scientific image here, what is the true long story? What is literally going on in the scientific image? A large part of the answer emerges, I propose, from the predictive coding perspective.

Every organism, whether a bacterium or a member of *Homo sapiens,* has a set of things in the world that matter to it and which it (therefore) needs to discriminate and anticipate as best it can. Call this the ontology of the organism, or the organism's *Umwelt* (von Uexküll 1934/1957). This does not yet have anything to do with consciousness but is rather an "engineering" concept, like the ontology of a bank of elevators in a skyscraper: all the kinds of things and situations the elevators need to distinguish and deal with. An animal's *Umwelt* consists in the first place of *affordances* (Gibson 1979), things to eat or mate with, openings to walk through or look out of, holes to hide in, things to stand on, and so forth. We may suppose that the *Umwelt* of a starfish or worm or daisy is more like the ontology of the elevator than like our manifest image. What's the difference? What makes our manifest image *manifest* (to *us*)?

Here is where Bayesian expectations could play an iterated role: Our ontology (in the elevator sense) does a close-to-optimal job of representing the things in the world that matter to the behavior our brains have to control. Hierarchical Bayesian predictions accomplish this, generating affordances galore: We expect solid objects to have backs that will come into view as we walk around them, doors to open, stairs to afford climbing, cups to hold liquid, and so forth. But among the things in our *Umwelt* that matter to our well-being are *ourselves*! We ought to have good Bayesian expectations about what we will do next, what we will think next, and what we will *expect* next! And we do. Here's an example:

Think of the cuteness of babies. It is not, of course, an "intrinsic" property of babies, though it seems to be. What you "project" out onto the baby is in fact your manifold of "felt" dispositions to cuddle, protect, nurture, kiss, coo over, . . . that little cutie-pie. It's not just that when your cuteness detector (based on facial proportions, etc.) fires, you have urges to nurture and protect; you *expect* to have those very urges, and that manifold of expectations just *is* the "projection" onto the baby of the property of cuteness. When we expect to see a baby in the crib, we also expect to "find it cute" – that is, we *expect* to *expect* to feel the urge to cuddle it and so forth. When our expectations are fulfilled, the absence of prediction error signals is interpreted as confirmation that, indeed, the thing in the world we are interacting with has the properties we expected it to have. Cuteness as a property passes the Bayesian test for being an objective structural part of the world we live in, and that is all that needs to happen. Any further "projection" process would be redundant. What is special about properties like sweetness and cuteness is that their perception depends on particularities of the nervous systems that have evolved to make much of them. The same is of course also true of colors. This is what is left of Locke's (and Boyle's) distinction between primary and secondary qualities.

Grounding predictive coding models in empirical neuroscience research

doi:10.1017/S0140525X1200218X

Tobias Egner[a] and Christopher Summerfield[b]

[a]*Department of Psychology & Neuroscience, and Center for Cognitive Neuroscience, Duke University, Durham, NC 27708;* [b]*Department of Experimental Psychology, University of Oxford, Oxford OX1 3UD, United Kingdom.*

tobias.egner@duke.edu
http://sites.google.com/site/egnerlab/
christopher.summerfield@psy.ox.ac.uk
https://sites.google.com/site/summerfieldlab/home

Abstract: Clark makes a convincing case for the merits of conceptualizing brains as hierarchical prediction machines. This perspective has the potential to provide an elegant and powerful general theory of brain function, but it will ultimately stand or fall with evidence from basic neuroscience research. Here, we characterize the status quo of that evidence and highlight important avenues for future investigations.

The intuition that our brains harbor a predictive (forward) model linking visual percepts to their probable external causes (Helmholtz 1876) has been fleshed out over recent decades by sophisticated models (Friston 2005; Mumford 1992; Rao & Ballard 1999), inspiring the view that Clark puts forward in the target article, that *predictive coding* is a cardinal principle of neural systems (cf. Friston 2010; Hawkins & Blakeslee 2004). While this perspective offers elegant post-hoc explanations for a wide array of behavioral and neural phenomena, empirical studies directly testing the basic biological assumptions of predictive coding remain scarce. Specifically, the core empirical hypotheses derived from the predictive coding scheme are the presence of separable and hierarchically organized visual expectation and surprise computations (and associated neural units/signals) in the posterior brain (Friston 2005). These predictions are provocative, because they differ drastically from traditional views of visual neurons as mere bottom-up feature detectors (Hubel & Wiesel 1965; Riesenhuber & Poggio 2000). But what is the empirical evidence directly supporting these claims? We first address results from macroscopic, human neuroimaging studies, followed by microscopic data from invasive animal experiments.

At the macroscopic level of inquiry provided by whole-brain functional neuroimaging, there are at present modest but promising lines of empirical support for predictive coding's core propositions. Most firmly established is the finding of robust occipital responses evoked by the surprising presence or absence of visual stimuli, presumably attributable to the computation of prediction error (e.g., Alink et al. 2010; den Ouden et al. 2009; Egner et al. 2010). Similarly, "repetition suppression," the attenuated neural response to a repeated stimulus that predictive coding attributes to a decrease in prediction error (Friston 2005), has repeatedly been shown to be modulated by expectations, including in human functional magnetic resonance imaging (fMRI) (Summerfield et al. 2008), electroencephalographic (EEG) (Summerfield

et al. 2011), and magnetoencephalographic (MEG) (Todorovic et al. 2011) recordings. However, although evidence for visual surprise signals at the neural population level is fairly abundant, the attribution of these signals to local prediction error computations is not unequivocal, in that they could instead be argued to reflect attentional highlighting of unexpected stimuli (cf. Pearce & Hall 1980) driven by predictive processing elsewhere in the brain. In fact, the precise role that attention plays in the predictive coding machinery is currently under debate (Feldman & Friston 2010; Summerfield & Egner 2009) and represents an important line of recent (Kok et al. 2011; Wyart et al. 2012) and future investigations into the predictive brain hypothesis.

In contrast to this support for the existence visual surprise signals, the proposition that there are *simultaneous* computations of prediction and prediction error signals carried out by distinct neural populations in visual cortex is presently only poorly substantiated. One recent fMRI study showed that neural population responses in the ventral visual stream can be successfully modeled as reflecting the summed activity of putative prediction and prediction error signals (Egner et al. 2010; Jiang et al. 2012). Similarly, a recent computational model can account for a wide array of auditory EEG responses by supposing co-existing prediction and prediction error neurons (Wacongne et al. 2012). However, neither of these studies demonstrates unambiguously the simultaneous operation of distinct neural sub-populations coding for expectations and surprise, a finding that would greatly bolster the biological feasibility of predictive coding models. Finally, the purported hierarchical nature of the interplay between expectation and surprise signals has garnered indirect support from a handful of fMRI studies. For instance, Murray and colleagues demonstrated the "explaining away" of activity in lower-level visual regions by activity in higher-level visual cortex when presenting a coherent visual object compared to its dissembled constituent parts (Murray et al. 2002). Other investigators have employed effective connectivity analysis of fMRI data to probe how dynamic interactions between different brain regions may mediate prediction and surprise signals (den Ouden et al. 2009; 2010; Kok et al. 2011; Summerfield & Koechlin 2008; Summerfield et al. 2006). Nevertheless, a comprehensive demonstration of predictive coding "message passing" across several adjacent levels of the visual processing hierarchy remains lacking from the literature.

Perhaps most importantly, microscopic or cellular level data addressing the core tenets of the predictive coding hypothesis have been particularly scarce. In part, this may be for methodological reasons: For example, neurons with proposed "predictive fields" might be excluded from recording studies where cells are screened according to their bottom-up sensitivity. Moreover, the dynamics of the reciprocal interaction within the hierarchy might give rise to complex neural responses, making it hard to segregate prediction and error signals. Nevertheless, recent work has supplied some promising data. First, Meyer and Olson (2011) have recently described single neurons in monkey inferotemporal cortex that exhibit surprise responses to unexpected stimulus transitions, thus possibly documenting visual prediction error neurons in the ventral visual stream. Two other recent studies, one in monkeys (Eliades & Wang 2008) and one in mice (Keller et al. 2012), assessed neuronal activity in the context of sensorimotor feedback (e.g., the integration of movement with predicted changes in visual stimulation), observing putative prediction error signals in primary sensory cortices (for alternative interpretations, see Eliades & Wang 2008). Importantly, in Keller et al. (2012), these surprise signals co-occurred with both pure motor-related and sensory-driven signals, thus providing initial evidence for the possibility of co-habiting prediction and prediction error neurons in early visual cortex. Moreover, the putative prediction error neurons were found in supra-granular layers 2/3, which house precisely the superficial pyramidal cells that have been posited to support prediction error signaling by theoretical models of predictive coding (Friston 2008; Mumford 1992).

In conclusion, we submit that the extant data from studies that directly aimed at testing core tenets of the predictive coding hypothesis are few but generally supportive. Looking to the future, additional demonstrations of simultaneous prediction and surprise computations within a single processing stage (in particular from single-neuron electrophysiology), as well as evidence for hierarchical interactions with adjacent stages, are required. We hope that over coming years, neuroscientists will be inspired to collect these data.

Prediction, explanation, and the role of generative models in language processing

doi:10.1017/S0140525X12002312

Thomas A. Farmer,[a,b] Meredith Brown,[a] and Michael K. Tanenhaus[a]

[a]*Department of Brain and Cognitive Sciences and* [b]*Center for Language Sciences, University of Rochester, Rochester, NY 14627-0268.*

tfarmer@bcs.rochester.edu **mbrown@bcs.rochester.edu** **mtan@bcs.rochester.edu**

Abstract: We propose, following Clark, that generative models also play a central role in the perception and interpretation of linguistic signals. The data explanation approach provides a rationale for the role of prediction in language processing and unifies a number of phenomena, including multiple-cue integration, adaptation effects, and cortical responses to violations of linguistic expectations.

Traditional models of language comprehension assume that language processing involves recognizing patterns, for example, words, by mapping the signal onto existing representations, retrieving information associated with these stored representations, and then using rules based on abstract categories (e.g., syntactic rules) to build structured representations. Four aspects of the literature are inconsistent with this framework. First, listeners are exquisitely sensitive to fine-grained, sub-categorical properties of the signal, making use of this information rather than discarding it (McMurray et al. 2009). Second, comprehenders rapidly integrate constraints at multiple grains. Third, they generate expectations about likely input at multiple levels of representation. Finally, adaptation is ubiquitous in language processing. These results can be unified if we assume that comprehenders use internally generated predictions at multiple levels to *explain* the source of the input, and that prediction error is used to update the generative models in order to facilitate more accurate predictions in the future.

Extended to the domain of language processing, Clark's framework predicts that expectations at higher levels of representation (e.g., syntactic expectations) should constrain interpretation at lower levels of representation (e.g., speech perception). According to this view, listeners develop fine-grained probabilistic expectations about how lexical alternatives are likely to be realized in context (e.g., *net* vs. *neck*) that propagate from top to bottom through the levels of a hierarchically organized system representing progressively more fine-grained perceptual information. Provisional hypotheses compete to explain the data at each level, with the predicted acoustic realization of each alternative being evaluated against the actual form of the input, resulting in a residual feed-forward error signal propagated up the hierarchy. As the signal unfolds, then, the activation of a particular lexical candidate should be inversely proportional to the joint error signal at all levels of the hierarchy (i.e., the degree of divergence between the predicted acoustic realization of that candidate and the actual incoming signal), such that candidate words whose predicted realizations are most congruent with the acoustic signal are favored.

Hierarchical predictive processing therefore provides a potential explanatory framework for understanding a wide variety of context effects and cue integration phenomena in spoken word

recognition. Converging evidence suggests that the initial moments of competition between lexical alternatives are constrained by multiple sources of information from different dimensions of the linguistic input (e.g., Dahan & Tanenhaus 2004; Kukona et al. 2011), including information external to the linguistic system, such as visually conveyed social information (Hay & Drager 2010; Staum Casasanto 2008) and high-level information about a speaker's linguistic ability (Arnold et al. 2007). Crucially, lexical processing is influenced by information preceding the target word by several syllables or clauses (Dilley & McAuley 2008; Dilley & Pitt 2010) and this information affects listeners' expectations (Brown et al. 2011; 2012). The integration of these various constraints, despite their diversity, is consistent with the hypothesis that disparate sources of constraint are integrated within generative models in the language processing system.

Clark's framework also helps explain a recent set of results on context effects in reading that are surprising from the viewpoint of more traditional theories that emphasize the bottom-up, feed-forward flow of information. Farmer et al. (2006) demonstrated that when a sentential context conferred a strong expectation for a word of a given grammatical category (as in *The child saved the…*, where a noun is strongly expected), participants were slower to read the incoming noun when the *form* of it (i.e., its phonological/orthographic properties) was atypical with respect to other words in the expected category. In a subsequent MEG experiment, Dikker et al. (2010) showed that at about 100 msec post-stimulus onset – timing that is unambiguously associated with perceptual processing – a strong neural response was elicited when there was a mismatch between form and syntactic expectation. Moreover, the source of the effect was localized to the occipital lobe, suggesting that the visual system had access to syntactic representations. These results provide support for Clark's hypothesis that "if the predictive processing story is correct, we expect to see powerful context effects propagating quite low down the processing hierarchy" (sect. 3.1, para. 8). Linguistic context is used to generate expectations about form-based properties of upcoming words, and these expectations are propagated to perceptual cortices (Tanenhaus & Hare 2007).

This framework also serves to specify the functionality of the prediction error that arises when some degree of mismatch between a prediction and the incoming signal occurs. In behavioral and Event-Related Potential (ERP) experiments, prediction-input mismatch frequently results in increased processing difficulty, typically interpreted as evidence that prediction is being made. But, under Clark's framework, the error signal assumes functionality; in part, it serves to adjust higher-level models such that they better approximate future input. The explanatory power of this hypothesis can best be seen when considering the large amount of relatively recent literature on adaptation within linguistic domains. Whether in the domain of speech perception (Kleinschmidt & Jaeger 2011; Kraljic et al. 2008), syntactic processing (Farmer et al. 2011; Fine et al. under review; Wells et al. 2009), prosody (Kurumada et al. 2012), or pragmatics (Grodner & Sedivy 2011), it has become increasingly apparent that readers and listeners continually update their expectations about the likelihood of encountering some stimulus based on their exposure to the statistical regularities of a specific experimental context. Adaptation of expectations is predicted by Clark's framework, and it may be taken as evidence that prediction-input mismatch produces an error signal that is fed forward to update the relevant generative models.

In sum, Clark's hierarchical prediction machine hypothesis provides a framework that we believe will unify the literature on prediction in language processing. This unification will necessarily involve systematic examination of what aspects of the stimulus are predicted, when in the chain of processing these predictions are generated and assessed, and the precise form of these generative models. This task will be challenging because it is likely that generative models use signal-relevant properties that do not map to the standard levels of linguistic representation that are incorporated into most models of language processing.

Active inference and free energy

doi:10.1017/S0140525X12002142

Karl Friston

The Wellcome Trust Centre for Neuroimaging, Institute of Neurology, University College London, Queen Square, London WC1N 3BG, United Kingdom.
k.friston@ucl.ac.uk

Abstract: Why do brains have so many connections? The principles exposed by Andy Clark provide answers to questions like this by appealing to the notion that brains distil causal regularities in the sensorium and embody them in models of their world. For example, connections embody the fact that causes have particular consequences. This commentary considers the imperatives for this form of embodiment.

1. Introduction. It is a pleasure to comment upon Andy Clark's exposition of the Bayesian brain, predictive coding, and the free-energy principle. Clark describes modern thinking about the brain as a constructive and predictive machine in a compelling and accessible way. Furthermore, he develops the fundaments of this approach from basic questions about the nature of life and consciousness – remarkably, without recourse to mathematical equations.

Clark's synthesis is impressive – it highlights the consistency (and convergence) of the underlying ideas from many perspectives, ranging from the psychophysics of perceptual inference through to motor control and embodiment. The key thing that emerges from his treatment is that minimising surprise or surprisal (Tribus 1961) accommodates many intuitions and theories about brain function that have emerged over the past century or so. Had space allowed, other ideas could have been celebrated (developed) within this framework; for example, the principle of efficient coding (Barlow 1961); the notion of perception as hypothesis testing (Gregory 1980), and the action-perception cycle (Fuster 2001) – all rest on the premise that we build parsimonious models to explain our world (Dayan et al. 1995).

In what follows, I revisit three challenges – highlighted by Clark – to the free-energy principle, and its incarnations like predictive coding and the Bayesian brain. Specifically, these are: (1) the relationship between free-energy minimisation and predictive coding, (2) the dark room problem, and (3) explanatory power.

2. Free-energy and predictive coding. Clark frames surprise minimisation in terms of predictive coding in the Bayesian brain (Mumford 1992; Rao & Ballard 1999; Yuille & Kersten 2006). This works extremely well and is a useful way to introduce the ideas. However, it may detract from a simple but important point: *Predictive coding is a consequence of surprise minimisation, not its cause*. Free-energy is a mathematical bound on surprise, where prediction error is a measure of free-energy that is easy to compute (neurobiologically). Free-energy minimisation is an instance of the celebrated *principle of least action* – because the average energy over time is also called action. Furthermore, it entails the *maximum entropy principle* (Jaynes 1957) – because free-energy is expected energy minus the entropy of predictions. These principles will be familiar to anyone in physics or statistics because they govern the behaviour of known physical systems. The important thing – for self-organising systems – is that the long-term average of surprise is (almost surely) equal to the entropy of sensations. This means that minimising free-energy minimises sensory entropy. As articulated nicely by Clarke, we can minimise free-energy (prediction errors) by either changing our predictions (perception) or changing the things that we predict (action). The key thing that the free-energy principle brings to the table is that both perception and action minimise prediction error *but only action minimises surprise* (because surprise is an attribute of sensations actively sampled). This is active inference (Friston 2010). The imperative to minimise surprise rests on the need to resist a natural tendency to disorder – to minimise sensory entropy (Ashby 1947). The Bayesian brain and predictive

coding are then seen as a consequence of, or requirement for, this fundamental imperative – not as a causal explanation for how our brains work. This is important, because any evidence that suggests we are Bayes-optimal can be taken as evidence for active inference.

3. The dark room problem. Clark introduces and then (almost) dismisses the dark room problem by appeal to itinerant (exploratory) behaviours that minimise surprise over long periods of time (that is, minimise sensory entropy). I think that his discussion is exactly right; however, the "grain of truth" in the dark room problem can be dismissed in an even simpler way – by noting that predication errors are only defined in relation to predictions. For example, when we enter a dark room, the first thing we do is switch on a light. This is because we expect the room to be brightly lit (or more exactly, we expect our bodily movements to bring this about). In other words, the state of a room being dark is surprising because we do not expect to occupy dark rooms. This surprise depends upon (prior) expectations, but where do these prior beliefs come from? They come from evolution and experience, in the sense that if we did not have these prior beliefs, we would be drawn to dark rooms and die there. In short, a dynamic world can only support a generative model of that world (prior beliefs) that predicts the dynamics it encounters – predictions that action fulfils.

4. Evidence and explanatory power. Clark questions the evidence for surprise minimisation and its explanatory power. I am more complacent about this issue, because the free-energy formulation explains so much already. Potent examples rest on appreciating that an agent does not have a model of its world – it *is* a model. In other words, the form, structure, and states of our embodied brains do not contain a model of the sensorium – they are that model. This allows one to equate the long-term minimisation of surprise with the entropy of our physical (sensory) states – and explains our curious (biological) ability to resist the second law of thermodynamics (Ashby 1947). But what does this mean practically? It means that every aspect of our brain can be predicted from our environment. This seems a powerful explanation for neuroanatomy and neurophysiology. A nice example is the anatomical division into *what* and *where* pathways in visual cortex (Ungerleider & Mishkin 1982). Could this have been predicted from the free-energy principle? Yes – if anatomical structure in the brain recapitulates causal structure in the environment, then one would expect independent causes to be encoded in functionally segregated neuronal structures. Given that objects can be in different places, they possess separable attributes of "what" and "where." This translates into separate neuronal representations in segregated visual pathways. In summary, the evidence for the free-energy principle may not necessarily be in next month's scientific journals but may lie in the accumulated wealth of empirical neurobiological knowledge that Andy Clark has unpacked for us.

The brain is not an isolated "black box," nor is its goal to become one

doi:10.1017/S0140525X12002348

Tom Froese[a,b] and Takashi Ikegami[b]

[a]*Departamento de Ciencias de la Computación, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, Ciudad Universitaria, A.P. 20-726, 01000 México D.F., México;* [b]*Ikegami Laboratory, Department of General Systems Studies, Graduate School of Arts and Sciences, University of Tokyo, Meguro-ku, Tokyo 153-8902, Japan.*

t.froese@gmail.com **http://froese.wordpress.com**
ikeg@sacral.c.u-tokyo.ac.jp **http://sacral.c.u-tokyo.ac.jp/**

Abstract: In important ways, Clark's "hierarchical prediction machine" (HPM) approach parallels the research agenda we have been pursuing. Nevertheless, we remain unconvinced that the HPM offers the best clue yet to the shape of a unified science of mind and action. The apparent convergence of research interests is offset by a profound divergence of theoretical starting points and ideal goals.

We share with Clark a commitment to exploring the deep continuities of life, mind, and sociality (Froese & Di Paolo 2011). Similar to the enactive notion of "sense-making," Clark's "hierarchical prediction machine" (HPM) entails that perceiving cannot be separated from acting and cognizing. Nevertheless, we disagree with Clark's theoretical premises and their ideal consequences.

Clark begins with the assumption that the task of the brain is analogous to establishing a "view from inside the black box." On this view, the mind is locked inside the head and it follows that, as Clark puts it, "the world itself is thus off-limits" (sect. 1.2, para. 1). This is the premise of *internalism*, from which another assumption can be derived, namely that knowledge about the world must be indirect. Accordingly, there is a need to create an internal model of the external source of the sensory signals, or, in Clark's terms, of "the world hidden behind the veil of perception" (sect. 1.2, para. 6). This is the premise of *representationalism*.

It is important to realize that these two premises set up the basic problem space, which the HPM is designed to solve. Without them, the HPM makes little sense as a scientific theory. To be sure, internalism may seem to be biologically plausible. As Clark observes, all the brain "knows" about, in any direct sense, are the ways its own states (e.g., spike trains) flow and alter. However, the enactive approach prefers to interpret this kind of autonomous organization not as a black-box prison of the mind, but rather as a self-organized perspectival reference point that serves to enact a set of meaningful relations with its milieu (Di Paolo 2009). On this view, mind and action are complex phenomena that emerge from the nonlinear interactions of brain, body, and environment (Beer 2000). Such a dynamical perspective supports a relational, direct realist account of perception (Noë 2004; 2009).

An enactive approach to neuroscience exhibits many of the virtues of the HPM approach. Following the pioneering work of Varela (1999), it is also formalizable (in dynamical systems theory); it has explanatory power (including built-in context-sensitivity); and it can be related to the fundamental structures of lived experience (including multistable perceptions). Indeed, it accounts for much of the same neuroscientific evidence, since global self-organization of brain activity – for example, via neural synchrony – requires extensive usage of what Clark refers to as "backward connections" in order to impose top-down constraints (Varela et al. 2001).

Advantageously, the enactive approach avoids the HPM's essential requirement of a clean functional separation between "error units" and "representation units," and it exhibits a different kind of neural efficiency. Properties of the environment do not need to be encoded and transmitted to higher cortical areas, but not because they are already expected by an internal model of the world, but rather because the world is its own best model. The environment itself, as a constitutive part of the whole brain-body-environment system, replaces the HPM's essential requirement of a multilevel generative modeling machinery (cf. Note 16 in the target article).

The enactive approach also avoids absurd consequences of the HPM, which follow its generalization into an all-encompassing "free-energy principle" (FEP). The FEP states that "all the quantities that can change; i.e. that are part of the system, will change to minimize free-energy" (Friston & Stephan 2007, p. 427). According to Clark, the central idea is that perception, cognition, and action work closely together to minimize sensory prediction errors by selectively sampling, and actively sculpting, the stimulus array. But given that there are no constraints on this process (according to the FEP, everything is enslaved as long as it is part of the system), there are abnormal yet effective ways of reducing prediction error, for example by stereotypic self-stimulation,

catatonic withdrawal from the world, and autistic withdrawal from others. The idea that the brain is an isolated black box, therefore, forms not only the fundamental starting point for the HPM, but also its ideal end point. Ironically, raising the HPM to the status of a universal principle has the opposite effect: namely, making it most suitable as an account of patently pathological mental conditions.

Similar concerns about the overgeneralization of the FEP have been raised by others (Gershman & Daw 2012), and are acknowledged by Clark in his "desert landscape" and "dark room" scenarios. The general worry is that an agent's values need to be partially decoupled from prediction optimization, since reducing surprise for its own sake is not always in the organism's best interest. In this regard the enactive approach may be of help. Like Friston, it rejects the need for specialized value systems, as values are deemed to be inherent in autonomous dynamics (Di Paolo et al. 2010). But it avoids the FEP's problems by grounding values in the viability constraints of the organism. Arguably, it is the organism's precarious existence as a thermodynamically open system in non-equilibrium conditions which constitutes the meaning of its interactions with the environment (Froese & Ziemke 2009).

However, this enactive account forces the HPM approach to make more realistic assumptions about the conditions of the agent. Notably, it is no longer acceptable that the FEP requires a "system that is at equilibrium with its environment" (Friston 2010, p. 127). This assumption may appear plausible at a sufficiently abstract level (Ashby 1940), but only at the cost of obscuring crucial differences between living and non-living systems (Froese & Stewart 2010). Organisms are essentially non-equilibrium systems, and thermodynamic equilibration with the environment is identical with disintegration and death, rather than optimal adaptiveness. However, contra to the motivations for the FEP (Friston 2009, p. 293), this does not mean that organisms aim to ideally get rid of disorder altogether, either. Living beings are precariously situated between randomness and stasis by means of self-organized criticality, and this inherent chaos has implications for perception (Ikegami 2007). Following Bateson, we propose that it is more important to be open to perceiving differences that make a difference, rather than to eliminate differences that could surprise you.

Unraveling the mind

doi:10.1017/S0140525X1200235X

Philip Gerrans

Department of Philosophy, University of Adelaide, North Terrace Campus, SA 5005, Australia.

philip.gerrans@adelaide.edu.au **http://philipgerrans.com**

Abstract: A radical interpretation of the predictive coding approach suggests that the mind is "seamless" – that is, that cancellation of error signals can propagate smoothly from highest to lowest levels of the control hierarchy, dissolving a distinction between belief and perception. Delusions of alien control provide a test case. Close examination suggests that while they are evidence of predictive coding within the cortex, they are not evidence for the seamless interpretation.

Andy Clark describes delusions as the dark side of the seamless story for predictive coding in which, "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 internal models that are making the predictions" (sect. 2.3, para. 8).

Theorists who endorse the predictive coding model have argued that in delusions of alien control, patients *actually experience* being controlled by an external agent. As Gallagher puts it, "the attribution of agency to another *is a genuine result of what is truly experienced*" (Gallagher 2004, p. 17, my italics). Some experiments suggest that this experience is the result of a prior belief about the external origin of movement. This would be a nice vindication of the seamless story. I think, however, that the mind is not quite so seamless and that there is another explanation consistent with the predictive coding framework.

How could someone experience his or her own movements as alienated actions? The short answer is that right inferior parietal activation represents "surprisal" for intended movements. Surprisal is minimised for *intended* movements because the motor command from the supplementary motor area (SMA) attenuates activity in the right interior parietal cortex. On the seamless story, unpredicted/unattenuated parietal activation (surprisal) arising in the context of action observation is experienced as alienation: "The patients really had no cues (as inferred from the change in activity in the parietal lobe) about whether they saw their own movements *or those of an alien agent*" (Jeannerod 2006, my italics). Thus, they experience their own movements as alienated.

In an important experiment Daprati and collaborators had subjects trace a path from their body midline to a target directly in front of them. The subjects' view of their moving hands was occluded until the final 30% of the movement. For the first 70%, patients saw a computer-generated trace of the movement path. On some trials the experimenters introduced a deviation of 15% into the movement path so that if uncorrected the trace would veer off to the right. Both schizophrenic and neurotypical subjects were able to compensate for the perturbation, during the occluded section of the movement, with the result that when the hand came into view, the hand was to the left of the midline. Danckert et al. (2004) express the consensus in a large literature when they say that such cases show that "on-line monitoring and adjustment of action is unaffected in patients with schizophrenia" (p. 253).

In Daprati's experiment, the last 30% of the movement is not occluded. When the subject *sees* the hand it is 15 degrees to the left of a straight line to the target. Neurotypical subjects attributed this discrepancy to the computer, indicating that they were able to become aware that they had intended a different movement than the one they actually made. Schizophrenics with positive symptoms did not, leading to the conclusion that "online control can coexist with a tendency to *misattribute the source of error*" (Daprati et al. 1997, p. 253, emphasis theirs).

This tendency arises for schizophrenics when they visually attend to the movement. In this case they seem lose access to information about self-initiation. (Note: this is a problem of *degree not kind*. The dominance of visual attention over proprioceptive/motor information generates similar misattributions in many conditions).

Blakemore et al. (2003) hypnotized subjects whose arms were attached to a pulley apparatus and gave them two instructions. In the first they were told to raise their arms and in the second that the pulley would raise their arms. The pulley did not actually exert any force. Highly hypnotizable subjects moved their arms in response to both instructions but in the second case they reported no feeling of agency, attributing the movement to the pulley. In effect, hypnosis induced the experience of failed action monitoring characteristic of delusions of alien control. The authors explain: "The prediction made by the parietal cortex is concerned more with *high level prediction such as strategic planning actions.*" Furthermore, they suggest, "Perhaps the predictions made by the parietal cortex can be made available to consciousness" (Blakemore et al. 2003, p. 243, my italics). In other words we can experience ourselves as authors of our actions in virtue of attenuated parietal activity. Because schizophrenics cannot attenuate this activity, they cannot become aware of themselves as authors of their actions in some conditions.

Does it follow that unattenuated parietal activity represents that *someone else* is the author of the action? From what we have seen so far, the modulation of parietal activity only tells the subject

whether a movement is produced by the SMA. That is a very low level of cognitive processing from which information about agency is absent.

Evolution has not posed us with the problem of determining which movements are ours *rather than someone else's*. It has posed us with the problem of determining which aspects of a movement are consequences of motor intentions in order to compute and resolve error. Therefore, there seems no reason to think that we would need to use predictive coding to disambiguate the *agent* of an action rather than to simply control our own action. This is true both at the level of automatic and of controlled processing.

In general, then, I conclude that parietal activation is not specialised for determining *who* intended the action. Rather, it determines for any movement whether it is a consequence of a motor instruction. It evolved to control movement, not to identify the agent. Because schizophrenics cannot attenuate this activity when visually monitoring actions, they cannot experience themselves as authors of those actions. In both experiments, however, the context provides a default interpretation of alienation.

If the fabric of the mind is stitched together seamlessly with predictive coding threads we should be able to unravel it entirely from the top down. But the fact that online control in schizophrenia is intact suggests that the seam linking automatic and visually guided motor control, while flexible, has been robustly tailored by evolution.

Bayesian animals sense ecological constraints to predict fitness and organize individually flexible reproductive decisions

doi:10.1017/S0140525X12002385

Patricia Adair Gowaty and Stephen P. Hubbell

Department of Ecology and Evolutionary Biology, and Institute of Environment and Sustainability, Los Angeles, CA 90095; and Smithsonian Tropical Research Institute, Unit 9100, BOX 0948, DPO AA 34002-9998.

gowaty@eeb.ucla.edu shubbell@eeb.ucla.edu

http://www.eeb.ucla.edu/indivfaculty.php?FacultyKey=8418

http://www.eeb.ucla.edu/indivfaculty.php?FacultyKey=8416

Abstract: A quantitative theory of reproductive decisions (Gowaty & Hubbell 2009) says that individuals use updated priors from constantly changing demographic circumstances to predict their futures to adjust actions flexibly and adaptively. Our ecological/evolutionary models of ultimate causes seem consistent with Clark's ideas and thus suggest an opportunity for a unified proximate and ultimate theory of Bayesian animal brains, senses, and actions.

Reading Clark suggests possible connections between proximate causes of animal – not just human – perception, mind, and action and their ultimate causes. We suggest that it is worth considering that nonhuman animals, not just humans, are Bayesian too, and that the world also appears to them as a set of intertwined probability density distributions. We think of all animals as Bayesian and we define (Gowaty & Hubbell 2005; 2009) animals as adaptively flexible individuals who "predict" ("visualize," "imagine") alternatives and make choices among them "controlling plasticity" to serve fitness. We have argued previously that animals predict their futures and act as though they are indeed perceiving and responding to "intertwined set[s] of probability density distributions" (see target article, sect. 4.1, para. 3). We say explicitly that animals behave as if playing the odds of *fitness* against the odds of time. Thus, we argue that animals are flexible individuals who act behaviorally and physiologically in real ecological time, not just evolutionary time, to enhance their real-time fitness. Could it be that the intertwined set of probability density distributions associated with the main problems of individuals – surviving and reproducing – are on a continuum of connected ultimate and proximate causes and perhaps fuel the organization of perception and action? Do Bayesian animals predict the future from a set of constantly updated priors to produce predictions of most importance: finding a mate, finding a better mate, or dying?

Fitness is a relative concept and demography-dependent. Here, we direct readers to a theoretical scenario (Figure 1) with its mathematical analytical solutions for the evolution of human and nonhuman Bayesian individuals who perceive their real time alternatives, predict the fitness that would accrue or not from those alternatives and modify their behavior accordingly. One of our main assumptions is that individuals are able to predict (unconsciously or consciously) their own demographic circumstances (how they are doing/will do relative to others). To some of our readers, our assumptions have seemed otherworldly. Clark's article suggests that our assumptions are not so odd in the human cognitive sciences and they signal new empirical research about the meanings of animal behavior in the unified contexts of linked proximate and ultimate causes.

From a Darwinian evolutionary perspective (Darwin 1871), who among potential mates to accept and/or reject is one of the most important of reproductive decisions. To be fitness enhancing in contemporary time, reproductive decisions must be flexible and made against the unavoidable context of demography (Gowaty & Hubbell 2005). Demography is not static: things change; stochastic effects are inevitable. Potential mates enter and leave populations; some individuals may die and never appear again; and predators, parasites, and pathogens come and go, so that the survival likelihoods of decision-makers also change. The minimal set of parameters contributing to stochastic demography (Hubbell & Johnson 1987) are those providing sensory information about the availability of potential mates (encounter probability, e), the likelihood of continued life of decision-makers (survival probability, s), and the distribution within the population of fitness that would be conferred from mating with this or that potential mate (*w-distribution*). The minimal set of information necessary for making real-time, fitness-enhancing reproductive decisions is e, s and the *w-distribution*.

Gowaty and Hubbell (2005) hypothesized that individuals, not sexes, are under selection to flexibly modify their reproductive decisions moment-to-moment as their ecological and social circumstances change to enhance their instantaneous contributions to lifetime mean fitness (Fig. 1). Stochastic variation in e, s, and l (latency from the end of one mating, to onset, to receptivity, to the next mating) results in mean lifetime number of mates (MLNM). Variation in MLNM favors the evolution of *sensitivity* to e, s, and l, while variation in the *w-distribution* favors *assessment* of fitness that would be conferred through mating with this or that potential mate. Once sensitivity and assessment evolve, the stage is set for flexible individuals to modify their behavior in ways which their sensitivities and assessments predict are fitness enhancing. The analytical solution to this model is the Switch Point Theorem (SPT). An SPT graph shows the rule for acceptance and rejection of each potential mate, ranked from best at 1 to worst at n, by a single unique individual in the population, given variation in e, s, l, n and the *w-distribution*.

The assumptions of the analytical solution as to how many potential mates in a population will be acceptable or not to a given individual (Gowaty & Hubbell 2009) are as follows:

1. Before there was natural selection to accept or reject potential mates, there was stochastic variation in encounters with potential mates and with decision-makers' likelihood of survival.
2. The encounter probability and survival probability determine the mean lifetime number of mates and the variance in lifetime number of mates.
3. Potential mates come in n-qualities, where n = the number of potential mates in the population.
4. Mate assessment is self-referential and depends upon information learned during development about self relative to others.

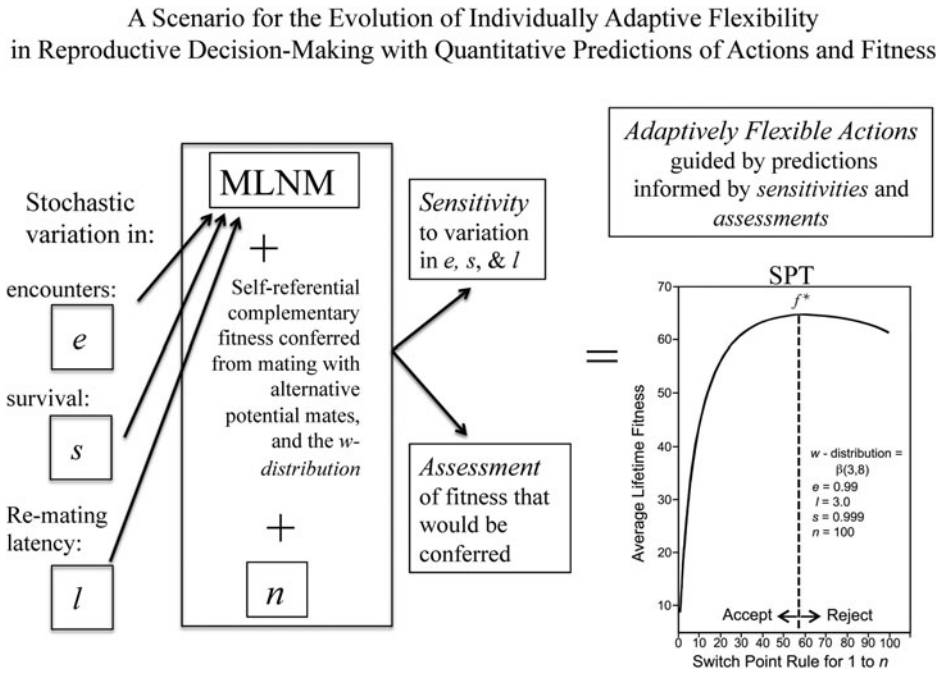

Figure 1 (Gowaty & Hubbell). The hypothesis for the evolution of adaptively flexible behavior (modified from figures in Gowaty & Hubbell 2005; 2009).

5. Individuals update their information to predict adaptive acceptance and rejection of potential mates thereby maximizing instantaneous contributions to lifetime fitness.

The analytical solution of whom to accept and reject for mating is the switch point theorem (SPT in Fig. 1).

Resistance to our assumptions from behavioral ecologists is perhaps not surprising, for we begin with individuals, rather than sexes, to predict sex differences. What surprises us, however, is that there are critics who resist our assumption that animals use probabilistic information as instantaneous clues to predict their next move, which the SPT proved theoretically is adaptive. The Bayesian updating that Clark describes as a fundamental aspect of neural processing of what the world is, suggests to us that his and our ideas are conceptually linked. Our use of the Bayesian metaphor suggests that there is something self-similar linking proximate and ultimate causes. But, what if animals too are Bayesians with linkages between how and why brains interpret the world?

We agree with Clark. What is on offer is a unified science of perception, attention, prediction, and flexibility of action. The SPT suggests that fitness drives all.

Personal narratives as the highest level of cognitive integration

doi:10.1017/S0140525X12002269

Jacob B. Hirsh,[a] Raymond A. Mar,[b] and Jordan B. Peterson[c]

[a]*Rotman School of Management, University of Toronto, Toronto, ON M5S 3E6, Canada;* [b]*Department of Psychology, York University, Toronto, ON M3J1P3, Canada;* [c]*Department of Psychology, University of Toronto, Toronto, ON M5S 3G3, Canada.*

jacob.hirsh@utoronto.ca **www.jacobhirsh.com**
mar@yorku.ca **www.yorku.ca/mar**
peterson@psych.utoronto.ca **www.psych.utoronto.ca/users/peterson**

Abstract: We suggest that the hierarchical predictive processing account detailed by Clark can be usefully integrated with narrative psychology by situating personal narratives at the top of an individual's knowledge hierarchy. Narrative representations function as high-level generative models that direct our attention and structure our expectations about unfolding events. Implications for integrating scientific and humanistic views of human experience are discussed.

Clark's article presents the hierarchical predictive processing account of human cognition as a unifying model for understanding mind and action. He also highlights the importance of bridging this perspective with our daily "folk" or "humanistic" conceptions of self and world. We propose that such a bridge is provided by the field of narrative psychology, with narrative models of the world occupying the highest levels of an individual's predictive hierarchy.

A growing body of theory and research indicates that the broadest and most integrative levels of an individual's knowledge system can be characterized as narrative descriptions of reality (Bruner 1986; 1991; McAdams 1997; Peterson 1999; Ricoeur et al. 1990; Sarbin 1986). Although narratives can take many different forms, they are distinguished by their ability to compress and encode a great deal of information about the world, including the causal relations between events over time (Graesser et al. 1997), the planning and sequencing of goal-directed actions (Schank & Abelson 1977), the emotional significance of an event within a temporal context (Oatley 1992), the unfolding nature of personal identity (McAdams 1997), and the dynamic intentions of multiple social agents (Mar & Oatley 2008). It is the integrative ability of narrative representations to coordinate vast domains of knowledge and behavior that has led some theorists to propose narrative as an organizing framework for understanding human psychology (Sarbin 1986). Narrative representations thus appear to function as high-level generative models of the sort that Clark describes, structuring our expectations about daily experiences and providing an organizing framework for interpreting incoming sensory information (Bruner 1986; Mandler 1984). Such representations are particularly crucial for anticipating the sequential unfolding of events over time, allowing for the prediction of actions and outcomes within a chain of events (Abelson 1981). Integrating narratives into predictive modeling

means that information consistent with an individual's currently active narrative schema will be "explained away" in the fashion that Clark describes; events that were not predicted by the schema, on the other hand, will require more detailed processing and accommodation.

Incorporating narrative psychology into the hierarchical predictive processing account brings with it an important advantage. In particular, narratives provide a point of contact between the predictive processing account and the socio-cultural context in which individual minds develop. Narrative representations are fundamentally social in nature, as children are socialized to adopt particular modes of narrative thought during development (Nelson & Fivush 2004). An individual's personal narrative representations of the world are selectively constructed from the many social and public narratives that are available within the broader cultural context (Nelson 2003). In placing these narrative structures at the top of the predictive hierarchy, an individual's cultural context is afforded a powerful influence on the top-down regulation of domain-specific knowledge structures and behavioral patterns (Kitayama & Cohen 2010).

More broadly, this hybrid narrative predictive processing account highlights the relevance of the humanities for the cognitive sciences, suggesting a unified framework for their integration. A primary function of the humanities is elaborating upon the "manifest" image of the world as it is directly experienced by us, in contrast to the "scientific" image that provides a depersonalized view of the world (Sellars 1963). Narrative psychology acknowledges the importance of these "manifest" images, as they guide an individual's expectations and shape the cascade of cognitive operations that give rise to subjective experience. Within such a framework, a full appreciation of an individual's subjectivity is thus crucial to adequately modeling her construal of and reactions to the world.

Although higher-order narratives influence cognitive processes, the coherence of these narrative representations varies from person to person, with some having more clearly articulated stories for situating their experiences than others (McAdams 2006). A crucial consequence of this variation is that those with only vague narrative representations of the world will have more difficulty selectively focusing attention on the most relevant aspects of the environment. From a predictive processing perspective, a lack of narrative coherence will produce an inability to generate an adequate predictive model of the world, hindering the ability to "explain away" the majority of the sensory information being received and producing a burdensome processing load. When no high-level generative model is available to adequately anticipate the ongoing unfolding of events, the cognitive system can very easily be overwhelmed by the large volume of "error" information being carried up the neural hierarchy (Hirsh et al. 2012). This has downstream consequences for the individual, as a lack of personal narrative integration is associated with reduced well-being (Baerger & McAdams 1999). In contrast, developing clearly articulated narrative accounts of one's experiences is associated with a number of positive health benefits (Pennebaker & Seagal 1999).

Although the affective significance of prediction errors was not highlighted in Clark's article, the narrative account and its base of subjectivity makes this clear, as prediction errors can reflect violations of basic life assumptions. Such errors are often experienced as aversive and threatening (Hajcak & Foti 2008) and can trigger a variety of attempts to minimize or suppress error information (Proulx et al. 2012), some of which veer toward the pathological (Peterson 1999). The emotional impact of expectancy violations also appears to vary depending on the level of the neural hierarchy at which they occur, such that relatively low-level errors are experienced as fairly benign while violations of one's core narratives about the world are often associated with severe forms of emotional trauma (Janoff-Bulman 1992). Within the narrative framework, the ability to flexibly maintain the integrity of one's high-level generative models (instantiated as narrative representations) is thus one of the core requirements for mental health and well-being. Insomuch as the humanities help to provide us with narrative representations that capture the emotional vicissitudes of daily life in a given cultural environment (Oatley 1999), they help to orient and constrain our predictive modeling and provide critical components of our adaptive functioning in the world. Integrating narrative psychology with the predictive processing account thus highlights the importance of humanistic approaches for arriving at a complete understanding of human cognitive science.

Whenever next: Hierarchical timing of perception and action

doi:10.1017/S0140525X12002336

Linus Holm and Guy Madison

Department of Psychology, University of Umeå, 901 87 Umeå, Sweden.
linus.holm@psy.umu.se **guy.madison@psy.umu.se**
http://www.psy.umu.se/om-institutionen/personal/guy-madison

Abstract: The target article focuses on the predictive coding of "what" and "where" something happened and the "where" and "what" response to make. We extend that scope by addressing the "when" aspect of perception and action. Successful interaction with the environment requires predictions of everything from millisecond-accurate motor timing to far future events. The hierarchical framework seems appropriate for timing.

Timing intrinsically involves prediction. Determining when to act upon a future event requires the ability to predict it. For instance, ensemble music performance requires precise estimation of the passage of time in order to synchronize and coordinate sounds to re-produce the musical structure.

A central idea in the predictive coding account of cognition is that prior knowledge is used to guide sensory interpretations and action decisions. Identifying the periodicity of an event in the world is typically an ill-posed problem: How does the agent know beforehand what constitutes the signals that indicate a period? To infer the beat in a complex musical piece, or when a quail will reappear from behind a bush, are underspecified problems in the sensory signal. In both cases, prior experience appears necessary to play to the beat or to catch the quail.

Another key idea in the predictive coding framework is information compression. Representing music or other temporally structured events as cycles reduces the entropy in the signal and allows for more efficient storage. Action can serve to further bootstrap timing. For instance, humans spontaneously tap along with their hands or feet to music (Brown 2003) and entrain their movements to other people's movements (Demos et al. 2012; Merker et al. 2009). Just like active interactions with an object improve perception (Harman et al. 1999), timed activities have been shown to improve the reliability of temporal perception (Grahn & McAuley 2009; Phillips-Silver & Trainor 2007). A benefit of having induced the rhythm is that violations of rhythm are easier to detect (Ladinig et al. 2009).

Bayesian inference of timing requires temporal uncertainties to be represented. The nature of the timing signal remains open to debate. One candidate is trace strength that decays with time (Buhusi & Meck 2005). A function of decay, trace strength conveys information about the time since it occurred. Another time signal candidate is populations of oscillating neurons. Timing could then be established by coincidence detection in the oscillating network (Matell & Meck 2004; Miall 1989). Regardless of the signal format, its representation is noisy and its uncertainty should reasonably increase with timing over long

durations. Indeed, human temporal perception and production do deteriorate monotonically with time scale (Buhusi & Meck 2005). Exactly how the human system deals with temporal signal uncertainty remains an open question.

A key notion in the target article is the hierarchical division of labor from bottom sensory to top associative cortical control. For timing, the scaling of time appears as a likely attribute to stretch across such a hierarchical structure. Millisecond control of motor timing cannot feasibly be carried out directly by the prefrontal cortical regions involved in working memory, due to transfer speed, and the accumulated signal error that such an extensive chain of transmission would involve. Instead, millisecond control might be represented closer to the action output (e.g., cortical effector representation and the cerebellum) and involve a more direct pathway between sensory input and motor output. In contrast, when observation and action become more detached in time, the window of opportunity for planning opens up, involving more prefrontal processing.

Consistently, many studies support the view that there is a distinction in neural representation, for example, above and below about one second (Gooch et al. 2001; Lewis & Miall 2003; Madison 2001). Furthermore, time representation for subsecond intervals appears at least to some extent to be sensory specific (Morrone et al. 2005; Nagarajan et al. 1998), and under some conditions even limited to spatial locations (Burr et al. 2007; Johnston 2006). Additionally, there appear to be breakpoints in interval discrimination such that there are scalar properties in timing performance for intervals above about one second, but nonlinear relationships between time and perception below one second (Karmarkar & Buonomano 2007; Rammsayer 1999) – further supporting the notion that longer time intervals are controlled by different brain regions from those involved in sub-second timing. Also, with longer time periods under consideration, a larger part of the prefrontal cortex gets activated (Lewis & Miall 2006; Simons et al. 2006). This timing-related frontal lobe network is also largely overlapping with those employed by working memory and executive control processes (Jahanshahi et al. 2000; Owen et al. 2005), suggesting that timing constitutes a general cognitive control problem at longer time durations. The hierarchical organization from accurate and dedicated timing devices at sensory levels and less accurate but flexible timing at longer time frames in the prefrontal cortex might be accounted for by signal averaging in the time domain from sensory to frontal cortical regions (Harrison et al. 2011). Harrison and colleagues suggested that decay rate is faster close to the sensory input level and slower at later stages in the visual hierarchy, thus allowing for a differentiation across time scale and brain region. Taken together, there is abundant support for the differentiation of brain regions involved in timing at different time scales.

Communication of temporal information across the levels of the outlined timing hierarchy is currently rather unclear. Intuitively, the more temporally extended control processes associated with prefrontal working memory processes might still influence control at shorter time frames without interfering in direct control, such as in initiation of a drumming exercise, without employing moment to moment volitional control of the individual beats. Recent findings from our research group suggest that executive functions are indirectly related to motor timing via, for example, effector coordination (Holm et al., in press). Furthermore, there is a well-established yet poorly specified relationship between intelligence and simple motor timing (Galton 1883; Madison et al. 2009). More research is clearly needed to identify how high-level temporal expectations might influence brief interval timing. Another important question is how the brain identifies the time scales from noisy input and learns how to treat those signals. The predictive account of cognition seems like a useful theoretical framework for understanding timing, and the Bayesian formalism is a promising tool to investigate and explain its operation.

Two kinds of theory-laden cognitive processes: Distinguishing intransigence from dogmatism

doi:10.1017/S0140525X12002403

Elias L. Khalil

Department of Economics, Monash University, Clayton, Victoria 3800, Australia.

elias.khalil@monash.edu www.eliaskhalil.com

Abstract: The brain is involved in theory-laden cognitive processes. But there are two different theory-laden processes. In cases where the theory is based on facts, more facts can either falsify or confirm a theory. In cases where the theory is about the choice of a benchmark or a standard, more facts can only make a theory either more or less warranted.

Clark offers a review of a view of the brain where the brain processes input information in a way that confirms its priors or its predictions. This does not mean that the brain creates its own reality. The brain, rather, processes input data, but it does so in light of its own priors. The brain is a bidirectional hierarchical structure. While the top layers generate priors, the lower layers process input data. The brain amounts to the dynamics of image-making, where the top-down process generates unified images, while the bottom-up process, which takes data, corrects the images.

Such an iterative cognitive process is not simple. The top-layer generated priors greatly determine the assimilated inputs. But the input data are not fully manipulated by the priors. As such, it is best to characterize the brain as a medium that tries to balance between two competing needs: First, the brain needs to generate a unified, that is, meaningful, image of the real world. The top layers, which generate the priors or the predictions, function to fulfill the need for unity. Second, the brain needs to accommodate raw input data to stay as truthful as possible to the given real world. If the brain performs only the first function, that is, preserving the unity of the image, the brain would generate images that, although unified, are disconnected from reality. On the other hand, if the brain performs only the second function, that is, preserving the details of the world, the brain would generate images that, although detailed, are tremendously messy and meaningless.

As a result of trying to meet these two competing needs, the images that cognitive processes generate are theory-laden. This has long been understood by the emerging new philosophy of science, most epitomized by the contribution of Thomas Kuhn, and can even be traced to Immanuel Kant. This is not the place to review the history of philosophy of science, characterized ultimately as a conflict between rationalism (demanding unity of image) and empiricism (demanding detailed images) (see Khalil 1989). What is germane here is that Clark fails to note two different kinds of theory-laden cognitive processes: the first, which can be called "perception-laden" processes, where one's theory can be ultimately corrected by sensory input; the second, which can be called "conception-laden" processes, where one's theory cannot be ultimately corrected by sensory input.

Perception-laden beliefs, for example, let one predict stormy weather or that the Earth is flat. In light of sensory input, and using Bayes' rule, one may adjust such a prediction and reach the conclusion that the weather will be stable and the Earth is round. Many people may not adjust quickly and insist on "explaining away" the data to justify their priors. But such manipulation can be delineated from the normal course of belief adjustment. When perception-laden processes are at issue, priors must ultimately adjust to correspond to the mounting evidence. The legal system, and everyday science, cannot function without the adherence to the possibility of belief-free grounds that can allow sensory data, in the final analysis, to dominate top-down priors.

Conception-laden beliefs, for example, let one view a picture such as the famous Rubin Vase, where the brain switches between perceiving the vase and perceiving the two profiles. The image depends on what the brain judges to be the background. If the background is judged to be white, the brain sees the two profiles. If the background is judged to be black, the brain sees the vase. No amount of data can compel the top level hierarchy of the brain to abandon its prior. The prior here cannot be confirmed or refuted by evidence because it is *not* based on evidence as with perception-laden processes. The choice of background, the basis of conception, is similar to the choice of a benchmark, where one can judge a glass to be either half-full or half-empty. Likewise, one judges one's income as satisfactory or non-satisfactory depending on one's benchmark. Happiness seems to depend, at least partially, on the choice of an arbitrary income as the benchmark income.

The conflation of the perception- and conception-laden processes leads to the commitment of a Bayesian fallacy. The fallacy arises from the supposition that all beliefs are perception-laden and, hence, can be corrected by further empirical investigation (Khalil 2010). It is imperative to distinguish conceptions from perceptions. Aside from allowing us to understand happiness, the distinction sheds light on two kinds of stubbornness: intransigence, related to perception-laden beliefs, and dogmatism, related to conception-laden beliefs. Belief in a flat Earth and in conspiracy theories illustrates intransigence. In contrast, to insist on a background, despite the rising evidence to the contrary, illustrates dogmatism. To use the Rubin Vase example, if a person chooses the black as the background and, hence, the image is the vase, but continues to choose the black despite contrary added evidence – such as added eyes and moustache – the person would be dogmatic. While the dogmatic belief cannot be judged as true or false, it can be judged as warranted or unwarranted given the details of the profiles. The choice of background, to remind ourselves, is non-empirical and, hence, cannot be characterized as true or false.

Predictions in the light of your own action repertoire as a general computational principle

doi:10.1017/S0140525X12002294

Peter König,[a,b] Niklas Wilming,[a] Kai Kaspar,[a] Saskia K. Nagel,[a] and Selim Onat[c]

[a]*Institute of Cognitive Science, University Osnabrück, 49076 Osnabrück, Germany;* [b]*Department of Neurophysiology and Pathophysiology, University Medical Center Hamburg-Eppendorf, 20251 Hamburg, Germany;* [c]*Department of Systems Neuroscience, University Medical Center Hamburg-Eppendorf, 20251 Hamburg, Germany.*

koenig@uni-osnabrueck.de **nwilming@uni-osnabrueck.de**
kkaspar@uni-osnabrueck.de **snagel@uni-osnabrueck.de**
sonat@uos.de
http://cogsci.uni-osnabrueck.de/~NBP/
http://cogsci.uni-osnabrueck.de/~nwilming/
http://kai-kaspar.jimdo.com/
http://cogsci.uni-osnabrueck.de/en/changingbrains/people/saskia
www.selimonat.com

Abstract: We argue that brains generate predictions only within the constraints of the action repertoire. This makes the computational complexity tractable and fosters a step-by-step parallel development of sensory and motor systems. Hence, it is more of a benefit than a literal constraint and may serve as a universal normative principle to understand sensorimotor coupling and interactions with the world.

Present cognitive science is characterized by a dichotomy separating sensory and motor domains. This results in a perceived gap between perception and action and is mirrored in leading theories of cognition. For illustration we consider the visual neurosciences, a paradigmatic field for the investigation of sensory processes. A discourse given by standard textbooks depicts a world external to the agent, with a set of pre-established attributes and objects. Sensory processing starts with transmitting these attributes by low-level neurons to subsequent stages. There, more elaborate computations extract patterns of stimulus features and objects. Up to this point, processing focuses on a veridical representation of the external world, serving for later decisions and actions. We argue in favor of a radical change of this view, assigning a central role to predictions of sensory consequences of one's own actions and thereby eliminating the strict separation of sensory and motor processing.

In the target article, Andy Clark beautifully describes the central role of predictions in sensory processing. We endorse this view – yet two complementary aspects are needed. First, predictability of sensory signals serves as a normative principle guiding sensory processing and as a boundary constraint in the selection of information to process. Second, predictions are performed only in the context of the agents' action repertoire (König & Krüger 2006). These two specifications have crucial implications.

The information content of the primary sensory signal is enormous, and extraction of information without further constraint is an ill-posed problem. However, it is not the task of the sensory systems to process all possible details, and a reduction of information is paramount. Even in simple model systems, taking into account a limited behavioral repertoire converts demanding sensory processing into a tractable problem (Wyss et al. 2004). Applying the normative principle of predictability generalizes this idea and serves as a selection criterion for features to process and variability to ignore. Indeed, within the hierarchy of the visual system, neuronal response properties are invariant to more and more parametric changes of the sensory input (Tanaka 1996). Even category learning at higher levels of the visual system can be interpreted within this framework. The commonalities between different instances of the same category relate to similar sensorimotor patterns generated by the interaction with these "objects." Finally, actions are directly related to the agent's survival and thereby processing features that change predictably, given chosen actions, are more relevant than those that do not. Hence, processing of sensory signals is guided by the relevance for behavior, and relevance is expressed by the ability to predict sensory changes contingent on the own action repertoire.

A paradigm is based on the active interpretation of incoming sensory information such that it makes sense for the agent. Hence, it is intended to replace a passive representationalist view. In such a paradigm, the predicted future state of the world is important insofar as it interacts with own actions and variables of importance are co-determined by the action repertoire. A demonstration of the integration of new sensory information (magnetic north) that is co-determined by own movements (yaw-turns) is given by the feelSpace project (Kärcher et al. 2012; Nagel et al. 2005). Comparing different species, for example, cat and human, with similar visual input (Betsch et al. 2004; Einhäuser et al. 2009), the remarkable differences in the sensory hierarchy appear to be at odds with a passive representationalist view and await an explanation. Here, differences in behavioral repertoire offer themselves. Pointedly, we speculate that the huge action repertoire of humans, due to, for example, opposable thumbs, might foster the illusion of a veridical perception of the world. It has been emphasized early on that cognitive and motor capabilities develop in parallel and mutual dependence (Piaget 1952). To grow up means to harden specific action routines, on the one hand, but to lose the bulk of alternative action capabilities and cognitive flexibility, on the other hand. Furthermore, a large variability of perceptual interpretation of identical physical stimuli is found between humans of the same culture area as well as between different cultures (Segall et al. 1963). A critical view of our own culture reveals many aspects that serve

to increase the reliability of predictions. In summary, agents with identical sensory organs but different action repertoires might have very different views of the world.

Is the concept of normative principles plausible in view of our knowledge of cortical networks? Neuronal computations are constrained by properties of the brain in the form of number of neurons and synapses, and space and energy consumption. The latter has served as an argument for sparse coding – that is, low mean activity at constant variance of activity (Barlow 1961). The insight that receptive fields of simple cells in primary visual cortex form such an optimally sparse representation of natural images drastically increased interest in normative models (Olshausen & Field 1996; Simoncelli & Olshausen 2001). Properties of the second major neuron type in primary visual cortex, complex cells, can be understood along similar lines as optimizing stable representations (Berkes & Wiskott 2005; Körding et al. 2004). Importantly, both optimization principles can be easily implemented by recurrent connectivity within a cortical area (Einhäuser et al. 2002). Hence, existing normative models of the early visual system are plausible in view of anatomical and physiological data.

A critical test of the concept will be the application well beyond processing in the primary visual cortex. The step from sparseness and stability to predictability as an optimization principle requires critical extensions. Phillips et al. (1995) put forward a very promising proposal: Coherent infomax selects and coordinates activities as a function of their predictive relationships and current relevance. The relation of this approach (see Phillips' commentary in this issue) to the free energy principle (Friston 2010) and optimal predictability (König & Krüger 2006) has to be investigated. These developments hold the promise to apply to "higher" cognitive functions as well as giving rise to a true theory of cognitive science.

Maximal mutual information, not minimal entropy, for escaping the "Dark Room"

doi:10.1017/S0140525X12002415

Daniel Ying-Jeh Little and Friedrich Tobias Sommer

Redwood Center for Theoretical Neuroscience, University of California–Berkeley, Berkeley, CA 94720-3198.

dylittle@berkeley.edu **fsommer@berkeley.edu**

http://redwood.berkeley.edu/wiki/Daniel_Little

http://redwood.berkeley.edu/wiki/Fritz_Sommer

Abstract: A behavioral drive directed solely at minimizing prediction error would cause an agent to seek out states of unchanging, and thus easily predictable, sensory inputs (such as a dark room). The default to an evolutionarily encoded prior to avoid such untenable behaviors is unsatisfying. We suggest an alternate information theoretic interpretation to address this dilemma.

We would like to compliment Clark for his comprehensive and insightful review of the strengths and limitations of hierarchical predictive processing and its application to modeling actions as well as perception. We agree that the search for fundamental theoretical principles will be key in explaining and uniting the myriad functions of the brain. Here, we hope to contribute to the discussion by reconsidering a particular challenge to the minimum prediction error (MPE) principle identified by Clark, which we dub the "Dark Room Dilemma," and by offering an alternate solution that captures both the drive to reduce errors and the drive to seek out complex and interesting situations.

As described by Clark, a common challenge to extending the principle of minimum prediction error (MPE) to action selection is that it would drive an animal to seek out a dark room where predicting sensory inputs becomes trivial and precise. In response, Clark suggests that "animals like us live and forage in a changing and challenging world, and hence 'expect' to deploy quite complex 'itinerant' strategies" (sect. 3.2, para. 2). At first, this response seems tautological: We act so that we can predict the outcome of our actions; we predict that our actions will be complex and interesting; and therefore we act in complex and interesting ways. The tautology is broken by invoking a prior expectation on action, one presumably hardwired and selected for by evolutionary pressures. But, such an assumption would seem to remove the explanatory power of the MPE principle in describing complex behaviors. Furthermore, it goes against the common view that the evolutionary advantage of the brain lies in the ability to be adaptive and alleviate much of the need for hardwired pre-programming (pre-expectations) of behavior. A more satisfying solution to the "Dark Room Dilemma" may potentially be found in a different information theoretic interpretation of the interaction between action and perception.

Clark turns to the free-energy formulation for an information theoretic interpretation of the MPE principle (Friston & Stephan 2007). Within this framework, average prediction error is captured by the information theoretic measure entropy, which quantifies an agent's informational cost for representing the sensory input by its internal model. An alternative quantification of the predictive accuracy of an internal model would be to consider its mutual information (MI) with the sensory inputs. MI quantifies the information shared between two distributions – in this case, the informational content the internal states of the brain hold regarding its future sensory inputs. MI and entropy are in a sense converses of one another. Entropy is the informational cost of a (bad) internal model, while MI is the informational gains of a (good) internal model. When selecting a model, minimizing entropy and maximizing MI both yield minimal prediction error. When selecting actions, however, these two principles yield very different results.

Actions allow an agent, through the sensor-motor loop, to change the statistics of its sensory inputs. It is in response to such changes that the principles of maximizing MI and minimizing entropy differ. This difference can be highlighted by a hypothetical extreme, in which an agent acts to remove all variation in its sensory inputs – that is, it dwells in a "Dark Room." Here, a trivial model can perfectly predict sensory inputs without any information cost. Entropy thus goes to zero satisfying the principle of minimal entropy. Similarly, MI also goes to zero in a Dark Room. Without variation in sensory inputs there is no information for the internal model to try to capture. This violates the maximal MI principle. Instead, of entering a "Dark Room," an agent following a principle of maximal MI would seek out conditions in which its sensory inputs vary in a complex, but still predictable, fashion. This is because MI is bounded below by the variability in sensory input and bounded above by its ability to predict. Thus, MI balances predictability with complexity. Passively, maximizing MI accomplishes the same objective as minimizing entropy, namely the reduction of prediction error, but actively it encourages an escape from the Dark Room.

The prediction–complexity duality of MI and its importance to learning has been a recurring finding in computational methods. Important early implementations of a maximal MI principle in modeling passive learning include the Computational Mechanics approach for dynamical systems of Crutchfield and Young (1989) and the Information Bottleneck Method of Tishby et al. (1999) for analyzing time series. Recently, the Information Bottleneck method has been extended to action selection by Still (2009). Further, the Predictive Information Model of Ay et al. (2008) has shown that complex behaviors can emerge from simple manipulations of action controllers towards maximizing the mutual information between states. And our own work utilizes MI to drive exploratory behaviors (Little & Sommer 2011).

The principle of minimum prediction error and the related hierarchical prediction models offer important insights that should not be discounted. Our aim is not to suggest otherwise. Indeed, we favor the view that hierarchical prediction models could explain the motor implementation of intended actions. But we also believe its explanatory value is limited. Specifically, it would be desirable for a theoretical principle of the brain to address and not spare the intriguing question of what makes animals, even the simplest ones, venture out of their dark rooms.

Backwards is the way forward: Feedback in the cortical hierarchy predicts the expected future

doi:10.1017/S0140525X12002361

Lars Muckli, Lucy S. Petro, and Fraser W. Smith

Centre for Cognitive Neuroimaging, Institute of Neuroscience and Psychology, University of Glasgow, Glasgow G12 8QB, United Kingdom.

Lars.Muckli@glasgow.ac.uk **lucyp@psy.gla.ac.uk**
Fraser.Smith@glasgow.ac.uk **http://muckli.psy.gla.ac.uk/**

Abstract: Clark offers a powerful description of the brain as a prediction machine, which offers progress on two distinct levels. First, on an *abstract* conceptual level, it provides a unifying framework for perception, action, and cognition (including subdivisions such as attention, expectation, and imagination). Second, hierarchical prediction offers progress on a *concrete* descriptive level for testing and constraining conceptual elements and mechanisms of predictive coding models (estimation of predictions, prediction errors, and internal models).

Abstract level description. Understanding the brain as a prediction machine offers a compelling framework for perception, action, and cognition. Irrespective of the neuronal implementation, the framework ascribes a function to internal models and neuronal processes to best prepare for the anticipated future. At an abstract level, the predictive coding framework also draws attention to two blind spots in neuroscience: (1) internal cortical communication (i.e., maintaining internal models) and (2) the brain processes prior to stimulation onset (i.e., predictive processing).

A starting point to explore internal communication is by investigating cortical feedback (Van Essen 2005; Muckli & Petro 2013). Conventional paradigms struggle, however, to isolate cortical feedback during sensory processing (which includes both feedforward and feedback information). We have demonstrated such separation by blocking feedforward stimulation using visual occlusion and reading out rich information content (multivariate patterns) from within non-stimulated regions of the retinotopic cortex (which receive cortical feedback activation; Muckli & Petro 2013; Smith & Muckli 2010). By decoding cortical feedback, we begin to shed light on internal processing. With regard to investigating brain processes prior to stimulation onset, we have shown that motion predictions are carried over to new retinal positions after saccadic eye-movements (Vetter et al. 2012), which confirms that saccadic updating incorporates predictions generated during pre-saccadic perception. This is an important proof of concept of predictive coding in saccadic viewing conditions. Moreover, Hesselmann et al. (2010), have shown that variations in baseline activity influence subsequent perception, and a causal role of V5 in generating predictions sent to V1 can be demonstrated using transcranial magnetic stimulation (TMS). Pilot data show that TMS interferes with predictive codes during the baseline prior to stimulation onset (Vetter et al., under revision). If the brain would be seen as a "representation machine" instead of a "prediction machine," one would not look for predictive brain processing before stimulus onset and important information about cortico-cortical communication would remain concealed. Motivating the search for predictive signals in the system is therefore another important contribution of the conceptual framework.

Concrete level description. On the *concrete* conceptual level, hierarchical cortical prediction provides a scaffold on which we can constrain variants of predictive coding models. Predictions are proposed to explain away the incoming signal or filter away the unexpected noise (Grossberg 2013). Rao and Ballard (1999) proposed a model in which forward connections convey prediction errors only, and internal models are updated on the basis of the prediction error (Rao & Ballard 1999). Grossberg on the other hand proposes Adaptive Resonance Theory (ART) models that update internal models based on recognition error. It remains an empirical question which combination of these models suffices to explain the rich and diverse cortical response properties. A recent brain imaging study shows that under conditions of face repetition, some voxels show repetition suppression consistent with the concept that the prediction error is reduced with every repetition of the identical image, while others (30%) show repetition enhancement (De Gardelle et al. 2012). Repetition enhancement in a subpopulation of fusiform face area (FFA) voxels could reinforce the internal model of the face identity and be used to stabilize the prediction. The claim that the brain is a prediction machine might be true regardless of the precise implementation of predictive coding mechanism. Internal models might update on error, stabilize on confirmation or scrutinize on attention (Hohwy 2012). A recent brain imaging study investigated whether expectation induced signal suppression coincides with sharpening of the underlying neuronal code (Kok et al. 2012). Consistent with the predictive coding framework, auditory-cued stimuli led to reduced V1 fMRI activity. Although the overall activity was reduced, the activation profile was more distinct, "sharpened," for the expected conditions as measured using multivariate decoding analysis. The study concludes that expectation helps to explain away the signal while attention amplifies the remaining prediction error (Hohwy 2012; Spratling 2008b).

Another concrete level aspect of predictive coding relates to the question of spatial precision. Are the back-projected predictions at the precision level of the "sending" brain area (i.e., coarse), or at the precision level of the "receiving" brain area (i.e., spatially precise)? We have evidence in favor of both; V5 feedback signals spread out to a large region in primary visual cortex (de-Wit et al. 2012; Muckli et al. 2005) but spatio-temporal predictions in V1 which have been relayed by V5 can also be spatially precise (Alink et al. 2010). The optimal way to account for this discrepancy is by assuming an architecture that combines coarse feedback with the lateral spread of feedforward signals (Erlhagen 2003). If this principle holds true, it helps to explain why the architecture of cortical feedback as described by Angelucci et al. (2002) contributes to precise predictions even though it is divergent.

The examples above show that on an abstract level important new research is motivated by the hierarchical predictive coding framework and on a concrete conceptual level, the many interactions of cortical feedback of predictions, processing of prediction errors, and different accounts of feedforward connections (some stabilizing the internal model, others explaining away signal discrepancies) await further empirical scrutiny. However, the developing narrative of predictive coding becomes increasingly compelling with attention from sophisticated human neuroimaging and animal neurophysiological studies (Muckli & Petro 2013). Not only is extending our knowledge of cortical feedback and its encapsulated predictions essential for understanding cortical function, but important opportunities will arise to investigate deviations of predictive coding in aging and neuropsychiatric diseases such as schizophrenia (Sanders et al. 2012).

Skull-bound perception and precision optimization through culture

doi:10.1017/S0140525X12002191

Bryan Paton,[a] Josh Skewes,[b] Chris Frith,[c] and Jakob Hohwy[a]

[a]*Philosophy and Cognition Laboratory, Philosophy Department, Monash University, Clayton, VIC3800, Australia;* [b]*Department of Culture and Society, Aarhus University, DK8000 Aarhus C, Denmark, and Interacting Minds Centre, Aarhus University Hospital, DK8000 Aarhus C, Denmark;* [c]*Institute of Neurology, University College London, London, WC1E 6BT, and All Souls College, Oxford University, Oxford OX1 4AL, United Kingdom.*

Bryan.Paton@monash.edu **filjcs@hum.au.dk** **c.frith@ucl.ac.uk**
Jakob.Hohwy@monash.edu
https://sites.google.com/site/bryanpaton/home
http://www.cfin.au.dk/menu538-en
https://sites.google.com/site/chrisdfrith/Home
https://sites.google.com/site/jakobhohwy/

Abstract: Clark acknowledges but resists the indirect mind–world relation inherent in prediction error minimization (PEM). But directness should also be resisted. This creates a puzzle, which calls for reconceptualization of the relation. We suggest that a causal conception captures both aspects. With this conception, aspects of situated cognition, social interaction and culture can be understood as emerging through precision optimization.

Andy Clark acknowledges the "challenging vision" of prediction error minimization (PEM), according to which representation is inner and skull-bound such that perception is a fantasy that coincides with reality (Frith 2007). This view does not require homunculi and sense-data but does convey a somehow *indirect* mind–world relation.

Clark resists indirectness. He states that PEM "makes structuring our worlds genuinely continuous with structuring our brains and sculpting our actions" (sect. 3.4, para. 1), and that "*what* we perceive is not some internal representation or hypothesis but (precisely) the world" (sect. 4.4, para. 3, emphasis Clark's).

The sentiment is right, but caution about directness is needed. Without indirectness we ignore how the mind is always precariously hostage to the urge to rid itself of prediction error. This urge forces very improbable and fantastical perceptions upon us when the world does not collaborate in its usual, uniform way. For example, in the contemporary swathe of rubber-hand and full-body illusions, we easily and compellingly experience having a rubber hand (or two), occupying another's body or a little doll's body, or having magnetic forces or spectral guns operating on our skin (Hohwy & Paton 2010; Lenggenhager et al. 2007; Petkova & Ehrsson 2008). Moreover, more stable and fundamental aspects of mind, such as our sense of agency, privileged access to self, and mentalizing, all seem to make sense only in terms of perceptual fantasizing (Frith 2007).

This leaves a puzzle. On PEM, the perceptual relation cannot be direct. But neither is it wholly indirect. The challenge is then to reconceive the mind–world relation to encompass both aspects. We suggest a causal conception, and use its internal aspect to leverage an understanding of situated and social cognition.

The implicit inversion of a generative model happens when prediction error is minimized between the *model* maintained in the brain and the *sensory input* (how the world impinges on the senses). This yields causal inference on the hidden *causes* (the states of affairs in the world) of the sensory input. This is a distinctly causal conception of how the brain recapitulates – provides a multilayered mirror image of – the causal structure of the world. This representational relation is *direct* in the sense that causation is direct: There is an invariant relation between the model and world, such that, given how the model is, it changes in certain ways when the world changes in certain ways. But, seen from the inside, there is *indirectness* in the sense that causal relata are distinct existences, giving rise to a need for causal inference on hidden, environmental causes.

Though the brain can optimize precisions on its prediction error, it is hostage to the causal link from environmental causes to sensory input. If the variance in the signal from the world to the senses is large, then there is only so much the brain can do there and then to ensure optimal encoding. Precisely because the mind is destined to be behind the veil of sensory input, it then makes sense for it to devise ways of optimizing the information channel from the world to the senses. Thus, through active inference prediction error is minimized, not only by selective sampling, but also by optimizing its precision: removing sources of noise in the environment and amplifying sensory input.

Many of the technical, social and cultural ways we interact with the world can be characterized as attempts to make the link between sensory input and environmental causes less volatile. We see this in the benefits of the built environment (letting us engage in activities unperturbed by wind and weather), in technical and electronic devices (radio lets us hear things directly rather than through hearsay), and in language (communicating propositional content). This picture relies on the internal nature of the neural mechanism that minimizes prediction error, relative to which all our cultural and technological trappings are external. Culture and technology situate the mind closer to the world through improving the reliability of its sensory input. But perception remains an inferred fantasy about what lies behind the veil of input.

By maintaining focus on the internal nature of perceptual processes, in this causal setting, we can appreciate another perspective on social interaction and culture than the "mutual prediction error reduction" that Clark rightly points to.

As Locke insisted, communication is the sharing of each other's hidden ideas. Ideas are well-hidden causes, so PEM is the tool for inferring them through a mix of prediction ("after saying A, he tends to say B") and active inference (asking something to elicit a predicted answer). An overlooked aspect here is how this is facilitated not just by representing the other's mental states but also by aligning our mental states with each other in a process of neural hermeneutics – a fusion of expectation horizons. We do this, not to change the sensory input itself, but to enhance the precision with which we can probe each other's current mental states, perhaps to such an extent that the receiver in a social interaction ends up having more precise information about the sender's mental states than the sender him- or herself (Frith & Wentzer, in press).

Perhaps culture too, in a very wide sense, can be seen as, at least partly, a tool for precision optimization through shared context. Ritual, convention, and shared practices enhance mutual predictability between people's hidden mental states. This would make sense of cultural diversity because this process is concerned with signal reliability rather than with what the signals are about, and there are many different ways of using cultural tools to align our mental states. Furthermore, when precision has been optimized, alignment enables simple, information rich signaling and thereby communication efficiency.

If alignment of mental states is an integral part of how culture optimizes precision and communication efficiency, then culture should be seen as providing a set of frameworks for interpretation, rather than merely for scaffolding interpretation. If the brain is a hierarchical Bayesian network providing a perceptual fantasy of the world, then culture determines and constrains the hyperpriors needed by such a neural system.

Neuronal inference must be local, selective, and coordinated

doi:10.1017/S0140525X12002257

William A. Phillips

Psychology Department, University of Stirling, FK9 4LA Stirling, Scotland, United Kingdom, and Frankfurt Institute of Advanced Studies, 60438 Frankfurt am Main, Germany.

wap1@stir.ac.uk
http://www.psychology.stir.ac.uk/staff/staff-profiles/honorary-staff/bill-phillips

Abstract: Life is preserved and enhanced by coordinated selectivity in local neural circuits. Narrow receptive-field selectivity is necessary to avoid the curse-of-dimensionality, but local activities can be made coherent and relevant by guiding learning and processing using broad coordinating contextual gain-controlling interactions. Better understanding of the functions and mechanisms of those interactions is therefore crucial to the issues Clark examines.

Much in Clark's review is of fundamental importance. Probabilistic inference is crucial to life in general and neural systems in particular, but does it have a single coherent logic? Jaynes (2003) argued that it does, but for that logic to be relevant to brain theory, it must be shown how systems built from local neural processors can perform essential functions that are assumed to be the responsibility of the scientist in Jaynes' theory (Fiorillo 2012; Phillips 2012).

Most crucial of those functions are selection of the information relevant to the role of each local cell or microcircuit and coordination of their multiple concurrent activities. The information available to neural systems is so rich that it cannot be used for inference if taken as a single, multi-dimensional whole because the number of locations in multi-dimensional space increases exponentially with dimensionality. Most events that actually occur in high-dimensional spaces are therefore novel and distant from previous events, precluding learning based on sample probabilities. This constraint, well-known to the machine-learning community as the *curse-of-dimensionality*, has major consequences for psychology and neuroscience. It implies that for learning and inference to be possible large data-bases must be divided into small subsets, as amply confirmed by the clear selectivity observed within and between brain regions at all hierarchical levels. Creation of the subsets involves both prespecified mechanisms, as in receptive field selectivity, and dynamic grouping as proposed by Gestalt psychology (Phillips et al. 2010). The criteria for selection must be use-dependent because information crucial to one use would be fatal to another, as in the contrast between dorsal and ventral visual pathways. Contextual modulation is also crucial because interpretations with low probability overall may have high probability in certain contexts. Therefore, the activity of local processors must be guided by the broader context, and their multiple concurrent decisions must be coordinated if they are to create coherent percepts, thoughts, and actions.

Most models of predictive coding (PC) and Bayesian inference (BI) assume that the information to be coded and used for inference is a given. In those models, it is – by the modelers. Modelers may assume that in the real world this information is given by the external input, but that provides more information than could be used for inference if taken as a whole. Self-organized selection of the information relevant to particular uses is therefore crucial. Efficient coding strategies, such as PC, are concerned with ways of transmitting information through a hierarchy, not with deciding what information to transmit. They assume lossless transmission of all input information to be the goal, and so provide no way of extracting different information for different uses. Models using BI show how to combine information from different sources when computing a single posterior decision; but they do not show how local neural processors can select the relevant information, nor do they show how multiple streams of processing can coordinate their activities. Thus, local selectivity, dynamic-grouping, contextual-disambiguation, and coordinating interactions are all necessary within cognitive systems, but are not adequately explained by the essential principles of either PC or BI.

Clark's review, however, does contain the essence of an idea that could help resolve the mysteries of selectivity and coordination, that is, context-sensitive gain-control, for which there are several widely-distributed neural mechanisms. A crucial strength of the free-energy theory is that it uses gain-controlling interactions to implement attention (Feldman & Friston 2010), but such mechanisms can do far more than that. For example, they can select and coordinate activities by amplifying or suppressing them as a function of their predictive relationships and current relevance. This is emphasized by the theory of Coherent Infomax (Kay et al. 1998; Kay & Phillips 2010; Phillips et al. 1995), which synthesizes evidence from neuroanatomy, neurophysiology, macroscopic neuroimaging, and psychophysics (Phillips & Singer 1997; von der Malsburg et al. 2010). That theory is further strengthened by evidence from psychopathology as reviewed by Phillips and Silverstein (2003), and extended by many subsequent studies. Körding and König (2000) argue for a closely related theory.

Free-energy theory (Friston 2010) and Coherent Infomax assume that good predictions are vital, and formalize that assumption as an information theoretic objective. Though these theories have superficial differences, with Coherent Infomax being formulated at the neuronal rather than the system level, it may be possible to unify their objectives as that of maximizing prediction success, which, under plausible assumptions, is equivalent to minimizing prediction error (Phillips & Friston, in preparation). Formulating the objective as maximizing the amount of information correctly predicted directly solves the "dark-room" problem discussed by Clark. That objective, however, does not necessarily imply that prediction errors are the fundamental currency of feedforward communication. Inferences could be computed by reducing prediction errors locally, and communicating inferences more widely (Spratling 2008a). That version of PC is supported by much neurobiological evidence, though it remains possible that neural systems use both versions.

Another important issue concerns the obvious diversity of brains and cognition. How could any unifying theory cast light on that? Though possible in principle, detailed answers to this question are largely a hope for the future. Coherent Infomax hypothesizes a local building-block from which endlessly many architectures could be built, but use of that to enlighten the obvious diversity is a task hardly yet begun. Similarly, though major transitions in the evolution of inferential capabilities seem plausible, study of what they may be remains a task for the future (Phillips 2012). By deriving algorithms for learning, Coherent Infomax shows in principle how endless diversity can arise from diverse lives, and it has been shown that the effectiveness of contextual-coordination varies greatly across people of different ages (Doherty et al. 2010), sex (Phillips et al. 2004), and culture (Doherty et al. 2008). Use of this possible source of variability to enlighten diversity across and within species still has far to go, however.

Overall, I expect theories such as those examined by Clark to have far-reaching consequences for philosophy, and human thought in general, so I fully endorse the journey on which he has embarked.

God, the devil, and the details: Fleshing out the predictive processing framework

doi:10.1017/S0140525X12002154

Daniel Rasmussen and Chris Eliasmith

Centre for Theoretical Neuroscience, University of Waterloo, Waterloo, ON N2L 3G1, Canada.

drasmuss@uwaterloo.ca **celiasmith@uwaterloo.ca**

Abstract: The predictive processing framework lacks many of the architectural and implementational details needed to fully investigate or evaluate the ideas it presents. One way to begin to fill in these details is by turning to standard control-theoretic descriptions of these types of systems (e.g., Kalman filters), and by building complex, unified computational models in biologically realistic neural simulations.

God is in the details
— Mies van der Rohe
The devil is in the details
— Anonymous

Despite their theologically contradictory nature, both of these statements are true: the first is noting that details are important, and the second that getting the details right is difficult. It is for exactly this pair of reasons that we believe the predictive processing framework is limited in its ability to contribute, in a deep way, to our understanding of brain function.

This is not to deny that the brain does prediction. This is a view that has been beautifully articulated by Clark, and lies in a great tradition. For instance, in his 1943 book, Kenneth Craik devotes several chapters to his central hypothesis that: "One of the most fundamental properties of thought is its power of predicting events" (Craik 1943, p. 50). The evidence for prediction-related signals is strong, and the high-level models are often tantalizing. However, we (and, in our experience, most neuroscientists) want more: We want specific neural mechanisms that are employed in specific circumstances, and we want to know how such models can be arranged to explain complex behavior (i.e., we want an architectural specification).

Unfortunately, as Clark himself points out, the predictive processing framework "fail[s] to specify the overall form of a cognitive architecture" and "leaves unanswered a wide range of genuine questions concerning the representational formats used by different brain areas" (sect. 3.3, para. 4). The extent of the predictive processing framework's architectural claims is that the brain is organized in a hierarchical manner, with error signals passing up the hierarchy and predictions of world state passing down. However, this description seems to miss all the interesting details: What is the specific form and function of the connections between levels of this hierarchy? In the human brain, along what neuroanatomical pathways should we expect to see this information flowing? And, more generally, how do different hierarchies interact? How does information pass between them? Is there a unifying representational format? The predictive processing framework leaves all of these details unspecified, but it strikes us that the filling-in of these details is where the framework would gain deep, empirical content.

It may seem as if some of these questions are answered. For instance, the primary method of representation in the brain is supposed to be through probability density functions across the possible states/concepts. However, as Clark mentions, these representations could be implemented with a "wide variety of different schemes and surface forms" (sect. 3.2, para. 4). For example, a probability density $p(x)$ could be represented as a histogram (which explicitly stores how many times each state x has occurred) or as a summary model (e.g., storing just the mean and variance of a normal distribution). These different schemes have enormously different resource implications for a physical implementation. As long as the characterization of representation is left at the level of specifying a general, abstract form, it is difficult to empirically evaluate.

Even what seems to be the most specific claim of the predictive processing framework – that there exist functionally distinct "error" and "representation" units in the brain – is ambiguous. Given multidimensional neuron tuning (Townsend et al. 2006; Tudusciuc & Nieder 2009), units could be simultaneously sensitive to both error and representation, and still perform the relevant computations (Eliasmith & Anderson 2003). This would be compatible with the neurophysiological evidence showing neurons responsive to prediction error, without requiring that there be a sharp division in the brain into these two different sub-populations. Again, the details matter.

One way to begin to fill in the missing details in the predictive processing framework is by being more specific as to what functions are computed. For example, Kalman filters[1] (Kalman 1960) are standard control-theoretic structures that maintain an internal representation of the state of the world, and then use the difference between the predictions of that internal state and incoming data to update the internal model (as the predictive processing framework uses the prediction error signal to update its representations). Clark claims that the predictive processing framework differs from these structures in that it contains a richer error signal (see Note 9 in the target article). However, the Kalman filter is often employed in a multidimensional form (Villalon-Turrubiates et al. 2004; Wu 1985), allowing the error signal to encode rich and complex information about the world. Making use of these parallels provides many potential advantages. For example, Clark describes the need to adjust the relative weight of the model's predictions versus the incoming information, but he does not indicate how that balance is to be achieved. This is a well-studied problem in Kalman filters, where there are specific mechanisms to adjust these weights depending on the measurement or estimate error (Brown & Hwang 1992). Thus, it may be possible to replace the poorly specified notion of "attention" used to control these weights in the predictive processing framework (sect. 2.3) with well-defined mechanisms, providing a more grounded and concrete description.

This is a way of providing computational details to the approach, but we advocate going further – providing implementational details as well. For instance, there is more than one way to implement a Kalman filter in a spiking neural network (Eliasmith & Anderson 2003, Ch. 9), each of which has different implications for the neurophysiological behavior of those networks. Once a neural implementation has been specified, detailed comparisons between computational models and empirical data can be made. More critically, for the grander suggestion that the predictive processing framework is unifying, the implementation of some small set of mechanisms should explain a wide swath of empirical data (see, e.g., Eliasmith et al. [2012] or Eliasmith [in press] for one such attempt).

The ideas presented by Clark are compelling, compatible with empirical data, and attempt to unify several interesting aspects of cognition. However, given the current lack of implementational detail or firm architectural commitments, it is impossible to determine whether the predictive processing framework is largely correct or empirically vacuous. The real test of these ideas will come when they are used to build a model that unifies perception, cognition, and action in a single system. Such an effort will require a deeper investigation of the details, and either fill them in with answers, or if answers are not to be found, require a reworking of the theory. Either way, the predictive processing framework will benefit enormously from the exercise.

NOTE

1. We have in mind here all the varieties of Kalman filters (e.g., extended, unscented, etc.).

Interactively human: Sharing time, constructing materiality

doi:10.1017/S0140525X12002427

Andreas Roepstorff

Interacting Minds Centre, and Centre for Functionally Integrative Neuroscience, Institute of Culture and Society, Aarhus University, DK-8000 Aarhus C, Denmark.

andreas.roepstorff@hum.au.dk

Abstract: Predictive processing models of cognition are promising an elegant way to unite action, perception, and learning. However, in the current formulations, they are species-unspecific and have very little particularly human about them. I propose to examine how, in this framework, humans can be able to massively interact and to build shared worlds that are both material and symbolic.

Andy Clark has written an impressive piece. Predictive processing ideas have been the hype in the neurocognitive community for

some years, for all the reasons that the target article's review identifies. They propose to unify models of perception, action, and learning within a framework – which is elegant, aligned with neuroanatomical and functional findings, computationally plausible, and able to generate empirical research with relatively clear hypotheses.

So far the ideas have been a well-kept secret within the community. This BBS article is likely to change that. As one of the first, Clark brings the predictive processing framework in touch with more general views in cognition and philosophy of mind in a format available to a wider audience. Stripping it of the mathematical formality without losing out on the conceptual stringency, opens for a wider discussion of potential implications for how we think of the brain and of ourselves. Key terms like anticipation, expectancy, models of reality, attention, agency, and surprise appear to move seamlessly between the neuronal, the mathematical, the phenomenological, and the behavioral. The ambition to extend this to a general model of human cognition is impressive, but this is also where the proposal becomes very open-ended. For, ultimately, how human-specific is this predictive framework? In the current formulation, hardly at all. The underlying neural models are basically species-unspecific, and the empirical cases move back and forth between many different model systems. This is not a weakness of the framework; on the contrary, the ambition is to lay out a general theory of brain function, cortical responses, predictive coding, free energy, and so forth. However, it leaves a lot of work open when gauging how this relates to a specific understanding of human action and cognition.

To begin this, one may need to ask what is characteristic of humans as a life form? We don't know for sure, but there are a few candidates. One is an unusual ability for *interaction* – people coordinate, couple, take turns – at many different levels (Levinson 2006). Through interactions, they come to share a structuring of activities in time, and, perhaps, bring brain internal processes in sync too. Another, probably not unrelated, is an amazing ability to *co-construct artefacts* and *build shared worlds* that are at the same time material and symbolic (Clark 2006b; Roepstorff 2008): worlds that exist outside the individual, and in time-windows, which extends beyond the here-and-now of interaction; worlds that, somehow, get internalized. Are these two principles uniquely human? Probably not: Other species also coordinate actions, and other species also modify their surroundings, building niches that are both material and cognitive, but the degree to which people do it is amazing, and we still need to figure out how this can come about, also at a cognitive level.

In sociology and anthropology, one influential attempt to relate interactions and the co-constructed shared worlds has been a focus on human practices (Bourdieu 1977; Roepstorff et al. 2010) as particular unfoldings of temporality set within specific materialities. Translated into predictive coding lingo, these practices may help establish priors or even hyperpriors, sets of expectations that shape perception and guide action (Roepstoff & Frith 2012). Following from this, human priors may not only be driven by statistical properties in the environment, picked up by individual experience, or hardwired into the developing cognitive system. They are also a result of shared expectations that are communicated in interactions, mediated by representations, solidified through materiality, and extended into an action space, going way beyond the physical body and into proximal and distal forms of technology.

This means that both the "predictive" and the "situated" in Clark's title may get a radical twist. It is not so much a matter of living inside a "socio-cultural cocoon," as Clark puts it (sect. 5.2, para. 4). This metaphor suggest that we will at some point grow up and come out of the cocoon into the real world. It is also not just a matter of "man" as "an animal suspended in webs of significance he himself has spun," as Clifford Geertz (1966), following Max Weber, famously suggested. This formulation overemphasizes the symbolic and the individualistic, and it fails to see that the webs "we" have spun are indeed also very material, and that the dimensions of materiality "we" can spin ourselves into seem to be constantly changing. Humans appear to live lives where both priors and possibilities for action – and perhaps also, increasingly, the world – are shaped by actions of others and constrained, stabilised, and afforded by those structures built in the process. But if "being human" in general is about living in unfolded practices, what, then, is it about our cognition that allows us to do that? We don't know. But something about how humans can bridge the material and the symbolic, and something about how they in and through interactions can share both external and internal time, may be critical.

The predictive framework, in "linking action, perception, and learning," is highly relevant also to researchers outside of the neurosciences. But at this stage, there is much to fill in for it to function as a general model of human cognition and action. Certainly, the free energy principle, the predictive hierarchical stuff, the putative links between action, perception, and learning seem to be good candidates for the new "rough guide" to brain function. However, these guiding principles appear to work equally well in rats, in macaques, and in humans. For those of us who are particularly interested in what humans do to themselves, to each other, and to their world, there seem to be a lot of lacunae to be explored, and a lot of gaps to be filled. Getting these right may perhaps also teach something about what humans, as interactive agents, embedded in sociocultural worlds, may do to their brains. Will this throw new light on neuroscience too? Perhaps. There is certainly much work to be done by researchers from many disciplines.

Action-oriented predictive processing and the neuroeconomics of sub-cognitive reward

doi:10.1017/S0140525X12002166

Don Ross

School of Economics, University of Cape Town, Rondebosch 7701, Cape Town, South Africa.

don.ross@uct.ac.za **http://uct.academia.edu/DonRoss**

Abstract: Clark expresses reservations about Friston's reductive interpretation of action-oriented predictive processing (AOPP) models of cognition, but he doesn't link these reservations to specific alternatives. Neuroeconomic models of sub-cognitive reward valuation, which, like AOPP, integrate attention with action based on prediction error, are such an alternative. They interpret reward valuation as an input to neocortical processing instead of reducing it.

Clark impressively surveys the prospects, based on current evidence and speculations tethered to clearly specified models, that action-oriented predictive processing (AOPP) accounts of cortical activity offer the basis for a deeply unified account of perception, cognition, and action. It is indeed clear that such accounts provide, at the very least, a fresh and stimulating framework for explaining the apparently expectation-driven nature of perception. And once one gets this far, it would be a strangely timid modeler who did not see value in exploring the hypothesis that such perception was closely linked to preparation of action and to monitoring of its consequences. However, Clark structures his critical discussion around the most ambitious efforts to use AOPP as the basis for a reductive unification of "all elements of systemic organization" in the brain (sect. 1.6, para. 3), efforts mainly associated with the work of Karl Friston and his co-authors. Clark expresses some reservations about this strong, over-arching hypothesis. My commentary amplifies some of these reservations, based on neglect of the role of specialized subsystems that may integrate valuation, attention, and motor preparation semi-independently of general cortical processing.

Clark's survey is notable for the absence of any discussion of relative reward-value computation. Studies of such valuation based on single-cell recordings in rat striatum were the original locus of models of neural learning as adjustment of synaptic weights and connections through prediction-error correction (Schultz et al. 1997). The temporal difference (TD) learning that has been progressively generalized in descendents of Schultz et al.'s model is a form of Rescorla-Wagner conditioning, not Bayesian equilibration, and so could not plausibly be expected to provide a general account of mammalian cognition. However, neuroeconomists have subsequently embedded TD learning in models of wider scope that exploit drift diffusion and meta-conditioning to track such complex targets as stochastic dominance of strategies in games with shifting mixed-strategy equilibria (Glimcher 2010; Lee & Wang 2009). Such models can effectively approximate Bayesian learning. However, as Clark reports, Friston's most recent work "looks to involve a strong commitment … to the wholesale replacement of value functions, considered as determinants of action, with expectations … about action" (see Note 12 in the target article).

One theorist's elimination is frequently another theorist's construct implementation. Neuroeconomic models of the striatal dopamine circuit do away with the need to posit learned or innate reward value hierarchies that provide targets for the learning of action and the training of attention. Like AOPP theory, such models effectively fuse attentional capture and entrenchment with reward, explaining both as functional products of the prediction error learning encoded by dopamine signals. Extensions of neuroeconomic models to account for pathologies of attention and valuation, such as addiction, have incorporated evidence for direct dopaminergic/striatal signaling to motor preparation areas. For example, Everitt et al. (2001) suggest that direct signals to motor systems to prepare to consume addictive targets when attention is drawn to predictors of their availability are the basis for the visceral cravings that, in turn, cause addictive preoccupation. More basically, Glimcher's (2003) proposal to model some neural response using economics was originally motivated by observations of activity in cells that control eye saccades when monkeys implement incentivized choices through gaze direction (Platt & Glimcher 1999).

This integration of attention and neural learning with action is crucial in the present context, because, like the prediction errors modeled in AOPP, this allows them to "carry information not just about the quantity of error but … about the mismatched content itself," as Clark says (Note 9 of the target article).

So far, we might seem to have only a semantic difference between neuroeconomics and Friston's radical interpretation of AOPP: Neuroeconomists take themselves to be furnishing a theory of neural value functions, while Friston proposes to eliminate them. But this in fact represents substantive divergences, all of which reflect worries that Clark notes but doesn't connect with particular alternative accounts.

First, consider the problem of why, if AOPP is the general account of cognitive dynamics, animals do not just sit still in dark rooms to maintain error-minimizing equilibria. Clark cites Friston's suggestion in response that "some species are equipped with prior expectations that they will engage in exploratory or social play" (Friston 2011a; see sect. 3.2, para. 2, in the target article). However, good biological methodology recommends against positing speculative innate knowledge as inferences to best explanations conditional on one's hypothesis. The neuroeconomic model of striatal valuation makes this posit unnecessary – or, on another philosophical interpretation, replaces the dubious IBE by evidence for a mechanism – by suggesting that discovery of mismatches between expectations and consequences of action is the basis of phasic dopamine release, and such release is the foundation of reward, attention, and further action.

Second, allowing for a relatively encapsulated and cognitively impenetrable pre-frontal mechanism in striatum that integrates attention and action in a way that is partly independent of general cognition, allows us to straightforwardly model the disconnect Clark identifies between surprise to the brain ("surprisal") and surprise to the agent. Clark's example is of a surprise-minimizing perceptual inference that surprises the agent. But disconnects in the other direction are also important. Gambling addiction may result from the fact that the midbrain reward circuit is incapable of learning that there is nothing to learn from repeatedly playing a slot machine, even after the mechanism's victim/owner has become sadly aware of this truth (Ross et al. 2008).

The suggestion here is that neuroeconomics is one resource – of course we should expect there to be others – for addressing Clark's concern that "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 … divides its cognitive labors between multiple cortical and subcortical areas" (sect. 3.3, para. 4). But in that case it seems most natural to join the neuroeconomists in understanding sub-cognitive valuation as an input to cognition, rather than as something that a model of cognitive activity should reduce away.

Affect and non-uniform characteristics of predictive processing in musical behaviour

doi:10.1017/S0140525X12002373

Rebecca S. Schaefer, Katie Overy, and Peter Nelson

Institute for Music in Human and Social Development (IMHSD), Reid School of Music, University of Edinburgh, Edinburgh EH8 9DF, United Kingdom.

r.schaefer@ed.ac.uk k.overy@ed.ac.uk p.nelson@ed.ac.uk

http://www.ed.ac.uk/schools-departments/edinburgh-college-art/music/research/imhsd/imhsd-home

Abstract: The important roles of prediction and prior experience are well established in music research and fit well with Clark's concept of unified perception, cognition, and action arising from hierarchical, bidirectional predictive processing. However, in order to fully account for human musical intelligence, Clark needs to further consider the powerful and variable role of affect in relation to prediction error.

The roles of prediction, expectation, and prior experience in musical processing are well established (Huron 2006; Large et al. 2002; Meyer 1956; Narmour 1990; Phillips-Silver & Trainor 2008; Vuust & Frith 2008), and indeed have led to the proposal that music has the capacity to create an environment of *minimized prediction error* within individuals and within groups (e.g., via a steady pulse) (Overy & Molnar-Szakacs 2009). Bayesian models have been shown to account for a range of phenomena in music perception (Temperley 2007) and have been used to bring together apparently diverging datasets from rhythm perception and production tasks (Sadakata et al. 2006). Moreover, it has been shown that the motor system is engaged during auditory rhythm perception (e.g., Grahn & Brett 2007), and that musical imagery evokes similar neural responses as perception (Schaefer et al. 2011a; 2011b). Clark's unified framework of perception, action, and cognition is thus well supported by recent music research.

However, the current account does not attempt to deal with the range of ways in which prediction error induces arousal and affect. The extent to which our predictions are met or violated, historically theorized to lead to an arousal response (Berlyne 1970), can make a piece of music more or less coherent, interesting, and satisfying. Aesthetically, this leads to the concept of an *optimal* level of surprisal, which (although initially formulated to describe liking or hedonic value for differing levels of musical complexity; e.g., North & Hargreaves 1995) can be described as an inverted U-shaped function in which, on the *x*-axis of

prediction error, there is a preferred level of surprisal that leads to a maximally affective response, plotted on the *y*-axis. However, this optimal surprisal level is not uniform over musical features (e.g., expressive timing, harmonic structure), but rather is closely coupled to the specific characteristics of that musical feature or behaviour. As Clark states, context sensitivity is fundamental, and in the case of music, different levels of constraint will exist simultaneously across different systems of pitch space and time. For example: Singing often has high constraints in terms of pitch, tuning, and scale, while timing constraints may be more flexible; but drumming usually involves strict timing constraints, with more flexibility in terms of pitch. Our perceptual systems are finely attuned to these constraints, to the point that rhythmic deviations that fit with certain aspects of perceived musical structure are less well detected (Repp 1999), and humanly produced deviations from a steady rhythm are preferred over randomly added noise (Hennig et al. 2011).

This tuning of our perceptual system to specific deviations from an internal model is seen not only in performance aspects of music (such as expressive microtiming), but also in compositional aspects found in the score (such as syncopation). Most musical styles require and indeed "play" with levels of surprisal in the temporal domain, from the musical rubato of Romantic piano performance, to the syncopated off-beat rhythms of jazz, to the complex polyrhythms of African percussion. Proficient musicians and composers are implicitly aware of these effects, and tailor their efforts to interact with the surprisal responses of the listener. This leads to what has been coined "communicative pressure" in creating music (Temperley 2004): an implicit knowledge of the musical dimension in which prediction can be manipulated stylistically, without leading to a lack of clarity of the musical ideas. While this complexity corresponds closely to what Clark refers to as a designed environment, it is important to note that different musical environments have different rules, that different listeners (due to their different exposure backgrounds, such as culture and training) seek different environments, and that the desired outcome is a complex affective response. Indeed, exposure has been shown to influence liking for a completely new musical system after only 30 minutes of exposure (Loui et al. 2010). This finding supports the idea of a strong personalized configuration of one's own preference for unpredictability, reflected in musical likes and dislikes, as well as one's own prediction abilities, shown to be quite stable over time per individual, affecting interpersonal coordination (Pecenka & Keller 2011). An individual personality might be thrill-seeking and seek out highly unpredictable new musical experiences, or, more commonly, might seek out highly predictable familiar, favorite musical experiences.

Thus, different kinds of musical experience, different musical styles, and personal musical preferences lead to different predictions, error responses, arousal, and affect responses across a range of musical dimensions and hierarchical levels. The upshot is that the surprisal response is non-uniform for music: The positioning of a curve describing "optimal surprisal" for affective or aesthetic reward will be determined by culture, training, or musical style, and its precise shape (e.g., kurtosis) may be specific to the type and level of the prediction or mental model. And while the characteristics of the optimal surprisal for each aspect of music differs, the commonality remains affect, which, we propose, plays a major part in what makes prediction error in music (large or small) meaningful, and indeed determines its value.

To the extent that prediction is established as a powerful mechanism in conveying musical meaning, it seems clear then that it is the affective response to the prediction error that gives the initial prediction such power. We thus propose that the valence of the prediction error, leading to a range of affective responses, is a necessary component of the description of how predictive processing can explain musical behaviour. The function of such affective predictability will require discussion elsewhere, but we postulate that this will include deep connections with social understanding and communication, from simple group clapping, a uniquely human behaviour requiring constant automatic adjustments of probabilistic representation (Molnar-Szakacs & Overy 2006; Overy & Molnar-Szakacs 2009), to more sophisticated rhythmic organization and self-expression (Nelson 2012) with an emphasis on "error" as positive, meaningful information.

Extending predictive processing to the body: Emotion as interoceptive inference

doi:10.1017/S0140525X12002270

Anil K. Seth[a,b] and Hugo D. Critchley[a,c]

[a]*Sackler Centre for Consciousness Science, University of Sussex, Brighton BN1 9QJ, United Kingdom;* [b]*Department of Informatics, University of Sussex, Brighton BN1 9QJ, United Kingdom;* [c]*Department of Psychiatry, Brighton and Sussex Medical School, Brighton BN1 9QJ, United Kingdom.*

a.k.seth@sussex.ac.uk **H.Critchley@bsms.ac.uk**
www.anilseth.com **www.sussex.ac.uk/sackler/**

Abstract: The Bayesian brain hypothesis provides an attractive unifying framework for perception, cognition, and action. We argue that the framework can also usefully integrate *interoception*, the sense of the internal physiological condition of the body. Our model of "interoceptive predictive coding" entails a new view of emotion as interoceptive inference and may account for a range of psychiatric disorders of selfhood.

In his compelling survey, Clark powerfully motivates predictive processing as a framework for neuroscience by considering the "view from inside the black box," the notion that the brain must discover information about the world without any direct access to its source. The ensuing discussion, and the large majority of the literature surveyed, is focused on just these relations between brain and (external) world. Perhaps underemphasized in this view is the question of how perceptions of the body and self arise. However, the brain's access to the facts of its embodiment and of its physiological milieu is arguably just as indirect as its access to the surrounding world. Here, we extend Clark's integrative analysis by proposing that *interoception* – the sense of the physiological condition of the body (see Craig 2003) – can also be usefully considered from the perspective of predictive processing. Our model of "interoceptive predictive coding" (Critchley & Seth 2012; Seth et al. 2011) suggests a new view of emotional feelings as interoceptive inference, and sheds new light on dissociative disorders of self-consciousness.

Interoceptive concepts of emotion were crystallized by James (1890) and Lange (1885/1912), who argued that emotions arise from perception of changes in the body. This basic idea remains influential more than a century later, underpinning frameworks for understanding emotion and its neural substrates, such as the "somatic marker hypothesis" (Damasio 2000) and the "sentient self" model (Craig 2009), both linked to the notion of "interoceptive awareness" or "interoceptive sensitivity" (Critchley et al. 2004). Despite the neurobiological insights emerging from these frameworks, interoception has remained generally understood along "feedforward" lines, similar to classical feature-detection or evidence-accumulation theories of visual perception as summarized by Clark. However, it has long been recognised that explicit cognitions and beliefs about the causes of physiological changes influence subjective feeling states and emotional behaviour. Fifty years ago, Schachter and Singer (1962) famously demonstrated that injections of adrenaline, proximally causing a state of physiological arousal, would give rise to either anger or elation depending on the concurrent context (an irritated or elated confederate). This observation was formalized in their "two factor" theory, in which emotional experience is determined by the combination

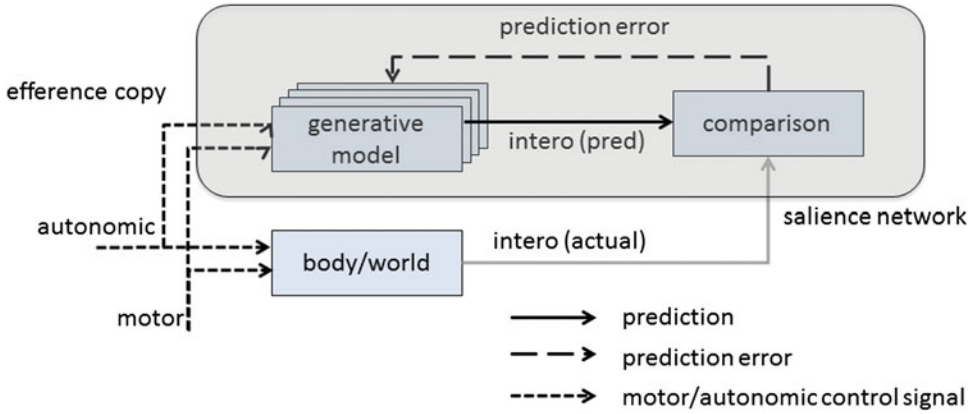

Figure 1 (Seth & Crichley). A model of interoceptive predictive coding according to which subjective feeling states are constituted by continually updated predictions of the causes of interoceptive input. Predictions are shaped by generative models informed by "efference copies" of visceral, autonomic, and motor control signals. These are generated, compared, and updated within a salience network anchored on the anterior insular and anterior cingulate cortices that engage brainstem regions as targets for visceromotor control and relays of afferent interoceptive signals. Adapted from Seth et al. (2011).

of physiological change and cognitive appraisal, that is, emotion as interpreted bodily arousal.

Though they involve expectations, two-factor theories fall considerably short of a full predictive processing model of emotion. By analogy with corresponding models of visual perception, predictive interoception involves hierarchically cascading top-down interoceptive predictions that counterflow with bottom-up interoceptive prediction errors. Subjective feeling states are then determined by the integrated content of these predictive representations across multiple levels (Seth et al. 2011). In other words, the model argues that emotional content is determined by a suite of hierarchically organized generative models that predict interoceptive responses to both external stimuli and the internal signals controlling bodily physiology (Fig. 1).

It is important to distinguish interoceptive predictive coding or processing from more generic interactions between prediction and emotion (e.g., Gilbert & Wilson 2009; Ploghaus et al. 1999). Crucially, predictive coding involves prediction at synchronic, fast time-scales, such that predictions (and prediction errors) are constitutive of content. For example, while Paulus and Stein (2006) hypothesize the existence of interoceptive prediction errors within insular cortex in the generation of anxiety, they do not contend, in the full predictive coding sense, that interoceptive predictions are the constitutive basis of emotions. Similarly, although Barrett and Bar (2009) propose that affective (interoceptive) predictions within orbitofrontal cortex shape visual object recognition at fast time-scales, they again do not describe interoceptive predictive coding per se.

Several strands of evidence lend support to our model and point to its implications for dissociative psychiatric symptoms such as depersonalization and chronic anxiety (Seth et al. 2011). Anterior insular cortex (AIC) in particular provides a natural locus for comparator mechanisms underlying interoceptive predictive coding, through its demonstrated importance for interoceptive representation (Craig, 2009; Critchley et al. 2004) and by the expression within AIC of prediction error signals across a variety of affect-laden contexts (Paulus & Stein 2006; Singer et al. 2009; Palaniyappan & Liddle 2011). Human AIC is also rich in Von Economo neurons (VENs), large projection neurons which are circumstantially associated with self-consciousness and complex social emotions (Craig 2009). In our model, fast VEN-mediated connections may enable the rapid registration of visceromotor and viscerosensory signals needed for efficient updating of generative models underlying interoceptive predictive coding. The recent discovery of VENs in the macaque monkey (Evrard et al. 2012) opens important new avenues for experimental tests of the potential role of VENs in this process and in conscious awareness more generally (Critchley & Seth 2012).

Disrupted interoceptive predictive coding may causally account for a range of psychiatric disorders. Chronic anxiety has been suggested to result from heightened interoceptive prediction error signals (Paulus & Stein 2006). By analogy with comparator models of schizophrenia (Frith 2012; Synofzik et al. 2010), we also suggest that dissociative symptoms, notably depersonalization and derealization arise from imprecise (as opposed to inaccurate) interoceptive prediction error signals. By the same token, the subjective sense of reality characteristic of normal conscious experience (i.e., "conscious presence") may depend on the successful suppression by top-down predictions of informative interoceptive signals (Seth et al. 2011).

In summary, subjective emotions and even conscious presence may be usefully conceptualized in terms of interoceptive predictive coding. A key test of our model will be to identify specific interoceptive prediction error responses in the AIC or elsewhere. This challenge is also yet to be met for predictive processing models of perception in general, and the relevant evidence would go a long way towards experimentally validating the Bayesian brain hypothesis.

Perception versus action: The computations may be the same but the direction of fit differs

doi:10.1017/S0140525X12002397

Nicholas Shea

Department of Philosophy, King's College London, Strand, London WC2R 2LS, United Kingdom.

nicholas.shea@kcl.ac.uk

http://www.kcl.ac.uk/artshums/depts/philosophy/people/staff/academic/shea/index.aspx

Abstract: Although predictive coding may offer a computational principle that unifies perception and action, states with different directions of fit are involved (with indicative and imperative contents, respectively). Predictive states are adjusted to fit the world in the course of perception, but in the case of action, the corresponding states act as a fixed target towards which the agent adjusts the world. This well-recognised distinction helps sidestep some problems discussed in the target article.

One of the central insights motivating Clark's interest in the potential for predictive coding to provide a unifying

computational principle is the finding that it can be the basis of effective algorithms in both the perceptual and motor domains (Eliasmith 2007, p. 380). That is surprising because perceptual inference in natural settings is based on a rich series of sensory inputs at all times, whereas a natural motor control task only specifies a final outcome. Many variations in the trajectory are irrelevant to achieving the final goal (Todorov & Jordan 2002), a redundancy that is absent from the perceptual inference problem. Despite this disanalogy, the two tasks are instances of the same general mathematical problem (Todorov 2006).

Clark emphasises the "deep unity" between the two tasks, which is justified but might serve to obscure an important difference. In the perceptual task, a prediction error is used to change expectations so as to match the input, whereas, as Clark notes, in the motor task the prediction error is used to drive motor behaviour that changes the input. In perception, prediction error is minimised by changing something internal (expectations), whereas in action prediction error is minimised by changing something external (acting on the world so as to alter sensory input). Although it is true in one sense that there is a common computational principle that does not distinguish between perceptual and motor tasks (sect. 1.5), we should not overlook the fact that those computations are deployed quite differently in the two cases. In the two cases state representations have what philosophers have called different "directions of fit." A motor task takes as input a goal state, which is held fixed; a motor program to attain that goal state is then calculated (Todorov 2004). These goal states have a world-to-mind direction of fit and imperative content. By contrast, the state descriptions in the perceptual task (expectations fed back from higher levels in the processing hierarchy) are continually adjusted so as to match the current sensory input more closely. They display a world-to-mind direction of fit and have indicative content. The difference is apparent in its consequences for the behaviour of the organism: Prediction errors in respect of indicative representations can be fully cancelled without the agent having to perform any action, whereas prediction errors in respect of imperative representations cannot be cancelled unless the agent moves in some way.

If these accounts are right, then the deep unity consists in the fact that both perception and action involve the reduction of prediction error. However, since they do so by quite different means, a deep difference between perception and action remains. Some sensorimotor accounts of our interactions with the world do indeed serve to dissolve the boundary between perception and action (Hurley 1998), but the predictive coding framework on its own does not. (It does, however, undermine a clear boundary between perception and cognition.) This gives rise to an important question for the predictive coding programme: What determines whether a given prediction/expectation is given a mind-to-world functional role, allowing it to be adjusted in the light of prediction errors, and what gives other expectations a world-to-mind functional role, such that prediction errors cause bodily movements/action? As the evidence for a common computational principle in perception and action mounts, the need becomes pressing to specify how this fundamental difference between its two modes of operation arises.

Clark goes on to consider whether an austere "desert landscape" description of the computational processing is possible that does away with goals and reward entirely (sect. 5.1), in the sense that neither are represented in the model. If action guidance requires states with a world-to-mind direction of fit, then states which function as goals have not been eliminated. Even if the difference is a matter of degree, with many cases in the middle, we are still operating with a continuum marked by the extent to which a state operates as a goal state at one end or as an indicative state at the other.

The distinction between indicative and imperative contents also throws light on the darkened room problem: Why don't agents minimise prediction error by just sitting still in a darkened room? If some subsystems are constrained to minimise prediction error not by changing expectations but by acting, then sitting still in a darkened room will be entirely ineffective in reducing such error signals. For example, if there is one of these goal state representations for the level of sugar in the blood, when sensory feedback fails to match the target the agent does not have the option of reducing the error signal by changing its expectation; instead, the agent must act so as to change the sensory feedback (i.e., to increase the level of sugar in the blood). This answer is complementary to Clark's observation that some forms of prior expectation could lead agents to engage in exploratory actions or social play. It is orthogonal to the distinction between exploratory and exploitative actions (which can, in any event, only be drawn relative to some set of goal states).

A final observation concerns the question of whether the expectations involved in predictive coding calculations refer to the external world. It is sometimes suggested that predictions and prediction errors only concern the states of other computational elements in the system. Goal states are perhaps the most obvious candidate for representations that refer to the external world. Since the feedback to which they are compared is changed by action on the world, it is plausible that they come to represent the external world affairs that must be changed if the prediction error is to be cancelled.

To conclude, Clark's persuasive case for the importance of predictive coding as a unifying computational principle, like any fruitful research agenda, brings new issues into focus. An important one is the question of what makes that computational principle operate in indicative (perceptual) mode in some subsystems and in imperative (action) mode in others.

Schizophrenia-related phenomena that challenge prediction error as the basis of cognitive functioning

doi:10.1017/S0140525X12002221

Steven M. Silverstein

University Behavioral HealthCare and Department of Psychiatry, Robert Wood Johnson Medical School, University of Medicine and Dentistry of New Jersey, Piscataway, NJ 08854.

silvers1@umdnj.edu

Abstract: There are aspects of schizophrenia that pose challenges for Clark's model. These include: (1) evidence for excitatory activity underlying self-organizing neural ensembles that support coordinating functions, and their impairment in schizophrenia; (2) evidence regarding hallucinations that suggest they are not due to excessive prediction error; and (3) the critical role of emotional factors as setting conditions for delusion formation.

Clark's model emphasizes the processing of prediction error, and, in section 4.2, this is applied to an understanding of hallucinations, delusions, and schizophrenia. This commentary emphasizes three points related to these themes, with the overall goal of demonstrating that Clark's view, at present, does not provide a fully adequate heuristic for understanding psychotic phenomena.

Clark's theory emphasizes anti-Hebbian feedforward processing, in which correlated activity across neurons is suppressed, presumably because no deviation from what is expected is present, therefore allowing any signals related to deviation from what is expected (i.e., prediction error) to become relatively more salient. While this would appear to be a useful data-compression strategy for coding invariant background information, it does not account for cases in which it is precisely the correlation between stimulus elements that codes their object properties, thereby signaling stimulus significance. Numerous demonstrations exist (e.g., Kinoshita et al. 2009; Silverstein et al. 2009; Singer 1995) wherein increasing the correlation between an aspect of elements (e.g., stimulus orientation in contour

integration paradigms) leads to increased signal strength. Of course, it is possible to argue, as Clark does, that this is due to a cancellation of the activity in error units and subsequent enhancement of the signal coding the contour or shape. However, it is not clear how these competing hypotheses could be pitted against each other in a definitive study.

Consistent with Clark's view, evidence exists that, for example, as random orientational jitter is applied to disconnected contour elements, increases in fMRI BOLD signal are observed (Silverstein et al. 2009). Clark's view is also consistent with Weber's (2002) view that much of our direct understanding of visual forms results from perception of "metamorphoses of geometry" or topological (isotopic) alterations of basic forms, a view consistent with evidence that topological invariants are the primitives to which our visual system responds most strongly (Chen 2005). However, it is also the case that compared to a non-informative background of randomly oriented Gabors, perception of a contour is associated with increased activity (Silverstein et al. 2009). Clarifying the extent to which these two forms of signal increase represent functioning of different circuits is an important task for future research. Until this is clarified, Clark's view appears to be most appropriate for understanding signaling of objects in the environment, as opposed to brain activity involved in creating representations of those objects. This is relevant for schizophrenia, as it is characterized by a breakdown in coordinating processes in perception and cognition (Phillips & Silverstein 2003; Silverstein & Keane 2011). A challenge for Clark's view is to account for these phenomena, which have been previously understood as reflecting a breakdown in Hebbian processing, and reduced self-organization at the local circuit level, involving reduced lateral (and re-entrant) excitation.

Clark notes that while perceptual anomalies alone will not typically lead to delusions, the perceptual and doxastic components should not be seen as independent. However, there are several syndromes (e.g., Charles Bonnet Syndrome, Dementia with Lewy Bodies, Parkinson's Disease Dementia) where visual hallucinations are prominent and delusions are typically absent (Santhouse et al. 2000). Moreover, it would appear to be difficult to explain the well-formed hallucinations characteristic of these syndromes as being due to prediction error, given their sometimes improbable content (e.g., very small people dressed in Victorian era attire), and apparent errors in size constancy (ffytche & Howard 1999; Geldmacher 2003) that argue against Bayes-optimal perception in these cases. There are also many cases of schizophrenia where delusions are present without hallucinations. Finally, while evidence of reduced binocular depth inversion illusions in schizophrenia (Keane et al., in press; Koethe et al. 2009) provides evidence, on the one hand, for a weakened influence of priors (or of the likelihood function) (Phillips 2012) on perception, this evidence also indicates *more* veridical perception of the environment. Therefore, these data suggest that, rather than prediction error signals being falsely generated and highly weighted (as Clark suggests), such signals appear not to be generated to a sufficient degree, resulting in a lack of top-down modulation, and bottom-up (but not error) signals being strengthened. Indeed, this is exactly what was demonstrated in recent studies using dynamic causal modeling of ERP and fMRI data from a hollow-mask perception task in people with schizophrenia (Dima et al. 2009; 2010). A developing impairment such as this would lead to subjective changes in the meaning of objects and the environment as a whole, and of the self – which, in turn, can spawn delusions (Mattusek 1987; Sass 1992; Uhlhaas & Mishara 2007), even though the delusional thoughts are unrelated to the likelihood functions and beliefs that existed prior to the onset of the delusion.

Finally, Clark's view of hallucinations is similar to many models of schizophrenia, in that it is based on computational considerations only. But, as noted, delusions often grow out of phenomenological changes and emotional reactions to these (see also Conrad 1958), and this cascade is typically ignored in computational models. It also must be noted that the delusions that patients develop are not about random events, but typically are framed in reference to the self, with appreciation of the statistical structure of the rest of the world being intact. Similarly, auditory hallucinations often involve negative comments about the self, and it has been suggested, due to the high prevalence of histories of childhood physical and sexual abuse in people with schizophrenia (Read et al. 2005), that voices are aspects of memory traces associated with the abuse experience that have been separated from other aspects of the memory trace due to hippocampal impairment secondary to chronic cortisol production (Read et al. 2001) (as opposed to being due to top-down expectancy driven processing). A purely computational theory of hallucinations and/or delusions is like a mathematical theory of music – it can explain aspects of it, but not why one piece of music creates a strong emotional response in one person yet not in another. Psychotic symptom formation must be understood within the context of personal vulnerability and emotional factors, and these are not well accounted for by a Bayesian view at present.

What else can brains do?

doi:10.1017/S0140525X12002439

Aaron Sloman

School of Computer Science, University of Birmingham, Birmingham B15 2TT United Kingdom.

a.sloman@cs.bham.ac.uk **http://www.cs.bham.ac.uk/~axs**

Abstract: The approach Clark labels "action-oriented predictive processing" treats all cognition as part of a system of on-line control. This ignores other important aspects of animal, human, and robot intelligence. He contrasts it with an alleged "mainstream" approach that also ignores the depth and variety of AI/Robotic research. I don't think the theory presented is worth taking seriously as a *complete* model, even if there is much that it explains.

Clark's paper deserves far more than 1,000 words, but I have to be brief and dogmatic. Characterizing brains as predicting machines ignores many abilities produced by evolution and development,[1] including mathematical discovery and reasoning, using evolved mechanisms (perhaps) shared by several species capable of the "representational redescription" postulated in Karmiloff-Smith (1992) and the meta-configured competences suggested in Chappell & Sloman (2007), including (largely unstudied) discoveries of "toddler theorems" (Sloman 2010). The "action-oriented predictive processing" approach treats everything as on-line control (Powers 1973), like "enactivist" theorists who usually ignore competences required to make predictions true and processes generating and choosing (sometimes unconsciously) between goals, plans, designs (for houses, machines, etc.), preferences, explanations, theories, arguments, story plots, forms of representation, ontologies, grammars, and proofs. Predictive processing doesn't explain termite cathedral building. (Compare Chittka & Skorupski 2011).

Simultaneous localisation and mapping (SLAM) robotic techniques, partly inspired by things animals do, create useful (topological, metrical, and possibly logical) representations of enduring extended environments. That's not learning about mappings between inputs and outputs. It's a special case of using actions, percepts, and implicit theories to derive useful information about the environment. Another is producing a theory of chemical valency.

Systematically varying how things are squeezed, stroked, sucked, lifted, rotated, and so forth, supports learning about kinds of matter, and different spatial configurations and processes involving matter (Gibson 1966). Predicting sensory signals is only one application. Others include creating future

structures and processes in the environment, and understanding processes. Choosing future actions often ignores sensory and motor details, since a different ontology is used (e.g., choosing between a holiday spent practising French and a music-making holiday, or choosing insulation for a new house). For more on "off-line" aspects of intelligence ignored by many "enactivist" and "embodied cognition" enthusiasts, see Sloman (1996; 2006; 2009). Even for on-line control, the use of servo-control with qualitative modifications of behavior responding to changing percepts reduces the need for probabilistic prediction: Head for the center of the gap, then as you get close use vision or touch to control your heading. Choosing a heading may, but need not, involve prediction: it could be a reflex action.

Predicting environmental changes need not use Bayesian inference, for example when you predict that two more chairs will ensure seats for everyone, or that the gear wheel rotating clockwise will make the one meshed with it rotate counter-clockwise. And some predictions refer to what cannot be sensed, for example most deep scientific predictions, or a prediction that a particular way of trying to prove Fermat's last theorem will fail.

Many things humans use brains for do not involve on-line intelligence, for example mulling over a conversation you had a week ago, lying supine with eyes shut composing a piano piece, trying to understand the flaw in a philosophical argument, or just day-dreaming about an inter-planetary journey.

I don't deny that many cognitive processes involve mixtures of top-down, bottom-up, middle-out (etc.) influence: I helped produce a simple model of such visual processing decades ago, Popeye (Sloman 1978, Ch. 9), and criticized over-simple theories of vision that ignored requirements for process perception and on-line control (Sloman 1982; 1989). David Hogg, then my student, used 3-D prediction to reduce visual search in tracking a human walker (Hogg 1983). Sloman (2008) suggests that rapid perception of complex visual scenes requires rapid activation and instantiation of many normally dormant, previously learnt model fragment types and relationships, using constraint propagation to rapidly assemble and instantiate multi-layered percepts of structures and processes: a process of *interpretation*, not *prediction* (compare parsing). Building working models to test the ideas will be difficult, but not impossible. Constraint propagation need not use Bayesian inference.

"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 it 'knows' about, in any direct sense, are the ways its own states (e.g., spike trains) flow and alter….The brain is one such black box" (sect. 1.2). This sounds like a variant of concept empiricism, defeated long ago by Kant (1781) and buried by philosophers of science.

Many things brains and minds do, including constructing interpretations and extending their own meta-cognitive mechanisms, are not concerned merely with predicting and controlling sensory and motor signals.

Evolutionary "trails", from very simple to much more complex systems, may provide clues for a deep theory of animal cognition explaining the many layers of mechanism in more complex organisms. We need to distinguish diverse *requirements* for information processing of various sorts, and also the different *behaviors* and *mechanisms*. A notable contribution is Karmiloff-Smith (1992). Other relevant work includes McCarthy (2008) and Trehub (1991), and research by biologists on the diversity of cognition, even in very simple organisms. I have been trying to do this this sort of exploration of "design space" and "niche space" for many years (Sloman 1971; 1978; 1979; 1987; 1993; 1996; 2002; 2011a; 2011b).

Where no intermediate evolutionary steps have been found, it may be possible to learn from alternative designs on branches derived from those missing cases. We can adopt the designer stance (McCarthy 2008) to speculate about testable mechanisms. (It is a mistake to disparage "just so" stories based on deep experience of struggling to build working systems, when used to guide research rather than replace it.) This project requires studying many types of environment, including not only environments with increasingly complex and varied *physical* challenges and opportunities, but also increasingly rich and varied interactions with *other information processing systems:* predators, prey, and conspecifics (young and old). Generalizing Turing (1952), I call this the "Meta-morphogenesis project" (Sloman 2013).

Clark compares the prediction "story" with "mainstream computational accounts that posit a cascade of increasingly complex feature detection (perhaps with some top-down biasing)" (sect. 5.1). This fits some AI research, but labelling it as "mainstream" and treating it as the only alternative, ignores the diversity of approaches and techniques including constraint-processing, SLAM, theorem proving, planning, case-based reasoning, natural language processing, and many more. Much human motivation, especially in young children, seems to be concerned with extensions of competences, as opposed to predicting and acting, and similar learning by exploration and experiment is being investigated in robotics.

A minor point: Binocular rivalry doesn't always lead to alternating percepts. For example look at an object with one eye, with something moving slowly up and down blocking the view from the other eye. The remote object can appear as if behind a textured window moving up and down.

Clark claims (in his abstract) that the "hierarchical prediction machine" approach "offers the best clue yet to the shape of a unified science of mind and action". But it unifies only the phenomena its proponents attend to.

NOTE

1. For more details, see http://www.cs.bham.ac.uk/research/projects/cogaff/12.html#1203.

Distinguishing theory from implementation in predictive coding accounts of brain function

doi:10.1017/S0140525X12002178

Michael W. Spratling

Department of Informatics, King's College London, University of London, London WC2R 2LS, United Kingdom.
michael.spratling@kcl.ac.uk

Abstract: It is often helpful to distinguish between a theory (Marr's computational level) and a specific implementation of that theory (Marr's physical level). However, in the target article, a single implementation of predictive coding is presented as if this were the theory of predictive coding itself. Other implementations of predictive coding have been formulated which can explain additional neurobiological phenomena.

Predictive coding (PC) is typically implemented using a hierarchy of neural populations, alternating between populations of error-detecting neurons and populations of prediction neurons. In the standard implementation of PC (Friston 2005; Rao & Ballard 1999), each population of prediction neurons sends excitatory connections forward to the subsequent population of error-detecting neurons, and also sends inhibitory connections backwards to the preceding population of error-detecting neurons. Similarly, each population of error-detecting neurons also sends information in both directions; via excitatory connection to the following population of prediction neurons, and via inhibitory connections to the preceding population of prediction neurons. (See, for example, Figure 2 in Friston [2005], or Figure 2b in Spratling [2008b]). It is therefore inaccurate for Clark to state (see sects. 1.1 and 2.1) that in PC the feedforward flow of information solely conveys prediction error, while feedback only

conveys predictions. Presumably what Clark really means to say is that the standard implementation of PC proposes that *inter-regional* feedforward connections carry error, whereas *inter-regional* feedback connections carry predictions (while information flow in the reverse directions takes place within each cortical area). However, this is simply one hypothesis about how PC should be implemented in cortical circuitry. It is also possible to group neural populations differently so that inter-regional feedforward connections carry predictions, not errors (Spratling 2008b).

As alternative implementations of the same computational theory, these two ways of grouping neural populations are compatible with the same psychophysical, brain imaging, and neurophysiological data reviewed in section 3.1 of the target article. However, they do suggest that different cortical circuitry may underlie these outward behaviours. This means that claims (repeated by Clark in sect. 2.1) that prediction neurons correspond to pyramidal cells in the deep layers of the cortex, while error-detecting neurons correspond to pyramidal cells in superficial cortical layers, are not predictions of PC in general, but predictions of one specific implementation of PC. These claims, therefore, do not constitute falsifiable predictions of PC (if they did then the idea that PC operates in the retina – as discussed in sect. 1.3 – could be rejected, due to the lack of cortical pyramidal cells in retinal circuitry!). Indeed, it is highly doubtful that these claims even constitute falsifiable predictions of the standard implementation of PC. The standard implementation is defined at a level of abstraction above that of cortical biophysics: it contains many biologically implausible features, like neurons that can generate both positive and negative firing rates. The mapping between elements of the standard implementation of PC and elements of cortical circuitry may, therefore, be far less direct than is suggested by the claim about deep and superficial layer pyramidal cells. For example, the role of prediction neurons and/or error-detecting neurons in the model might be performed by more complex cortical circuitry made up of diverse populations of neurons, none of which behave like the model neurons but whose combined action results in the same computation being performed.

The fact that PC is typically implemented at a level of abstraction that is intermediate between that of low-level, biophysical, circuits and that of high-level, psychological, behaviours is a virtue. Such intermediate-level models can identify common computational principles that operate across different structures of the nervous system and across different species (Carandini 2012; Phillips & Singer 1997); they seek integrative explanations that are consistent between levels of description (Bechtel 2006; Mareschal et al. 2007), and they provide *functional* explanations of the empirical data that are arguably the most relevant to neuroscience (Carandini et al. 2005; Olshausen & Field 2005). For PC, the pursuit of consistency across levels may prove to be a particularly important contribution to the modelling of Bayesian inference. Bayes' theorem states that the posterior is proportional to the product of the likelihood and the prior. However, it places no constraints on how these probabilities are calculated. Hence, any model that involves multiplying two numbers together, where those numbers can be plausibly claimed to represent the likelihood and posterior, can be passed off as a Bayesian model. This has led to numerous computational models which lay claim to probabilistic respectability while employing mechanisms to derive "probabilities" that are as ad-hoc and unprincipled as the non-Bayesian models they claim superiority over. It can be hoped that PC will provide a framework with sufficient constraints to allow principled models of hierarchical Bayesian inference to be derived.

A final point about different implementations is that they are not necessarily all equal. As well as implementing the PC theory using different ways of grouping neural populations, we can also implement the theory using different mathematical operations. Compared to the standard implementation of PC, one alternative implementation (PC/BC) is mathematically simpler while explaining more of the neurophysiological data: Compare the range of V1 response properties accounted for by PC/BC (Spratling 2010; 2011; 2012a; 2012b) with that simulated by the standard implementation of PC (Rao & Ballard 1999); or the range of attentional data accounted for by the PC/BC implementation (Spratling 2008a) compared to the standard implementation (Feldman & Friston 2010). Compared to the standard implementation, PC/BC is also more biologically plausible; for example, it does not employ negative firing rates. However, PC/BC is still defined at an intermediate-level of abstraction, and therefore, like the standard implementation, provides integrative and functional explanations of empirical data (Spratling 2011). It can also be interpreted as a form of hierarchical Bayesian inference (Lochmann & Deneve 2011). However, it goes beyond the standard implementation of PC by identifying computational principles that are shared with algorithms used in machine learning, such as generative models, matrix factorization methods, and deep learning architectures (Spratling 2012b), as well as linking to alternative theories of brain function, such as divisive normalisation and biased competition (Spratling 2008a; 2008b). Other implementations of PC may in future prove to be even better models of brain function, which is even more reason not to confuse one particular implementation of a theory with the theory itself.

Sparse coding and challenges for Bayesian models of the brain

doi:10.1017/S0140525X12002300

Thomas Trappenberg and Paul Hollensen

Faculty of Computer Science, Dalhousie University, Halifax, NS B3H 4R2, Canada.

tt@cs.dal.ca **paulhollensen@gmail.com**

www.cs.dal.ca/~tt

Abstract: While the target article provides a glowing account for the excitement in the field, we stress that hierarchical predictive learning in the brain requires sparseness of the representation. We also question the relation between Bayesian cognitive processes and hierarchical generative models as discussed by the target article.

Clark's target article captures well our excitement about predictive coding and the ability of humans to include uncertainty in making cognitive decisions. One additional factor for representational learning to match biological findings that has not been stressed much in the target article is the importance of sparseness constraints. We discuss this here, together with some critical remarks on Bayesian models and some remaining challenges quantifying the general approach.

There are many unsupervised generative models that can be used to learn representations to reconstruct input data. Consider, for example, photographs of natural images. A common method for dimensionality reduction is principle component analysis that represents data along orthogonal feature vectors of decreasing variance. However, as nicely pointed out by Olshausen and Field (1996), the corresponding filters do not resemble receptive fields in the brain. In contrast, if a generative model has the additional constraint to minimize not only the reconstruction error but also the number of basis functions that are used for any specific image, then filters emerge that resemble receptive fields of simple cells in the primary visual cortex.

Sparse representation in the neuroscientific context actually has a long and important history. Horace Barlow pointed out for years that the visual system seems to be remarkably set up for sparse representations (Barlow 1961), and probably the first systematic model in this direction was proposed by his student Peter

Földiák (1990). It seems that nearly every generative model with a sparseness constraint can reproduce receptive fields resembling simple cells (Saxe et al. 2011), and Ng and colleagues have shown that sparse hierarchical Restricted Boltzmann Machines (RBMs) resembles features of receptive fields in V1 and V2 (Lee et al. 2008). In our own work, we have shown how lateral inhibition can implement sparseness constrains in a biological way while also promoting topographic representations (Hollensen & Trappenberg 2011).

Sparse representation has great advantages. By definition, it means that only a small number of cells have to be active to reproduce inputs in great detail. This not only has advantages energetically, it also represents a large compression of the data. Of course, the extreme case of maximal sparseness corresponding to grandmother cells is not desirable, as this would hinder any generalization ability of a model. Experimental evidence of sparse coding has been found in V1 (Vinje & Gallant 2000) and hippocampus (Waydo et al. 2006).

The relation of the efficient coding principle to free energy is discussed by Friston (2010), who provides a derivation of free energy as the difference between complexity and accuracy. That is, minimizing free energy maximizes the probability of the data (accuracy), while also minimizing the difference (cross-entropy) between the causes we infer from the data and our prior on causes. The fact that the latter is termed *complexity* reflects our intuition that causes in the world lie in a smaller space than their sensory projections. Thus, our internal representation should mirror the sparse structure of the world.

While Friston shows the equivalence of Infomax and free energy minimization *given* a sparse prior, a fully Bayesian implementation would treat the prior itself as a random variable to be optimized through learning. Indeed, Friston goes on to say that the criticism of where these priors come from "dissolves with hierarchical generative models, in which the priors themselves are optimized" (Friston 2010, p. 129). This is precisely what has not yet been achieved: a model which learns a sparse representation of sensory messages due to the world's sparseness, rather than due to its architecture or static priors. Of course, we are likely endowed with a range of priors built-in to our evolved cortical architecture in order to bootstrap or guide development. What these native priors are and the form they take is an interesting and open question.

There are two alternatives to innate priors for explaining the receptive fields we observe. First, there has been a strong tendency to learn hierarchical models layer-by-layer, with each layer learning to reconstruct the output of the previous without being influenced by top-down expectations. Such top-down modulation is the prime candidate for expressing empirical priors and influencing learning to incorporate high-level tendencies. Implementing a model that balances conforming to both its input and top-down expectations while offering efficient inference and robustness is a largely open question (Jaeger 2011). Second, the data typically used to train our models on differs substantially from what we are exposed to. The visual cortex experiences a stream of images with substantial temporal coherence and correlation with internal signals such as eye movements, limiting the conclusions we can draw from comparing its representation to models trained on static images (see, e.g., Rust et al. 2005).

The final comment we would like to make here concerns the discussion of Bayesian processes. Bayesian models such as the ideal observer have received considerable attention in neuroscience since they seem to nicely capture human abilities to combine new evidence with prior knowledge in the "correct" probabilistic sense. However, it is important to realize that these Bayesian models are very specific to limited experimental tasks, often with only a few possible relevant states, and such models do not generalize well to changing experimental conditions. In contrast, the Bayesian model of a Boltzmann machine represents general mechanistic implementations of information processing in the brain that we believe can implement a general learning machine. While all these models are Bayesian in the sense that they represent causal models with probabilistic nodes, the nature of the models are very different. It is fascinating to think about how such specific Bayesian models as the ideal observer can emerge from general learning machines such as the RBM. Indeed, such a demonstration would be necessary to underpin the story that hierarchical generative models support the Bayesian cognitive processing as discussed in the target article.

Authors' Response

Are we predictive engines? Perils, prospects, and the puzzle of the porous perceiver

doi:10.1017/S0140525X12002440

Andy Clark

School of Philosophy, Psychology, and Language Sciences, University of Edinburgh, Edinburgh EH12 5AY, Scotland, United Kingdom.

andy.clark@ed.ac.uk

http://www.philosophy.ed.ac.uk/people/full-academic/andy-clark.html

Abstract: The target article sketched and explored a mechanism (action-oriented predictive processing) most plausibly associated with core forms of cortical processing. In assessing the attractions and pitfalls of the proposal we should keep that element distinct from larger, though interlocking, issues concerning the nature of adaptive organization in general.

R1. Introduction: Combining challenge and delight

The target article ("Whatever next? Predictive brains, situated agents, and the future of cognitive science" – henceforth WN for short) drew a large and varied set of responses from commentators. This has been a source of both challenge and delight. Challenge, because the variety and depth of the commentaries really demands (at least) a book-length reply, not to mention far more expertise than I possess. Delight, because the wonderfully constructive and expansive nature of those responses already paints a far richer picture of both the perils and the prospects of the emerging approach to cortical computation that I dubbed "action-oriented predictive processing" (henceforth PP for short). In what follows I respond, at least in outline, to three main types of challenge (the "perils" referred to in the title) that the commentaries have raised. I then offer some remarks on the many exciting suggestions concerning complementary perspectives and further applications (the prospects). I end by addressing a kind of conceptual puzzle (I call it "the puzzle of the porous perceiver") that surfaced in different ways and that helps focus some fundamental questions concerning the nature (and plausibility) of the implied relation between thought, agent, and world.

R2. Perils of prediction

The key perils highlighted by the commentaries concern (1) the proper "pitch" of the target proposal (is it about

implementation, algorithm, or something more abstract?); (2) the relation between PP and various other strategies and mechanisms plausibly implicated in human cognitive success; and (3) the nature and adequacy of the treatment of attention as a mechanism for "precision-weighting" prediction error.

R2.1 Questioning pitch

Rasmussen & Eliasmith raise some important worries concerning content and pitch. They agree with the target article on the importance and potency of action-oriented predictive processing (PP), and describe the ideas as compelling, compatible with the empirical data, and potentially unifying as well. But the compatibility, they fear, comes at a price. For, the architectural commitments of PP as I defined it are, they argue, too skimpy as yet to deliver a testable model unifying perception, action, and cognition. I agree. Indeed (as they themselves note) much of the target article argues that PP does not serve to specify the detailed form of a cognitive architecture. I cannot agree with them, however, that the commitments PP does make therefore run the risk of being "empirically vacuous." Those commitments include the top-down use of a hierarchical probabilistic generative model for both perception and action, the presence of functionally distinct neural populations coding for representation (prediction) and for prediction-error, and the suggestion that predictions flow backwards through the neural hierarchy while only information concerning prediction error flows forwards. The first of these (the widespread, top-down use of probabilistic generative models for perception and action) constitutes a very substantial, but admittedly quite abstract, proposal: namely, that perception and (by a clever variant – see WN, sect. 1.5) action both depend upon a form of "analysis by synthesis" in which observed sensory data is explained by finding the set of hidden causes that are the best candidates for having generated that sensory data in the first place.

Mechanistically, PP depicts the top-down use of (hierarchical) probabilistic generative models as the fundamental form of cortical processing, accommodating central cases of both perception and action, and makes a further suggestion concerning the way this is achieved. That suggestion brings on board the data compression strategy known as "predictive coding" (WN, sect. 1.1) from which it inherits – or so I argued, but see below – a distinctive image of the flow of information: one in which predictions (from the generative model) flow downwards (between regions of the neural hierarchy) and only deviations from what is predicted (in the form of residual errors) flow forwards between such regions. The general form of this proposal (as **Bridgeman** properly stresses) is not new. It has a long history in mainstream work in neuroscience and psychology that depicts cortex as coding not for properties of the stimulus but for the differences (hence the "news") between the incoming signal and the expected signal.

PP goes further, however, by positing a specific processing regime that seems to require functionally distinct encodings for prediction and prediction error. **Spratling** notes, helpfully, that the two key elements of this complex (the use of a hierarchical probabilistic generative model, and the predictive coding data compression device) constitute what he describes as an "intermediate-level model": one that still leaves unspecified a great many important details concerning implementation. Unlike **Rasmussen & Eliasmith,** however, Spratling notes that: "Such intermediate-level models can identify common computational principles that operate across different structures of the nervous system … and they provide *functional* explanations of the empirical data that are arguably the most relevant to neuroscience" (emphasis Spratling's). WN aimed to present *just such an intermediate-level model*. In so doing, it necessarily fell short of providing a detailed architectural specification of the kind Rasmussen & Eliasmith seek. It does, however, aim to pick out a space of models that share some deep assumptions: assumptions that already have (or so I argued – see WN, sect. 2) many distinctive conceptual and empirical consequences.

Spratling then worries (in a kind of inversion of the doubts raised by **Rasmussen & Eliasmith**) that in one respect, at least, the presentation in WN is rather *too specific*, too close to one possible (but not compulsory) implementation. The issue here concerns the depiction of error as flowing forwards (i.e., between regions in the hierarchy) and predictions as flowing backwards. WN depicts this as a direct consequence of the predictive coding compression technique. But it is better seen, Spratling convincingly argues, as a feature of one (albeit, as he himself accepts, the standard) implementation of predictive coding. Spratling is right to insist upon the distinction between theory and implementation. It is only by considering the space of alternative implementations that we can start to ask truly pointed experimental questions, (of the kind highlighted by Rasmussen & Eliasmith) of the brain: questions that may one day favour one implementation of the key principles, or even none at all. One problem, I suspect, will be that resolving the "what actually flows forward?" issue looks crucial to adjudicating between various close alternatives. But that depends (as Spratling's work shows) upon how we carved the system into levels in the first place, since that determines what counts as flow within a level versus flow between levels. This is not going to be as easy as it sounds, since it is not gross cortical layers but something much more functional (cortical columns, something else?) that is at issue. Experimenters and theorists will thus need to work together to build detailed, testable models whose assumptions (especially concerning what counts as a region or level) are agreed in advance.

Egner & Summerfield describe a number of empirical studies that support the existence both of (visual) surprise signals and of the hierarchical interplay between expectation and surprise. Some of this evidence (e.g., the work by Egner et al. 2010 and by Murray et al. 2002) is discussed in the text, but new evidence (see, e.g., Wyart et al. 2011) continues to emerge. In their commentary Egner & Summerfield stress, however, that complex questions remain concerning the origins of such surprise. Is it locally computed or due to predictions issuing from elsewhere in the brain? My own guess is that both kinds of computation occur, and that complex routing strategies (see Phillips et al. 2010 and essays in von der Marlsberg et al. 2010) determine, on a moment-to-moment basis, the bodies of knowledge and evidence relative to which salient (i.e., precise, highly weighted) prediction error is calculated. It is even possible that these routing effects are themselves

driven by prediction errors of various kinds, perhaps in the manner sketched by den Ouden et al. (2010). Egner & Summerfield go on to note (see WN, sect. 3.1) the continued absence of firm cellular-level evidence for the existence of functionally distinct neural encodings for expectation and surprise. More positively, they highlight some recent studies (Eliades & Wang 2008; Keller et al. 2012; Meyer & Olson 2011) that offer tantalizing hints of such evidence. Especially striking here is the work by Keller et al. (2012) offering early evidence for the existence of prediction-error neurons in supra-granular layers 2/3, which fits nicely with the classic proposals (implicating superficial pyramidal cells) by Friston (2005), Mumford (1992), and others. Such work represents some early steps along the long and complex journey that cognitive science must undertake if it is deliver evidence of the kind demanded by **Rasmussen & Eliasmith.**

Muckli, Petro, & Smith (Muckli et al.) continue this positive trend, describing a range of intriguing and important experimental results that address PP at both abstract and more concrete levels of description. At the abstract level, they present ongoing experiments that aim to isolate the contributions of cortical feedback (downward-flowing prediction) from other processing effects. Such experiments lend considerable support to the most basic tenets of the PP model. Moving on to the more concrete level they suggest, however, that the standard implementation of predictive coding may not do justice to the full swathe of emerging empirical data, some of which (Kok et al. 2012) shows both sharpening of some elements of the neuronal signal, as well as the kind of dampening mandated by successful "explaining away" of sensory data. However, as mentioned in WN sect. 2.1 (see also comments on **Bowman, Filetti, Wyble, & Olivers** [**Bowman et al.**] below), this combination is actually fully compatible with the standard model (see, e.g., remarks in Friston 2005), since explaining away releases intra-level inhibition, resulting in the correlative sharpening of some parts of the activation profile. I agree, however, that more needs to be done to disambiguate and test various nearby empirical possibilities, including the important questions about spatial precision mentioned later in Muckli et al's interesting commentary. Such experiments would go some way towards addressing the related issues raised by **Silverstein,** who worries that PP (by suppressing well-predicted signal elements) might not gracefully accommodate cases in which *correlations* between stimulus elements are crucial (e.g., when coding for objects) and need to be highlighted by increasing (rather than suppressing) activity. It is worth noting, however, that such correlations form the very heart of the generative models that are used to predict the incoming sensory patterns. This fact, combined with the co-emergence of both sharpening and dampening, makes the PP class of models well-suited to capturing the full gamut of observed effects.

I turn now to the relation between key elements of PP and the depiction of the brain as performing Bayesian inference. **Trappenberg & Hollensen** note that the space of Bayesian models is large, and they distinguish between demonstrations of Bayesian response profiles in limited experimental tasks and the much grander claim that that such specifics flow from something much more general and fundamental. The latter position is most strongly associated with the work of Karl **Friston**, and is further defended in his revealing commentary. PP is, however, deliberately pitched between these two extremes. It is committed to a general cortical processing strategy that minimizes surprisal using sensorimotor loops that sample the environment while deploying multilevel generative models to predict the ongoing flow of sensation.

Friston's focus is on a presumed biological imperative to reduce surprisal: an imperative obeyed by reducing the organism-computable quantity free energy. Both predictive coding and the Bayesian brain are, Friston argues, results of this surprise minimization mandate. The kinds of processing regime PP describes are thus, Friston claims, the *results* of surprisal minimization rather than its *cause*. Friston may be right to stress that, assuming the free energy story as he describes it is correct, predictive coding and the Bayesian brain emerge as direct consequences of that story. But I do not think the target article displays confusion on this matter. Instead, the issue turns on where we want to place our immediate bets, and perhaps on the Aristotelian distinction between proximate and ultimate causation. Thus, the *proximal cause* (the mechanism) of large amounts of surprisal reduction may well be the operation of a cortical predictive processing regime, even if the *ultimate cause* (the explanation of the presence of that very mechanism) is a larger biological imperative for surprisal minimization itself. This seems no stranger than saying that the reproductive advantages of distal sensing (an ultimate cause) explain the presence of various specific mechanisms (proximal causes) for distal sensing, such as vision and audition. WN, however, deliberately took no firm position on the full free energy story itself.

Friston also notes, importantly, that other ideas that fit within this general framework include ideas about efficient coding. This is correct, and I regard it as a shortfall of my treatment that space precluded discussion of this issue. For, as **Trappenberg & Hollensen** nicely point out, dimensionality reduction using generative models will only yield neurally plausible encodings (filters that resemble actual receptive fields in the brain) if there is pressure to minimize both prediction error *and* the complexity of the encoding itself. The upshot of this is pressure towards various forms of "sparse coding" running alongside the need to reduce prediction error at multiple spatial and temporal scales, and in some acceptably generalizable fashion. Trappenberg & Hollensen suggest that we still lack any concrete model capable of learning to form such sparse representations "due to the world's sparseness" rather than due to the pre-installation of some form of pressure (e.g., an innate hyperprior) towards sparse encodings. But this may be asking too much, given the quite general utility of complexity reduction. Reflecting on the sheer metabolic costs of creating and maintaining internal representations, such a bias seems like a very acceptable ingredient of any "minimal nativism" (Clark 1993).

R2.2. Other mechanisms

I move now to a second set of perils, or challenges. These challenges concern the relation between PP and various other strategies and mechanisms plausibly implicated in human cognitive success. **Ross** draws our attention to a

large and important body of work on "neuroeconomic models of sub-cognitive reward valuation." Such models (e.g., Lee & Wang 2009; Glimcher 2010) posit pre-computed reward valuations (outputs from specialized subsystems performing "striatal valuation") as the inputs to more flexible forms of cortical processing. But despite my intended emphasis on cortical processing, nothing in the PP story was meant to stand in the way of such modes of influence. To be sure, Friston's own ("desert landscape" – see WN, sects. 1.5 and 5.1) attempt to replace reward and value signals with multilevel expectations may at first sight seem inimical to such approaches. But Friston's account ends up accommodating such modes of influence (see, e. g., Friston 2011b), albeit with an importantly different functional and terminological spin. Here (see WN, sect. 3.2, and the commentary by **Friston**), it is important to recognize that predictions and expectations, in Friston's large-scale free energy treatments, are determined by the shape and nature of the whole agent (morphology, reflexes, and subcortical organization included) and are not merely the products of probabilistic models commanded by sensory and motor cortex. Insights concerning the importance of the mid-brain circuitry are compatible both with PP and with the full "desert landscape" version of Friston's own account. This means, incidentally, that the kind of non-cortical route to a (partial) resolution of the Darkened Room problem suggested by Ross (and hinted at also by **Shea**) is in fact equally available to Friston. It is also consistent with (though it is not implied by) the more restricted perspective offered by PP, understood as an account of cortical processing.

Ross's concern that PP may be losing sight of the crucial role played by non-cortical (e.g., environmental, morphological, and subcortical) organization is amplified by **Anderson & Chemero**, who fear that PP puts us on a slippery slope back to full-blown epistemic internalism of the kind I am supposed to have roundly and convincingly (Clark 1997; 2008) rejected. That slope is greased, Anderson & Chemero suggest, by the conflation of two very different senses of prediction. In the first sense, prediction amounts to nothing more than correlation (as in "height predicts weight"), so we might find "predictive processing" wherever we find processing that extracts and exploits correlations. This sense Anderson & Chemero regard as innocent because (involving merely "simple relationships between numbers") it can be deployed without reliance upon inner models, in what they call a model-free or even "knowledge-free" (I would prefer to say "knowledge-sparse") fashion so as to make the most of, for example, reliable cross-modal relationships among sensed information. The second sense is more loaded and "allied with abductive inference and hypothesis testing." It involves the generation of predictions using internal models that posit hidden variables tracking complex causal structure in the body and world. Prediction thus construed is, let us agree, knowledge-rich. Evidence for the utility and ubiquity of prediction in the knowledge-free (or knowledge-sparse) sense provides, just as Anderson & Chemero insist, no evidence for the ubiquity and operation (nor even for the biological possibility) of predictive processing in the second (knowledge-rich) sense.

This is undeniably true and important. But nowhere in the target article did I make or mean to imply such a claim. In displaying the origins of this kind of use of generative models in cognitive science (e.g., Dayan et al.'s [1995] work on the Helmholz machine, leading to all the work on "deep learning" – see Bengio 2009) I was careful to highlight their role in dealing with cases where successful learning required deriving new representations tracking hidden variables. As the story progressed, the role of complex multilevel models learnt and deployed using bidirectional hierarchies (as most clearly implemented by the cortex) was constantly center stage. The larger free energy story, to be sure, covers both the knowledge-rich and knowledge-sparse cases. From the free energy minimization perspective we might even choose to consider (as does **Friston**) the whole embodied, embedded agent as "the model" relative to which surprise is (long-term) minimized. But that story, in turn, does not conflate the two senses of prediction either, since it fluidly covers both. **Anderson & Chemero** suggest that I somehow rely on the (very speculative) model of binocular rivalry to make an illegitimate move from a knowledge-free to a knowledge-rich understanding of prediction. Here, the exposition in WN must be at fault. It may be that they think the account of rivalry plays this role because I preceded it with some remarks on dynamic predictive coding by the retina. But the retinal case, which may indeed be understood as essentially knowledge-sparse and internal-model-free prediction, was meant to illustrate only the predictive coding data compression technique, and not the full PP apparatus. Nor did I intend anything much to turn on the binocular rivalry story itself, which was meant merely as a helpful illustration of how the hypothesis-testing brain might deploy a multi-layered model. It is clear that much more needs to be done to defend and flesh out that account of binocular rivalry (as also pointed out by **Sloman**).

Anderson & Chemero believe that an account might be given that delivers the rivalry response by appealing solely to "low-level, knowledge-free, redundancy-reducing interactions between the eyes." This might turn out to be true, thus revealing the case as closer to that of the retinal ganglion cells than to any case involving hierarchical predictive processing as I defined it. There are, however, very many cases that simply cry out for an inner model–invoking approach. Thus, consider the case of handwritten digit recognition. This is a benchmark task in machine learning, and one that Hinton and Nair (2006) convincingly treat using a complex acquired generative model that performs recognition using acquired knowledge about production. The solution is knowledge-rich because the domain itself is highly structured, exhibiting (like the external world in general) many stacked and nested regularities that are best tracked by learning that unearths multiple interacting hidden variables. I do not think that such cases can be dealt with (at least in any remotely neurally plausible fashion) using resources that remain knowledge-free in the sense that Anderson & Chemero suggest. What seems true (Clark 1989; 1997; 2008) is that to whatever extent a system can avoid the effort and expense of learning about such hidden causes, and rely instead on surface statistics and clever tricks, it will most likely do so. Much of the structure we impose (this relates also to the comments by **Sloman**) upon the designed world is, I suspect, a device for thus reducing elements of the problems we confront to simpler forms (Clark & Thornton 1997). Thus, I fully agree that not all human cognition

depends upon the deployment of what Anderson & Chemero call "high-level, knowledge-rich predictive coding."

What kind of overall cognitive organization, it might be asked, does the embodied, embedded agent then display? Is that organization multiply and impenetrably fractured and firewalled, comprising a motley within which a few principled, knowledge-rich responses bed down with unwashed legions of just-good-enough ploys and stratagems? Surely such a motley is incompatible with the hope for any kind of unifying treatment? This issue (let's call it the Motley Challenge) is among the deepest unresolved questions in cognitive science. **Buckingham & Goodale** join **Ross** and **Anderson & Chemero**, and (as I discuss later) **Sloman** and **Froese & Ikegami**, in pressing the case for the cognitive motley. Following a crisp description of the many successes of Bayesian (i.e., optimal cue integration, given prior probabilities) models in the field of motor control and psychophysics, Buckingham & Goodale turn to some problem cases – cases where Bayesian style optimal integration seems to fail – using these to argue for a fractured and firewalled cognitive economy displaying "independent sets of priors for motor control and perceptual/cognitive judgments, which ultimately serve quite different functions." Poster-child for this dislocation is the size-weight illusion in which similar-looking objects appear weight-adjusted so that we judge the smaller one to feel heavier than the larger despite their identical objective weights (a pound of lead feels heavier, indeed, than a pound of feathers). Buckingham & Goodale survey some intriguing recent work on the size-weight illusion, noting that although Bayesian treatments do manage to get a grip on lifting behavior itself, they fail to explain the subjective comparison effect which some describe as "anti-Bayesian" since prior expectancies and sensory information there seem contrasted rather than integrated (Brayanov & Smith 2010).

Is this a case of multiple, independently operating priors governing various forms of response under various conditions? Perhaps. The first point I would make in response is that nothing either in PP or in the full free-energy formulation rules this out. For the underlying architecture, by dint of evolution, lifetime learning, or both, may come to include "soft modules" partially insulating some response systems from others. To the extent that this is so, that may be traceable, as **Friston** suggests, to the relative statistical independence of various key tracked variables. Information about what an object is, for example, tells us little about where it is, and vice versa, a fact that might explain the emergence of distinct (though not fully mutually insulated – see Schenk & McIntosh 2010) "what" and "where" pathways in the visual brain. Returning to the size-weight illusion itself, Zhu and Bingham (2011) show that the perception of relative heaviness marches delicately in step with the affordance of maximum-distance throwability. Perhaps, then, what we have simply labeled as the experience of "heaviness" is, in some deeper ecological sense, the experience of optimal weight-for-size to afford long-distance throwability? If that were true, then the experiences that **Buckingham & Goodale** describe re-emerge as optimal percepts for throwability, albeit ones that we routinely misconceive as simple but erroneous perceptions of relative object weight. The Zhu and Bingham account is intriguing but remains quite speculative. It reminds us, however, that what looks from one perspective to be a multiple, fragmented, and disconnected cognitive economy may, on deeper examination, turn out to be a well-integrated (though by no means homogeneous) mechanism responding to organism-relevant statistical structure in the environment.

Gerrans continues the theme of fragmentation, resisting the claim that prediction error minimization proceeds seamlessly throughout the cortical hierarchy. His test cases are delusions of alien control. I agree with Gerrans that nothing in the simple story about prediction error minimization explains why it seems that someone else is in control, rather than simply (as in the other cases he mentions) that the action is not under our own control. It is not clear to me, however, why that shortfall should be thought to cast doubt on the more general ("seamlessness") claim that perception phases gently into cognition, and that the differences concern scale and content rather than underlying mechanism.

Silverstein raises some important challenges both to the suggestion that PP provides an adequately general account of the emergence of delusions and hallucinations in schizophrenia, and (especially) to any attempt to extend that account to cover other cases (such as Charles Bonnet syndrome) in which hallucinations regularly emerge without delusions. Importantly, however, I did not mean to suggest that the integrated perceptuo-doxastic account that helps explain the co-emergence of the two positive symptoms in schizophrenia will apply across the board. What might very reasonably be expected, however, is that other syndromes and patterns (as highlighted by **Gerrans**) should be explicable using the same broad apparatus, that is, as a result of different forms of compromise to the very same kind of prediction-error–sensitive cognitive economy. In Charles Bonnet syndrome (CBS), gross damage to the visual system input stream (e.g., by lesions to the pathway connecting the eye to the visual cortex, or by macular degeneration) leads to complex hallucinations without delusion. But this pattern begins to makes sense if we reflect that the gross damage yields what are effectively localized random inputs that are then subjected to the full apparatus of learnt top-down expectation (see Stephan et al. 2009, p. 515). Recent computational work by Reichert et al. (2010) displays a fully implemented model in which hallucinations emerge in just this broad fashion, reflecting the operation of a hierarchical generative (predictive) model of sensory inputs in which inputs are compared with expectations and mismatches drive further processing. The detailed architecture used by Reichert et al. was, however, a so-called Deep Boltzmann Machine architecture (Salakhutdinov & Hinton 2009), a key component of which was a form of homeostatic regulation in which processing elements learn a preferred activation level to which they tend, unless further constrained, to return.

Phillips draws attention to the important question of how a PP-style system selects the right sub-sets of information upon which to base some current response. Information that is critical for one task may be uninformative or counter-productive for another. Appeals to predictive coding or Bayesian inference alone, he argues, cannot provide this. One way in which we might cast this issue, I suggest, is by considering how to select what, at any given moment, to try to predict. Thus, suppose we have an incoming sensory signal and an associated set of prediction errors. For almost any given purpose, it will be best not to

bother about some elements of the sensory signal (in effect, to treat prediction failures there as noise rather than signal). Other aspects, however, ones crucial to the task at hand, will have to be got exactly right (think of trying to spot the four-leaf clover among all the others in a field). To do this, the system must treat even small prediction errors, in respect of such crucial features, as signal and use them to select and nuance the winning top-down model. Within the PP framework, the primary tool for this is, Phillips notes, the use of *context-sensitive gain control*. This amplifies the effects of specific prediction error signals while allowing other prediction errors to self-cancel (e.g., by having that error unit self-inhibit). The same mechanism allows estimates of the relative reliability of different aspects of the sensory signal to be factored in, and it may underpin the recruitment of problem-specific temporary ensembles of neural resources, effectively gating information flow between areas of the brain (see den Ouden et al. [2009] and essays in von der Marlsburg et al. [2010]). On-the-hoof information selection and information coordination of these kinds is, Phillips then argues, a primary achievement of the neurocomputational theory known as "Coherent Infomax" (Kay & Phillips 2010; Phillips et al. 1995). Both Coherent Infomax and PP emphasize the role of prediction in learning and response, and it remains to be determined whether Coherent Infomax is best seen as an *alternative* or (more likely) a *complement* to the PP model, amounting perhaps to a more detailed description of a cortical microcircuit able to act as a repeated component in the construction of a PP architecture.

R2.3. Attention and precision

This brings us to our third set of perils: perils relating to the treatment of attention as a device for upping the gain on (hence the estimated "precision" of) selected prediction errors. **Bowman et al.** raise several important issues concerning the scope and adequacy of this proposal. Some ERP (event-related potential) components (such as P1 and N1), Bowman et al. note, are *increased* when a target appears repeatedly in the same location. Moreover, there are visual search experiments in which visual distractors, despite their rarity, yield little evoked response, yet predescribed, frequently appearing, targets deliver large ones. Can such effects be explained directly by the attention-modulated precision weighting of residual error? A recent fMRI study by Kok et al. (2012) lends elegant support to the PP model of such effects by showing that these are just the kinds of interaction between prediction and attention that the model of precision-weighted prediction error suggests. In particular, Kok et al. show that predicted stimuli that are unattended and task-irrelevant result in reduced activity in early visual cortex (the "silencing" of the predicted, as mandated by simple predictive coding) but that "this pattern reversed when the stimuli were attended and task-relevant" (Kok et al. 2012, p. 2198). The study manipulated spatial attention and prediction by using independent prediction and spatial cues (for the details, see the original paper by Kok et al.) and found that attention reversed the silencing effect of prediction upon the sensory signal, in just the way the precision-weighting account would specify. In addition, Feldman and Friston (2010) present a detailed, simulation-based model in which precision-modulated prediction error is used to optimize perceptual inference in a way that reproduces the ERP and psychophysical responses elicited by the Posner spatial cueing paradigm (see Posner 1980).

Bowman et al. go on to press an important further question concerning feature-based attention. For, feature-based attention seems to allow us to enhance response to a given feature even when it appears at an unpredicted location. In their example, the command to find an instance of bold type may result in attention being captured by a nearby spatial location. If the result of that is to increase the precision-weighting upon prediction error from that spatial location (as PP suggests) that seems to depict the precision weighting as a *consequence* of attending rather than a *cause or implementation of attending*. The resolution of this puzzle lies, I suggest, in the potential assignment of precision-weighting at many different levels of the processing hierarchy. Feature-based attention corresponds, intuitively, to increasing the gain on the prediction error units associated with the identity or configuration of a stimulus (e.g., increasing the gain on units responding to the distinctive geometric pattern of a four-leaf clover). Boosting that response (by giving added weight to the relevant kind of sensory prediction error) should enhance detection of that featural cue. Once the cue is provisionally detected, the subject can fixate the right spatial region, now under conditions of "four-leaf-clover-there" expectation. Residual error is then amplified for that feature at that location, and high confidence in the presence of the four-leaf clover can (if you are lucky!) be obtained. Note that attending to the wrong spatial region (e.g., due to incongruent spatial cueing) will actually be counter-productive in such cases. Precision-weighted prediction error, as I understand it, is thus able to encompass both mere-spatial and feature-based signal enhancement.

Block & Siegel claim that predictive processing (they speak simply of predictive coding, but they mean to target the full hierarchical, precision-modulated, generative-model based account) is unable to offer any plausible or distinctive account of very basic results such as the attentional enhancement of perceived contrast (Carrasco et al. 2004). In particular, they claim that the PP model fails to capture changes due to attending that *precede* the calculation of error, and that it *falsely predicts* a magnification of the changes that follow from attending (consequent upon upping the gain on some of the prediction error). However, I find Block & Siegel's attempted reconstruction of the PP treatment of such cases unclear or else importantly incomplete. In the cases they cite, subjects fixate a central spot with contrast gratings to the left and right. The gratings differ in absolute (actual) contrast. But when subjects are cued to attend (even covertly) to the lower contrast grating, their perception of the contrast there is increased, yielding the (false) judgment that, for example, an attended 70% (actual value) contrast grating is the same as an unattended 82% grating. Block & Siegel suggest that the PP account cannot explain the initial effect here (the false perception of an 82% contrast for the covertly attended 70% contrast grating) as the only error signal – but this is where they misconstrue the story – is the difference between the stable pre-attentive 70% registration and the post-attentive 82% one. But this difference, they worry, wasn't available until after attention had done its work! Worse still, once that difference is

available, shouldn't it be amplified once more, as the PP account says that gain on the relevant error units is now increased?

This is an ingenious challenge, but it is based on a misconstrual of the precision-weighting proposal. It is not the case that PP posits an error signal calculated on the basis of a difference between the unattended contrast (registered as 70%) and the subsequently attended contrast (now apparently of 82%). Rather, what attention alters is the *expectation of precise sensory information* from the attended spatial location. Precision is the inverse of the variance, and it is our "precision expectations" that attention here alters. What seems to be happening, in the case at hand, is that the very fact that we covertly attend to the grating on the left (say) increases our expectations of a precise sensory signal. Under such conditions, the expectation of precise information induces an inflated weighting for sensory error and our subjective estimate of the contrast is distorted as a result. The important point is that the error is not computed, as **Block & Siegel** seem to suggest, as a *difference* between some prior (in this case unattended) percept and some current (in this case attended) one. Instead, it is computed directly for the present sensory signal itself, but weighted in the light of our expectation of precise sensory information from that location. Expectations of precision are what, according to PP, is being manipulated by the contrast grating experiment, and PP thus offers a satisfying and distinctive account of the effect itself. This same mechanism explains the general effect of attention on spatial acuity, especially in cases where we alter fixation and where more precise information is indeed then available. Block & Siegel are right to demand that the PP framework confront the full spectrum of established empirical results in this area. But they underestimate the range of apparatus (and the distinctiveness of the accounts) that PP can bring to bear. This is not surprising, since these are early days and much further work is needed. For an excellent taste, however, of the kind of detailed, probing treatment of classic experimental results that is already possible, see Hohwy's (2012) exploration of conscious perception, attention, change blindness, and inattentional blindness from the perspective of precision-modulated predictive processing.

R3. Prospects

I have chosen to devote the bulk of this Response to addressing the various perils and pitfalls described above and to some even grander ones to be addressed in section 4 further on. A reassuringly large number of commentators, however, have offered illuminating and wonderfully constructive suggestions concerning ways in which to improve, augment, and extend the general picture. I'm extremely grateful for these suggestions, and plan to pursue several of them at greater length in future work. For present purposes, we can divide the suggestions into two main (though non-exclusive) camps: those which add detail or further dimensions to the core PP account, extending it to embrace additional mental phenomena, such as timing, emotion, language, personal narrative, and high-level forms of "self-expectation"; and those which reach out to the larger organizational forms of music, culture, and group behaviors.

R3.1. New dimensions

Shea usefully points out that perception and action, even assuming they indeed share deep computational commonalities, would still differ in respect of their "direction of fit." In (rich, world-revealing) perception, we reduce prediction error by selecting a model that explains away the sensory signal. In world-engaging action, we reduce prediction error by altering body and world to conform to our expectations. This is correct, and it helps show how the PP framework, despite offering a single fundamental model of cortical processing, comports with the evident multiplicity and variety of forms of cognitive contact with the world.

Farmer, Brown, & Tanenhaus (**Farmer et al.**) suggest (this was music to my ears) that the hierarchical prediction machine perspective provides a framework that might one day "unify the literature on prediction in language processing." They describe, in compelling detail, the many applications of prediction-and-generative-model-based accounts to linguistic phenomena. Language, indeed, is a paradigm case of an environmental cause that exhibits a complex, multilevel structure apt for engagement using hierarchical, generative models. Farmer et al. stress several aspects of language comprehension that are hard to explain using traditional models. All these aspects revolve (it seems to me) around the fact that language comprehension involves not "throwing away" information as processing proceeds, so much as using all the information available (in the signal, in the generative model, and in the context) to get a multi-scale, multi-dimensional grip on the evolving acoustic and semantic content. All manner of probabilistic expectation (including speaker-specific lexical expectations formed "on-the-hoof" as conversation proceeds) are thus brought to bear, and impact not just recognition but production (e.g., your own choice of words), too. Context effects, rampant on-the-hoof probability updating, and cross-cueing are all grist to the PP mill.

The PP framework, **Holm & Madison** convincingly argue, also lends itself extremely naturally to the treatment of timing and of temporal phenomena. In this regard, Holm & Madison draw our attention to large and important bodies of work that display the complex distribution of temporal control within the brain, and that suggest a tendency of later processing stages and higher areas to specialize in more flexible and longer time-scale (but correlatively less dedicated, and less accurate) forms of time-sensitive control. Such distributions, as they suggest, emerge naturally within the PP framework. They emerge from both the hierarchical form of the generative model and the dynamical and multi-scale nature of key phenomena. More specifically, the brain must learn a generative model of coupled dynamical processes spanning multiple temporal scales (a nice example is Friston and Kiebel's [2009] simulation of birdsong recognition). Holm & Madison (and see comments by **Schaefer, Overy, & Nelson** [**Schaefer et al.**]) also make the excellent point that action (e.g., tapping with hands and feet) can be used to bootstrap timing, and to increase the reliability of temporal perception. This provides an interesting instance of the so-called "self-structuring of information" (Pfeifer et al. 2007), a

key cognitive mechanism discussed in Clark (2008) and in the target article (see WN, sect. 3.4).

Gowaty & Hubbell suggest that *all* animals are Bayesians engaged in predicting the future on the basis of flexibly updated priors, and that they "imagine" (the scare quotes are theirs) alternatives and make choices among them. This is an intriguing hypothesis, but it is one that remains poised (thanks in part to those scare quotes) between two alternative interpretations. On the one hand, there is the (plausible) claim that elements in the systemic organization of all animals respond sensitively, at various timescales, to environmental contingencies so as to minimize free energy and allow the animals to remain within their envelope of viability. On the other hand, there is the (to me less plausible) claim that all animals ground flexible behavioral response in the top-down deployment of rich, internally represented generative models developed and tuned using prediction-driven learning routines of the kind described by PP. I return to this issue in section 4.

Seth & Critchley sketch a powerful and potentially very important bridge between PP-style work and new cognitive scientific treatments of emotion and affect. The proposed bridge to emotion relies on the idea that interoception (the "sense of the physiological condition of the body") provides a source of signals equally apt for prediction using the kinds of hierarchical generative models described in the target article The step to emotion is then accomplished (according to their "interoceptive predictive coding" account – see Seth et al. 2011) by treating emotional feelings as determined by a complex exchange between driving sensory (especially interoceptive) signals and multilevel downwards predictions. Of special interest here are signals and predictions concerning visceral, autonomic, and motor states. Attention to predictions (and pathologies of prediction) concerning these states provides, Seth & Critchley plausibly suggest, a major clue to the nature and genesis of many psychiatric syndromes. Dissociative syndromes, for example, may arise from mistaken assignments of precision (too little, in these cases) to key interoceptive signals. But are emotional feelings here constructed by successful predictions (by analogy to the exteroceptive case)? Or are feelings of emotion more closely tied (see also the comments by **Schaefer et al.** regarding prediction error in music) to the prediction errors themselves, presenting a world that is an essentially moving target, defined more by what it is not than by what it is? Or might (this is my own favorite) the division between emotional and non-emotional components itself prove illusory, at least in the context of a multi-dimensional, generative model – nearly every aspect of which can be permeated (Barrett & Bar 2009) by goal and affect-laden expectations that are constantly checked against the full interoceptive and exteroceptive array?

Dennett's fascinating and challenging contribution fits naturally, it seems to me, with the suggestions concerning interoceptive self-monitoring by **Seth & Critchley**. Just how could some story about neural prediction illuminate, in a deep manner, our ability to experience the baby as cute, the sky as blue, the honey as sweet, or the joke as funny? How, in these cases, does the way things seem to us (the daily "manifest image") hook up with the way things actually work? The key, Dennett suggests, may lie in our expectations about our own expectations. The cuteness of the baby, if I read Dennett correctly, is nothing over and above our expectations concerning our probable reactions (themselves rooted, if the PP story is correct, in a bunch of probabilistic expectations) to imminent baby-exposure. We expect to feel like cooing and nurturing, and when those expectations (which can, in the manner of action-oriented predictive processing, be partially self-fulfilling) are met, we deem the *baby itself* cute. This is what Dennett (2009) describes as a "strange inversion," in which we seem to project our own reactive complexes outward, populating our world with cuteness, sweetness, blueness, and more. I think there is something exactly right, and something that remains unclear, in Dennett's sketch. What seems exactly right is that we ourselves turn up as one crucial item among the many items that we humans model when we model our world. For, we ourselves (not just as organisms but as individuals with unique histories, tendencies, and features) are among the many things we need to get a grip upon if we are to navigate the complex social world, predicting our own and others' responses to new situations, threats, and opportunities.

To that extent (see also Friston 2011a), **Dennett** is surely right: We must develop a grip (what Dennett describes as a set of "Bayesian expectations") upon how we ourselves are likely to react, and upon how others model us. Our *Umwelt*, as Dennett says, is thus populated not just with simple affordances but with complex expectations concerning our own nature and reactions. What remains unclear, I think, is just how this complex of ideas hooks up the question with which Dennett precedes it, namely, "what makes our manifest image *manifest* (to us)?" For this, on the face of it, is a question about the origins of consciously perceived properties: the origins of awareness, or of something like it – something special that we have and that the elevator (in Dennett's example) rather plausibly lacks. It does not strike me as impossible that there might be a link here, perhaps even a close one. But how does it go? Is the thought that any system that models itself and has expectations about its own reactive dispositions, belongs to the class of the consciously aware? That condition seems both too weak (too easily satisfied by a simple artificial system) and too strong (as there may be conscious agents who fail to meet it). Is it that any system that models itself in that way will at least judge (perhaps self-fulfillingly) that it is consciously aware of certain things, such as the cuteness of babies? That's tempting, but we need to hear more. Or is this really just a story – albeit a neat and important one – about how, *assuming* a system is somehow conscious of some of the things in its world, those things might (if you are a sufficiently bright and complex social organism under pressure to include yourself in your own generative model) come to include such otherwise elusive items as cuteness, sweetness, funniness, and so on?

Hirsh, Mar, & Peterson (**Hirsh et al.**) suggest that an important feature of the predictive mosaic, when accounting for distinctively human forms of understanding, might be provided by the incorporation of personal narratives as high-level generative models that structure our predictions in a goal- and affect-laden way. This proposal sits well with the complex of ideas sketched by **Dennett** and by **Seth & Critchley**, and it provides, as they note, a hook into the important larger sociocultural circuits (see also comments by **Roepstorff**, and section 4 further on) that also sculpt

and inform human behavior. Personal narratives are often co-constructed with others, and can feed the structures and expectations of society back in as elements of the generative model that an individual uses to make sense of their own acts and choices. Hirsh et al., like **Dennett**, are thus concerned with bridging the apparent gap between the manifest and scientific image, and accounts that integrate culturally inflected personal narratives into the general picture of prediction-and generative-model based cognition seem ideally placed to play this important role. Narrative structures, if they are correct, lie towards the very top of the predictive hierarchy, and they influence and can help coordinate processing at every level beneath. It is not obvious to me, however, that personal narrative needs to be the concern of a clearly demarcated higher level. Instead, a narrative may be defined across many levels of the processing hierarchy, and supported (in a graded rather than all-or-none fashion) by multiple interacting bodies of self-expectation.

R3.2. Larger organizational forms

This brings us to some comments that directly target the larger organizational forms of music, culture, and group behaviors. Many aspects of our self-constructed sociocultural world, **Paton, Skewes, Frith, & Hohwy** (**Paton et al.**) argue, can be usefully conceptualized as devices that increase the reliability of the sensory input, yielding a better signal for learning and for online response. A simple example might be the use of windscreen wipers in the rain. But especially illuminating, in this regard, are their comments on conversation, ritual, convention, and shared practices. In conversation, speakers and listeners often align their uses (e.g., lexical and grammatical choices – see Pickering & Garrod 2007). This makes good sense under a regime of mutual prediction error reduction. But conversants may also, as Paton et al. intriguingly add, align their mental states in a kind of "fusion of expectation horizons." When such alignment is achieved, the otherwise blunt and imprecise tools of natural language (see Churchland 1989; 2012) can be better trusted to provide reliable information about another's ideas and mental states. Such a perspective ("neural hermeneutics"; Frith & Wentzer, in press) extends naturally to larger cultural forms, such as ritual and shared practice, which (by virtue of being shared) enhance and ensure the underlying alignment that improves interpersonal precision. Culture, in this sense, emerges as a prime source of shared hyperpriors (high-level shared expectations that condition the lower-level expectations that each agent brings to bear) that help make interpersonal exchange both possible and fruitful. Under such conditions (also highlighted by **Roepstorff**) we reliably discern each other's mental states, inferring them as further hidden causes in the interpersonal world. Natural hermeneutics may thus contribute to the growing alignment between the humanities and the sciences of mind (**Hirsh et al.**). At the very least, this offers an encompassing vision that adds significant dimensions to the simple idea of mutual prediction error reduction.

Schaefer et al. combine the themes of mutual prediction error reduction, culture, and affect. Their starting point is the idea that music (both in perception and production) creates a context within which prediction error – mutual prediction error, in the group case – is reduced. But this simple idea, they argue, needs augmenting with considerations of arousal, affect, and the scaffolding effects of cultural, training, and musical style. There is, Schaefer et al. suggest, an optimal or preferred level of surprisal at which musical experience leads to maximal (positive) affective response. That level is not uniform across musical types, musical features, or even individuals, some of whom may be more "thrill-seeking" than others. The commentary provides many promising tools for thinking about these variations, but makes one claim that I want to question (or at any rate probe a little), namely, that affect is what "makes prediction error in music … meaningful, and indeed determines its value." This is tricky ground, but I suspect it is misleading (see also comments on **Seth & Critchley**) to depict prediction error as, if you like, something that is given in experience, and that itself *generates* an affective response, rather than as that which (sub-personally) *determines* the (thoroughly affect-laden) experience itself. I am not convinced, that is to say, that I experience my own prediction errors (though I do, of course, sometimes experience surprise).

R4. Darkened rooms and the puzzle of the porous perceiver

R4.1. Darkened rooms

Several commentators (**Anderson & Chemero**, **Froese & Ikegami**, **Sloman**, and to a lesser extent **Little & Sommer**) have questioned the idea of surprisal minimization as the underlying imperative driving all forms of cognition and adaptive response. A recurrent thread here is the worry that surprisal minimization alone would incline the error-minimizing agent to find a nice "darkened room" and just stay there until they are dead. Despite explicitly bracketing the full free-energy story, WN did attempt (in sects. 3.2–3.4) to address this worry, with apparently mixed results. Little & Sommer argue that the solution proffered depends unwholesomely upon innate knowledge, or at least upon pre-programmed expectations concerning the shape (itinerant, exploratory) of our own behavior. Froese & Ikegami contend (contrary to the picture briefly explored in WN, sect. 3.2) that good ways of minimizing surprisal will include "stereotypic self-stimulation, catatonic withdrawal from the world, and autistic withdrawal from others."

Hints of a similar worry can be found in the comments by **Schaefer et al.,** who suggest that musical appreciation involves not the simple quashing of prediction error (perhaps that might be achieved by a repeated pulse?) but attraction towards a kind of sweet spot between predictability and surprise: an "optimal level of surprisal," albeit one that varies from case to case and between individuals and musical traditions. As a positive suggestion, **Little & Sommer** then suggest we shift our attention from the minimization of prediction error to the maximization of mutual information. That is to say, why not depict the goal as *maximizing the mutual information* (on this, see also **Phillips**) between an internal model of estimated causes and the sensory inputs? Minimizing entropy (prediction error) and maximizing mutual information (hence prediction success), Little & Sommer argue, each deliver minimal prediction error but differ in how they select

actions. A system that seeks to maximize mutual information won't, they suggest, fall into the dark room trap. For, it is driven instead towards a sweet spot between predictability and complexity and will "seek out conditions in which its sensory inputs vary in a complex, but still predictable, fashion."

Many interesting issues arise at this point. For example, we might also want to *minimize* mutual information (redundancy) among outputs (as opposed to between inputs and model) so as to achieve sparse, efficient coding (Olshausen & Field 1996). But for present purposes, the main point to make is that any improvement afforded by the move to mutual information is, as far as I can determine, merely cosmetic. Thus, consider a random system driven towards some sweet spot between predictability and complexity. For that system, there will be some complex set of inputs (imagine, to be concrete, a delicate, constantly changing stream of music) such that the set of inputs affords, for that agent, the perfect balance between predictability and complexity. The musical stream can be as complex as you like. Perhaps it must be so complex as never quite to repeat itself. Surely the agent must now enter the "musical room" and stay there until it is dead? The musical room, I contend, is as real (and, more important, as unreal) a danger as the darkened room. Notice that you can ramp up the complexity at will. Perhaps the sweet spot involves complex shifts between musical types. Perhaps the location of the sweet spot varies systematically with the different types. Make the scenario as complex as you wish. For that complexity, there is some musical room that now looks set to act as a death trap for that agent.

There is, of course, a perfectly good way out of this. It is to notice, with **Friston**, that all the key information-theoretic quantities are defined and computed relative to a type of agent – a specific kind of creature whose morphology, nervous system, and neural economy already render it (but *only* in the specific sense stressed by Friston; more on this shortly) a complex model of its own adaptive niche. As such, the creature, simply because it is the creature that it is, already embodies a complex set of "expectations" concerning moving, eating, playing, exploring, and so forth. It is because surprisal at the very largest scale is minimized against the backdrop of this complex set of creature-defining "expectations" that we need fear neither darkened nor musical (nor meta-musical, nor meta-meta-musical) rooms. The free-energy principle thus subsumes the mutual information approach (for a nice worked example, see Friston et al. 2012). The essential creature-defining backdrop then sets the scene for the deployment (sometimes, in some animals) of PP-style strategies of cortical learning in which hierarchical message passing, by implementing a version of "empirical Bayes," allows effective learning that is barely, if at all, hostage to initial priors. That learning requires ongoing exposure to rich input streams. It is the backdrop "expectations," deeply built-in to the structure of the organism (manifesting as, for example, play, curiosity, hunger, and thirst) that keep the organism alive and the input stream rich, and that promote various beneficial forms of "self-structuring" of the information flow – see Pfeifer et al. (2007).

This means that the general solution to the darkened room worry that was briefly scouted in WN, section 3.2, is mandatory, and that we must embrace it whatever our cosmetic preferences concerning entropy versus mutual information. This also means that the suggestion (**Froese & Ikegami**) that enactivism offers an alternative approach, with a distinctive resolution of the dark room issue, is misguided. Indeed, the "two" approaches are, with respect to the darkened room issue at least, essentially identical. Each stresses the autonomous dynamics of the agent. Each depicts agents as moving through space and time in ways determined by "the viability constraints of the organism." Each grounds value, ultimately, in those viability constraints (which are the essential backdrop to any richer forms of lifetime learning).

Froese & Ikegami also take PP (though they dub it HPM: the "Hierarchical Prediction Machine" story) to task for its commitment to some form of representationalism. This commitment leads, they fear, to an unacceptable internalism (recall also the comments from **Anderson & Chemero**) and to the unwelcome erection of a veil between mind and world. This issue arises also (although from essentially the opposite direction) in the commentary by **Paton et al.** Thus, Froese & Ikegami fear that the depiction of the cerebral cortex as commanding probabilistic internal models of the world puts the world "off-limits," while Paton et al. suggest that my preferred interpretation of the PP model makes the mind–world relation too direct and obscures the genuine sense in which "perception remains an inferred fantasy about what lies behind the veil of input." I find this strangely cheering, as these diametrically opposed reactions suggest that the account is, as intended, walking a delicate but important line. On the one hand, I want to say that perception – rich, world-revealing perception of the kind that we humans enjoy – involves the top-down deployment of generative models that have come, courtesy of prediction-driven learning within the bidirectional cortical hierarchy, to embody rich, probabilistic knowledge concerning the hidden causes of our sensory inputs. On the other hand, I want to stress that those same learning routines make us extremely porous to the statistical structure of the actual environment, and put us perceptually in touch, in as direct a fashion as is mechanistically possible, with the complex, multilayered, world around us.

R4.2. The puzzle of the porous perceiver

This, then, is the promised "puzzle of the porous perceiver": Can we both experience the world via a top-down generative-model based cascade and be in touch not with a realm of internal fantasy but, precisely, with the world? One superficially tempting way to try to secure a more direct mind–world relation is to follow **Froese & Ikegami** in rejecting the appeal to internally represented models altogether (we saw hints of this in the comments by **Anderson & Chemero** too). Thus, they argue that "Properties of the environment do not need to be encoded and transmitted to higher cortical areas, but not because they are already expected by an internal model of the world, but rather because the world is its own best model." But I do not believe (nor have I ever believed: see, e.g., Clark 1997, Ch. 8) that this strategy can cover all the cases, or that, working alone, it can deliver rich, world-revealing perception of the kind we humans enjoy – conscious perception of a world populated by (among other things) elections, annual rainfall statistics, prayers, paradoxes, and poker hands. To experience a world rich in such multifarious hidden causes we must do

some pretty fancy things, at various time-scales, with the incoming energetic flux: things at the heart of which lie, I claim, the prediction-driven acquisition and top-down deployment of probabilistic generative models. I do not believe that prayers, paradoxes, or even poker hands can be their own best model, if that means they can be known without the use of internal representations or inner models of external hidden causes. Worse still, in the cases where we might indeed allow the world, and directly elicited actions upon the world, to do most of the heavy lifting (the termite mound-building strategies mentioned by **Sloman** are a case in point) it is not obvious that there will – simply by virtue of deploying such strategies alone – be any world-presenting experience at all. What seems exactly right, however, is that brains like ours are masters of what I once heard Sloman describe as "productive laziness." Hence, we will probably not rely on a rich internal model when the canny use of body or world will serve as well, and many of the internal models that we do use will be partial at best, building in all kinds of calls (see Clark 2008) to embodied, problem-simplifying action. The upshot is that I did not intend (despite the fears of Anderson & Chemero) to depict all of cognition and adaptive response as grounded in the top-down deployment of knowledge-rich internal models. But I do think such models are among the most crucial achievements of cortical processing, and that they condition both online and offline forms of human experience.

Nor did I intend, as **Sloman** in a kind of reversal of the worry raised by **Anderson & Chemero** fears, to reduce all cognition to something like online control. Where Anderson & Chemero subtly mislocate the PP account as an attempt to depict *all* cognition as rooted in procedures apt only for high-level knowledge-rich response, Sloman subtly mislocates it as an over-ambitious attempt to depict *all* cognition as rooted in procedures apt only for low-level sensorimotor processing. Sloman thus contrasts prediction with interpretation, and stresses the importance to human (and animal) reasoning of multiple meta-cognitive mechanisms that (he argues) go far beyond the prediction and control of gross sensory and motor signals. In a related vein, **Khalil** interestingly notes that human cognition includes many "conception-laden processes" (such as choosing our own benchmark for a satisfactory income) that cannot be corrected simply by adjustments that better track gross sensory input.

Fortunately, there are no deep conflicts here. PP aims only to describe a core cortical processing strategy: a strategy that can deliver probabilistic generative models apt both for basic sensorimotor control and for more advanced tasks. The same core strategy can drive the development of generative models that track structure within highly abstract domains, and assertions concerning such domains can indeed resist simple perceptual correction. To say that the mechanisms of (rich, world-presenting) perception are continuous with the mechanisms of (rich, world-presenting) cognition is not to deny this. It may be, however, that learning about some highly abstract domains requires delivering structured symbolic inputs; for example, using the formalisms of language, science and mathematics. Understanding how prediction-driven learning interacts with the active production and uptake of external symbolic representations and with various forms of designer learning environments is thus a crucial challenge, as **Roepstorff** also notes.

PP thus functions primarily as an intermediate-level (see S**pratling**) description of the underlying form of cortical processing. This is the case even though the larger story about free energy minimization (a story I briefly sketched in WN, sect. 1.6, but tried to bracket as raising issues far beyond the scope of the article) aims to encompass far more. As a theory of cortical processing, PP suggests we learn to represent linked sets of probability density distributions, and that they provide the form of hierarchical generative models underlying both perception (of the rich, world-presenting variety) and many forms of world-engaging action. Importantly, this leaves plenty of space for other ploys and strategies to coexist with the core PP mechanism. I tried to celebrate that space by making a virtue (WN, sects. 3.2–3.4) out of the free-energy story's failure to specify the full form of a cognitive architecture, envisaging a cooperative project requiring many further insights from evolutionary, situated, embodied, and distributed approaches to understanding mind and adaptive response. Was it then false advertising to offer PP itself as a unifying account? Not, I fondly hope, if PP reveals common computational principles governing knowledge-rich forms of cortical processing (in both the sensory and motor realms), delivers a novel account of attention (as optimizing precision), and reveals prediction error minimization as the common goal of many forms of action, social engagement, and environmental structuring.

There is thus an important difference of emphasis between my treatment and the many seminal treatments by Karl Friston. For as the comments by **Friston** made clear, he himself sets little store by the difference between what I (like **Anderson & Chemero**) might describe as knowledge- and inner-model-rich versus knowledge-sparse ways of minimizing free energy and reducing surprisal. Viewed from the loftier perspective of free-energy minimization, the effect is indeed the same. Free-energy reduction can be promoted by the "fit" between morphology and niche, by quick-and-dirty internal-representation-sparse ploys, and by the costlier (but potent) use of prediction-driven learning to infer internally represented probabilistic generative models. But it is, I suspect, only that costlier class of approaches, capable of on-the-hoof learning about complex interanimated webs of hidden causes, that delivers a certain "cognitive package deal." The package deal bundles together what I have been calling "rich, world-presenting perception," offline imagination, and understanding (not just apt response) and has a natural extension to intentional, world-directed action (see Clark, forthcoming). Such a package may well be operative, as **Gowaty & Hubbell** suggested, in the generation of many instances of animal response. It need not implicate solely the neocortex (though that seems to be its natural home). But potent though the package is, it is not the only strategy at work, even in humans, and there may be some animals that do not deploy the strategy at all.

Thus, consider the humble earthworm. The worm is doubtless a wonderful minimizer of free energy, and we might even describe the whole worm (as the comments by **Friston** suggest) as a kind of free-energy minimizing model of its world. But does the worm command a model of its world parsed into distal causes courtesy of top-down expectations applied in a multilevel manner?

This is far from obvious. The worm is capable of sensing, but perhaps it does not thereby experience a perceptual world. If that is right, then not all ways of minimizing free energy are equal, and only some put you in perceptual touch (as opposed to mere causal contact) with a distal world characterized by multiple interacting hidden causes.

This brings us back, finally, to the vexed question of the mind–world relation itself. Where **Froese & Ikegami** fear that the PP strategy cuts us off from the world, inserting an internal model between us and the distal environment, I believe that it is only courtesy of such models that we (perhaps unlike the earthworm) can experience a distal environment at all! Does that mean that perception presents us (as **Paton et al.** suggest) with only a fantasy about the world? I continue to resist this way of casting things. Consider (recall the comments by **Farmer et al.**) the perception of sentence structure during speech processing. It is plausibly only due to the deployment of a rich generative model that a hearer can recover semantic and syntactic constituents from the impinging sound stream. Does that mean that the perception of sentence structure is "an inferred fantasy about what lies behind the veil of input"? Surely it does not. In recovering the right set of interacting distal causes (subjects, objects, meanings, verb-clauses, etc.) we see through the sound stream to the multilayered structure and complex purposes of the linguistic environment itself. This is possible because brains like ours are sensitive statistical sponges open to deep restructuring by the barrage of inputs coming from the world. Moreover, even apparently low-level structural features of cortex (receptive field orientations and spatial frequencies), as **Bridgeman** very eloquently reminds us, come to reflect the actual statistical profile of the environment, and do so in ways that are surprisingly open to variation by early experience.

Does this commit me to the implausible idea that perception presents us with the world "as it is in itself"? Here, the helpful commentary by **König, Wilming, Kaspar, Nagel, & Onat** (**König et al.**) seems to me to get the issue exactly right. Predictions are made, they stress (see also the comments by **Bridgeman**), in the light of our own action repertoire. This simple (but profound) fact results in reductions of computational complexity by helping to select what features to process, and what things to try to predict. From the huge space of possible ways of parsing the world, given the impinging energetic flux, we select the ways that serve our needs by fitting our action repertoires. Such selection will extend, as **Paton et al.** have noted (see also **Dennett**), to ways of parsing and understanding our own bodies and minds. Such parsing enables us to act on the world, imposing further structure on the flow of information, and eventually reshaping the environment itself to suit our needs.

Roepstorff's engaging commentary brings several of these issues into clearer focus by asking in what ways, if any, the PP framework illuminates specifically human forms of cognition. This is a crucial question. The larger free-energy story targets nothing that is specifically human, though (of course) it aims to encompass human cognition. The PP framework seeks to highlight a cortical processing strategy that, though not uniquely human, is plausibly essential to human intelligence and that provides, as mentioned above, a compelling "cognitive package deal." That package deal delivers, at a single stroke, understanding of complex, interacting distal causes and the ability to generate perception-like states from the top down. It delivers understanding because to perceive the world of distal causes in this way is not just to react appropriately to it. It is to know how that world will evolve and alter across multiple timescales. This, in turn, involves learning to generate perception-like states from the top-down. This double-innovation, carefully modulated by the precision-weighting of attention, lies (PP claims) at the very heart of many distinctively human forms of cognition. To be sure (recall **Gowaty & Hubbell**) the same strategy is at work in many nonhuman animals, delivering there too a quite deep understanding of a world of distal causes. What, then, is special about the human case?

Roepstorff points to a potent complex of features of human life, especially our abilities of temporally co-coordinated social interaction (see also commentaries by **Holm & Madison**, **Paton et al.**, and **Schaefer et al.**) and our (surely deeply related) abilities to construct artifacts and designer environments. Versions of all of this occur in other species. But in the human case, the mosaic comes together under the influence of flexible structured symbolic language and an almost obsessive drive to engage in shared cultural practices. We are thus enabled repeatedly to redeploy our core cognitive skills in the transformative context of exposure to patterned sociocultural practices, including the use of symbolic codes (encountered as "material symbols"; Clark 2006a) and complex social routines (Hutchins 1995; Roepstorff et al. 2010). If, as PP claims, one of the most potent inner tools available is deep, prediction-driven learning that locks on to interacting distal hidden causes, we may dimly imagine (WN, sect. 3.4; Clark 2006; 2008) a virtuous spiral in which our achieved understandings are given concrete and communicable form, and then shared and fed back using structured practices that present us with new patterns.

Such pattern-presenting practices should, as **Roepstorff** suggests, enable us to develop hierarchical generative models that track ever more rarefied causes spanning the brute and the manufactured environment. By tracking such causes they may also, in innocent ways, help create and propagate them (think of patterned practices such as marriage and music). It is this potentially rich and multilayered interaction between knowledge-rich prediction-driven learning and enculturated, situated cognition that most attracts me to the core PP proposal. These are early days, but I believe PP has the potential to help bridge the gap between simpler forms of embodied and situated response, the self-structuring of information flows, and the full spectrum of socially and technologically inflected human understanding.

References

[The letters "a" and "r" before author's initials stand for target article and response references, respectively]

Abelson, R. P. (1981) Psychological status of the script concept. *American Psychologist* 36(7):715–29. [JBH]

Adams, F. & Aizawa, K. (2001) The bounds of cognition. *Philosophical Psychology* 14(1):43–64. [aAC]

Alais, D. & Blake, R., eds. (2005) *Binocular rivalry*. MIT Press. [aAC]

Alais, D. & Burr, D. (2004) The ventriloquist effect results from near-optimal bimodal integration. *Current Biology* 14:257–62. [aAC]

Alink, A., Schwiedrzik, C. M., Kohler, A., Singer, W. & Muckli, L. (2010) Stimulus predictability reduces responses in primary visual cortex. *Journal of Neuroscience* 30:2960–66. [aAC, TE, LM]

Anderson, C. H. & Van Essen, D. C. (1994) Neurobiological computational systems. In: *Computational intelligence: Imitating life*, ed. J. M. Zurada, R. J. Marks & C. J. Robinson, pp. 213–22. IEEE Press. [aAC]

Anderson, M. L. (2006) Cognitive epistemic openness. *Phenomenology and the Cognitive Sciences* 5(2):125–54. [MLA]

Anderson, M. L. (2007) The massive redeployment hypothesis and the functional topography of the brain. *Philosophical Psychology* 20(2):143–74. [aAC]

Angelucci, A., Levitt, J. B., Walton, E. J. S., Hupe, J. M., Bullier, J. & Lund, J. S. (2002) Circuits for local and global signal integration in primary visual cortex. *Journal of Neuroscience* 22:8633–864. [LM]

Anton-Erxleben, K., Henrich, C. & Treue, S. (2007) Attention changes perceived size of moving visual patterns. *Journal of Vision* 7(11):1–9. [NB]

Arnold, J. E., Hudson, C. L. & Tanenhaus, M. K. (2007) If you say thee uh you are describing something hard: The on-line attribution of disfluency during reference comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 33:914–30. [TAF]

Arthur, B. (1994) *Increasing returns and path dependence in the economy*. University of Michigan Press. [aAC]

Ashby, W. R. (1940) Adaptiveness and equilibrium. *The British Journal of Psychiatry* 86:478–83. [TF]

Ashby, W. R. (1947) Principles of the self-organizing dynamic system. *Journal of General Psychology* 37:125–28. [KF]

Ay, N., Bertschinger, N., Der, R., Güttler, F. & Olbrich, E. (2008) Predictive information and explorative behavior of autonomous robots. *The European Physical Journal B – Condensed Matter and Complex Systems* 63(3):32939. [DYL]

Baerger, D. & McAdams, D. (1999) Life story coherence and its relation to psychological well-being. *Narrative Inquiry* 9:69–96. [JBH]

Bar, M. (2007) The proactive brain: Using analogies and associations to generate predictions. *Trends in Cognitive Sciences* 11(7):280–89. [aAC]

Barlow, H. B. (1961) Possible principles underlying the transformations of sensory messages. In: *Sensory communication*, ed. W. Rosenblith, pp. 217–34. (Chapter 13). MIT Press. [KF, PK, TT]

Barrett, L. F. (2009) The future of psychology: Connecting mind to brain. *Perspectives in Psychological Science* 4:326–39. [aAC]

Barrett, L. F. & Bar, M. (2009) See it with feeling: Affective predictions during object perception. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 364(1521):1325–34. [arAC, AKS]

Baugh, L. A., Kao, M., Johansson, R. S. & Flanagan, J. R. (2012) Material evidence: Interaction of well-learned priors and sensorimotor memory when lifting objects. *Journal of Neurophysiology* 108(5):1262–69. doi:10.1152/jn.00263.2012. [GB]

Bechtel, W. (2006) Reducing psychology while maintaining its autonomy via mechanistic explanation. In: *The matter of the mind: Philosophical essays on psychology, neuroscience and reduction*, ed. M. Schouten & H. L. de Jong, Ch. 8. Blackwell. [MWS]

Beer, R. D. (2000) Dynamical approaches to cognitive science. *Trends in Cognitive Sciences* 4(3):91–99. [TF]

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

Berkes, P. & Wiskott, L. (2005) Slow feature analysis yields a rich repertoire of complex cell properties. *Journal of Vision* 5(6):579–602. [PK]

Berlyne, D. E. (1970) Novelty, complexity and hedonic value. *Perception and Psychophysics* 8:279–86. [RSS]

Berniker, M. & Körding, K. P. (2008) Estimating the sources of motor errors for adaptation and generalization *Nature Neuroscience* 11:1454–61. [aAC]

Betsch, B. Y., Einhäuser, W., Körding, K. P. & König, P. (2004) The world from a cat's perspective – statistics of natural videos. *Biological Cybernetics* 90:41–50. [PK]

Bindra, D. (1959) Stimulus change, reactions to novelty, and response decrement. *Psychological Review* 66:96–103. [aAC]

Blake, R. (2001) A primer on binocular rivalry, including current controversies. *Brain and Mind* 2:5–38. [MLA]

Blakemore, S., Oakley, D. & Frith, C. D. (2003) Delusions of alien control in the normal brain. *Neuropsychologia* 41(8):1058–67. [PG]

Born, R. T., Tsui, J. M. & Pack, C. C. (2009) Temporal dynamics of motion integration, In: *Dynamics of visual motion processing*, ed. U. Ilg & G. Masson. pp. 37–54. Springer. [aAC]

Bourdieu, P. (1977) *Outline of a theory of practice*, trans. R. Nice. Cambridge University Press. [AR]

Brainard, D. (2009) Bayesian approaches to color vision. In: *The visual neurosciences*, 4th edition, ed. M. Gazzaniga, pp. 395–408. MIT Press. [aAC]

Brayanov, J. B. & Smith, M. A. (2010) Bayesian and "anti-Bayesian" biases in sensory integration for action and perception in the size–weight illusion. *Journal of Neurophysiology* 103(3):1518–31. [GB, rAC]

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

Brown, M., Dilley, L. C. & Tanenhaus, M. K. (2012) Real-time expectations based on context speech rate can cause words to appear or disappear. In: *Proceedings of the 34th Annual Conference of the Cognitive Science Society,* ed. N. Miyake, D. Peebles, & R. P. Cooper, pp. 1374–79. Cognitive Science Society. [TAF]

Brown, M., Salverda, A. P., Dilley, L. C. & Tanenhaus, M. K. (2011) Expectations from preceding prosody influence segmentation in online sentence processing. *Psychonomic Bulletin and Review* 18:1189–96. [TAF]

Brown, R. G. & Hwang, P. Y. C. (1992) *Introduction to random signals and applied Kalman filtering*, 2nd edition. Wiley. [DRa]

Brown, S. (2003) Biomusicology and the three paradoxes about music. *Bulletin of Psychology and the Arts* 4:14–17. [LH]

Bruner, J. (1986) *Actual minds, possible worlds*. Harvard University Press. [JBH]

Bruner, J. (1991) The narrative construction of reality. *Critical Inquiry* 18(1):1–21. [JBH]

Bubic, A., von Cramon, D. Y. & Schubotz, R. I. (2010) Prediction, cognition and the brain. *Frontiers in Human Neuroscience* 4(25):1–15. [aAC]

Buckingham, G. & Goodale, M. A. (2010) Lifting without seeing: The role of vision in perceiving and acting upon the size weight illusion. *PLoS ONE* 5(3):e9709. doi:10.1371/journal.pone.0009709. [GB]

Buhusi, C. V. & Meck, W. H. (2005) What makes us tick? Functional and neural mechanisms of interval timing. *Nature Reviews: Neuroscience* 6:755–65. [LH]

Burge, J., Fowlkes, C. & Banks, M. (2010) Natural-scene statistics predict how the figure–ground cue of convexity affects human depth perception. *Journal of Neuroscience* 30(21):7269–80. [aAC]

Burr, D., Tozzi, A. & Morrone, C. (2007) Neural mechanisms for timing visual events are spatially selective in real-world coordinates. *Nature Neuroscience* 10:423–25. [LH]

Carandini, M. (2012) From circuits to behavior: A bridge too far? *Nature Neuroscience* 15(4):507–509. [MWS]

Carandini, M., Demb, J. B., Mante, V., Tolhurst, D. J., Dan, Y., Olshausen, B. A., Gallant, J. L. & Rust, N. C. (2005) Do we know what the early visual system does? *Journal of Neuroscience* 25(46):10577–97. [MWS]

Carrasco, M. (2011) Visual attention: The past 25 years. *Vision Research* 51:1484–525. [NB]

Carrasco, M., Ling, S. & Read, S. (2004) Attention alters appearance. *Nature Neuroscience* 7:308–13. [NB, rAC]

Chappell, J. & Sloman, A. (2007) Natural and artificial meta-configured altricial information-processing systems. *International Journal of Unconventional Computing* 3(3):211–39. Available at: http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0609. [AS]

Chater, N. & Manning, C. (2006) Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences* 10(7):335–44. [aAC]

Chemero, A. (2009) *Radical embodied cognitive science*. MIT Press. [MLA]

Chen, L. (2005) The topological approach to perceptual organization. *Visual Cognition* 12:553–637. [SMS]

Chennu, S., Craston, P., Wyble, B. & Bowman, H. (2009) Attention increases the temporal precision of conscious perception: Verifying the neural ST^2 model. *PLOS Computational Biology* 5(11):1–13. [HB]

Chittka, L. & Skorupski, P. (2011) Information processing in miniature brains. *Proceedings of the Royal Society of London, B: Biological Sciences* 278 (1707):885–88. doi: 10.1098/rspb.2010.2699. [AS]

Churchland, P. M. (1989) *The neurocomputational perspective*. MIT/Bradford Books. [arAC]

Churchland, P. M. (2012) *Plato's camera: How the physical brain captures a landscape of abstract universals*. MIT Press. [arAC]

Clark, A. (1987) The kludge in the machine. *Mind and Language* 2(4):277–300. [aAC]

Clark, A. (1989) *Microcognition: Philosophy, cognitive science and parallel distributed processing*. MIT Press/Bradford Books. [arAC]

Clark, A. (1993) Minimal rationalism. *Mind* 102(408):587–610. [rAC]

Clark, A. (1997) *Being there: Putting brain, body and world together again*. MIT Press. [MLA, aAC]

Clark, A. (2006a) Language, embodiment and the cognitive niche. *Trends in Cognitive Sciences* 10(8):370–74. [arAC]

Clark, A. (2006b) Material symbols. *Philosophical Psychology* 19(3):291–307. [AR]

Clark, A (2008) *Supersizing the mind: Action, embodiment, and cognitive extension*. Oxford University Press. [arAC]

Clark, A. (2012) Dreaming the whole cat: Generative models, predictive processing, and the enactivist conception of perceptual experience. *Mind* 121(483):753–71. [rAC]

Clark, A. (forthcoming) Perceiving as predicting, In: *Perception and its modalities*, ed. M. Mohan, S. Biggs & D. Stokes. Oxford University Press. [arAC]

Clark, A. & Chalmers, D. (1998) The extended mind. *Analysis* 58(1):7–19. [aAC]

Clark, A. & Thornton, C. (1997) Trading spaces: Computation, representation, and the limits of uninformed learning. *Behavioral and Brain Sciences* 20(1):57–66. [rAC]

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* 47:3125–31. [aAC]

Coltheart, M. (2007) Cognitive neuropsychiatry and delusional belief (The 33rd Sir Frederick Bartlett Lecture). *The Quarterly Journal of Experimental Psychology* 60(8):1041–62. [aAC]

Conrad, K. (1958) *Die beginnende Schizophrenie*. Thieme Verlag. [SMS]

Corlett, P. R., Frith, C. D. & Fletcher, P. C. (2009a) From drugs to deprivation: A Bayesian framework for understanding models of psychosis. *Psychopharmacology (Berlin)* 206(4):515–30. [aAC]

Corlett, P. R., Krystal, J. K., Taylor, J. R. & Fletcher, P. C. (2009b) Why do delusions persist? *Frontiers in Human Neuroscience* 3:12. doi: 10.3389/neuro.09.012.2009. [aAC]

Corlett, P. R., Taylor, J. R., Wang, X. J., Fletcher, P. C. & Krystal, J. H. (2010) Toward a neurobiology of delusions. *Progress in Neurobiology* 92(3):345–69. [aAC]

Craig, A. D. (2003) Interoception: The sense of the physiological condition of the body. *Current Opinion in Neurobiology* 13(4):500–505. [AKS]

Craig, A. D. (2009) How do you feel – now? The anterior insula and human awareness. *Nature Reviews Neuroscience* 10(1):59–70. [AKS]

Craik, K. (1943) *The nature of explanation*. Cambridge University Press. [DRa]

Critchley, H. D. & Seth, A. K. (2012) Will studies of macaque insula reveal the neural mechanisms of self-awareness? *Neuron* 74(3):423–26. [AKS]

Critchley, H. D., Wiens, S., Rotshtein, P., Ohman, A. & Dolan, R. J. (2004) Neural systems supporting interoceptive awareness. *Nature Neuroscience* 7(2):189–95. [AKS]

Crutchfield, J. P. & Young, K. (1989) Inferring statistical complexity. *Physical Review Letters* 63:105–108. [DYL]

Dahan, D. & Tanenhaus, M. K. (2004) Continuous mapping from sound to meaning in spoken-language comprehension: Evidence from immediate effects of verb-based constraints. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 30:498–513. [TAF]

Damasio, A. (2000) *The feeling of what happens: Body and emotion in the making of consciousness*. Harvest Books. [AKS]

Danckert, J., Saoud, M. & Maruff, P. (2004) Attention, motor control and motor imagery in schizophrenia: implications for the role of the parietal cortex. *Schizophrenia Research* 70(2–3):241–61. [PG]

Daprati, E., Franck, N., Georgieff, N., Proust, J., Pacherie, E., Dalery, J. & Jeanerod, M. (1997) Looking for the agent: An investigation into consciousness of action and self-consciousness in schizophreniz patients. *Cognition* 65:71–86. [PG]

Darwin, C. (1871) *The descent of man and selection in relation to sex*. John Murray. [PAG]

Davidson, D. (1974) On the very idea of a conceptual scheme. *Proceedings and Addresses of the American Philosophical Association* 47:5–20. [MLA]

Dayan, P. (1997) Recognition in hierarchical models. In: *Foundations of computational mathematics*, ed. F. Cucker & M. Shub, pp. 43–57. Springer. [aAC]

Dayan, P. & Hinton, G. (1996) Varieties of Helmholtz machine. *Neural Networks* 9:1385–403. [aAC]

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

de Gardelle, V., Waszczuk, M., Egner, T. & Summerfield, C. (2012) Concurrent repetition enhancement and suppression responses in extrastriate visual cortex. *Cerebral Cortex*. [Epub ahead of print: July 18, 2012]. doi: 10.1093/cercor/bhs211. [LM]

Dehaene, S. (2009) *Reading in the brain*. Penguin. [aAC]

Demos, A. P., Chaffin, R., Begosh, K. T., Daniels, J. R. & Marsh, K. L. (2012) Rocking to the beat: Effects of music and partner's movements on spontaneous interpersonal coordination. *Journal of Experimental Psychology: General* 141:49–53. [LH]

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* 39:1–38. [aAC]

Deneve, S. (2008) Bayesian spiking neurons I: Inference. *Neural Computation* 20:91–117. [aAC]

Dennett, D. (1978) *Brainstorms: Philosophical essays on mind and psychology*. Bradford Books/MIT Press. [aAC]

Dennett, D. C. (1987) *The intentional stance*. MIT Press. [aAC]

Dennett, D. C. (1991) *Consciousness explained*. Little, Brown. [aAC]

Dennett, D. C. (2009) Darwin's "Strange Inversion of Reasoning". *Proceedings of the National Academy of Sciences USA* 106 (Suppl. 1):10061–65. [rAC, DCD]

den Ouden, H. E. M., Daunizeau, J., Roiser, J., Friston, K. J. & Stephan, K. E. (2010) Striatal prediction error modulates cortical coupling. *Journal of Neuroscience* 30:3210–19. [arAC, TE]

den Ouden, H. E. M, Friston, K. J., Daw, N. D., McIntosh, A. R. & Stephan, K. E. (2009) A dual role for prediction error in associative learning. *Cerebral Cortex* 19:1175–85. [rAC, TE]

Desimone, R. & Duncan, J. (1995) Neural mechanisms of selective visual attention. *Annual Review of Neuroscience* 18:193–222. [aAC]

de-Wit, L. H., Kubilius J., Wagemans J. & Op de Beeck, H. P. (2012) Bistable Gestalts reduce activity in the whole of V1, not just the retinotopically predicted parts. *Journal of Vision* 12:1–14. [LM]

de-Wit, L., Machilsen, B. & Putzeys, T. (2010) Predictive coding and the neural response to predictable stimuli. *Journal of Neuroscience* 30:8702–703. [aAC]

Dikker, S., Rabagliati, H., Farmer, T. A. & Pylkkanen, L. (2010) Early occipital sensitivity to syntactic category is based on form typicality. *Psychological Science* 21:629–34. [TAF]

Dilley, L. C. & McAuley, J. D. (2008) Distal prosodic context affects word segmentation and lexical processing. *Journal of Memory and Language* 59:294–311. [TAF]

Dilley, L. C. & Pitt, M. (2010) Altering context speech rate can cause words to appear or disappear. *Psychological Science* 21:1664–70. [TAF]

Dima, D., Dietrich, D. E., Dillo, W. & Emrich, H. M. (2010) Impaired top-down processes in schizophrenia: A DCM study of ERPs. *NeuroImage* 52:824–32. [SMS]

Dima, D., Roiser, J. P., Dietrich, D. E., Bonnemann, C., Lanfermann, H., Emrich, H. M. & Dillo, W. (2009) Understanding why patients with schizophrenia do not perceive the hollow-mask illusion using dynamic causal modeling. *NeuroImage* 46:1180–86. [SMS]

Di Paolo, E. A. (2009) Extended life. *Topoi* 28(1):9–21. [aAC, TF]

Di Paolo, E. A., Rohde, M. & De Jaegher, H. (2010) Horizons for the enactive mind: Values, social interaction, and play. In: *Enaction: Toward a new paradigm for cognitive science*, ed. J. Stewart, O. Gapenne & E. A. Di Paolo, pp. 33–87. MIT Press. [TF]

Doherty, M. J., Campbell, N. M., Tsuji, H. & Phillips, W. A. (2010) The Ebbinghaus illusion deceives adults but not young children. *Developmental Science* 13:714–21. doi:10.1111/j.1467-7687.2009.00931.x. [WAP]

Doherty, M. J., Tsuji, H. & Phillips, W. A. (2008) The context-sensitivity of visual size perception varies across cultures. *Perception* 37:1426–33. [WAP]

Doya, K., Ishii, S., Pouget, A. & Rao, R. eds. (2007) *Bayesian brain: Probabilistic approaches to neural coding*. MIT Press. [aAC]

Dumoulin, S. O. & Hess, R. F. (2006) Modulation of V1 activity by shape: image-statistics or shape-based perception? *Journal of Neurophysiology* 95:3654–64. [aAC]

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):16601–608. [arAC, TE]

Einhäuser, W., Kayser, C., König, P. & Körding, K. P. (2002) Learning the invariance properties of complex cells from their responses to natural stimuli. *European Journal of Neuroscience* 15:475–86. [PK]

Einhäuser, W., Moeller, G. U., Schumann, F., Conradt, J., Vockeroth, J., Bartl, K., Schneider, E. & König, P. (2009) Eye-head coordination during free exploration in human and cat. *Annals of the New York Academy of Sciences* 1164:353–66. [PK]

Eliades, S. J. & Wang, X. (2008) Neural substrates of vocalization feedback monitoring in primate auditory cortex. *Nature* 453:1102–106. [rAC, TE]

Eliasmith, C. (2007) How to build a brain: From function to implementation. *Synthese* 159(3):373–88. [aAC, NS]

Eliasmith, C. (in press) *How to build a brain: A neural architecture for biological cognition*. Oxford University Press. [DRa]

Eliasmith, C. & Anderson, C. (2003) *Neural engineering: Computation, representation, and dynamics in neurobiological systems*. MIT Press. [DRa]

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y. & Rasmussen, D. (2012) A large-scale model of the functioning brain. *Science* 338 (6111):1202–205. [DRa]

Engel, A. K., Fries, P. & Singer, W. (2001) Dynamic predictions: Oscillations and synchrony in top–down processing. *Nature Reviews: Neuroscience* 2:704–16. [aAC]

Erlhagen, W. (2003) Internal models for visual perception. *Biological Cybernetics* 88:409–17. [LM]

Ernst, M. O. (2010) Eye movements: Illusions in slow motion. *Current Biology* 20 (8):R357–59. [aAC]

Ernst, M. O. & Banks, M. S. (2002) Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 415:429–33. [aAC]

Everitt, B., Dickinson, A. & Robbins, T. (2001) The neuropsychological basis of addictive behavior. *Brain Research Reviews* 36:129–38. [DRo]

Evrard, H. C., Forro, T. & Logothetis, N. K. (2012) Von Economo neurons in the anterior insula of the macaque monkey. *Neuron* 74(3):482–89. [AKS]

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

Farmer, T. A., Christiansen, M. H. & Monaghan, P. (2006) Phonological typicality influences on-line sentence comprehension. *Proceedings of the National Academy of Sciences USA* 103:12203–208. [TAF]

Farmer, T. A., Monaghan, P., Misyak, J. B. & Christiansen, M. H. (2011) Phonological typicality influences sentence processing in predictive contexts: A reply to Staub et al. (2009) *Journal of Experimental Psychology: Learning, Memory, and Cognition* 37:1318–25. [TAF]

Feldman, H. & Friston, K. J. (2010) Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience* 4:215. doi:10.3389/fnmuh.2010.00215. [HB, arAC, TE, WAP, MWS]

Feldman, J. (2010) Cognitive science should be unified: Comment on Griffiths et al. and McClelland et al. *Trends in Cognitive Sciences* 14(8):341. [aAC]

ffytche, D. H. & Howard, R. J. (1999) The perceptual consequences of visual loss: "Positive" pathologies of vision. *Brain* 122:1247–60. [SMS]

Fine, A. B., Jaeger, T. F., Farmer, T. A. & Qian, T. (under review) Rapid expectation adaptation during syntactic comprehension. [TAF]

Fiorillo, C. D. (2012) Beyond Bayes: On the need for a unified and Jaynesian definition of probability and information within neuroscience. *Information* 3 (2):175–203. doi:10.3390/info3020175. [WAP]

Flanagan, J. R. & Beltzner, M. A. (2000) Independence of perceptual and sensorimotor predictions in the size-weight illusion. *Nature Neuroscience* 3(7):737–41. doi:10.1038/76701. [GB]

Flanagan, J. R., Bittner, J. P. & Johansson, R. S. (2008) Experience can change distinct size-weight priors engaged in lifting objects and judging their weights. *Current Biology: CB* 18(22):1742–47. doi:10.1016/j.cub.2008.09.042. [GB]

Fletcher, P. & Frith, C. (2009) Perceiving is believing: A Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews: Neuroscience* 10:48–58. [aAC]

Földiák, P. (1990) Forming sparse representations by local anti-Hebbian learning. *Biological Cybernetics* 64:165–70. [TT]

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* 20:757–62. [aAC]

Friston K. (2002) Beyond phrenology: What can neuroimaging tell us about distributed circuitry? *Annual Review of Neuroscience* 25:221–50. [aAC]

Friston, K. (2003) Learning and inference in the brain. *Neural Networks* 16(9):1325–52. [aAC]

Friston, K. (2005) A theory of cortical responses. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 360(1456):815–36. [arAC, TE, MWS]

Friston, K. (2008) Hierarchical models in the brain. *PLoS Computational Biology* 4: e1000211. [TE]

Friston, K. (2009) The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences* 13(7):293–301. [aAC, TF, AR]

Friston, K. J. (2010) The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience* 11(2):127–38. [aAC, TE, KF, TF, PK, WAP, TT]

Friston, K. (2011a) Embodied inference: Or I think therefore I am, if I am what I think. In: *The implications of embodiment* (*Cognition and Communication*), ed. W. Tschacher & C. Bergomi, pp. 89–125. Imprint Academic. [arAC, DRo]

Friston, K. (2011b) What is optimal about motor control? *Neuron* 72:488–98. [arAC]

Friston, K., Adams, R. A., Perrinet, L. & Breakspear, M. (2012) Perceptions as hypotheses: Saccades as experiments. *Frontiers in Psychology* 3:151. doi: 10.3389/fpsyg.2012.00151. [rAC]

Friston, K. J., Daunizeau, J. & Kiebel, S. J. (2009) Reinforcement learning or active inference? *PLoS (Public Library of Science) One* 4(7):e6421. [aAC]

Friston, K. J., Daunizeau, J., Kilner, J. & Kiebel, S. J. (2010) Action and behavior: A free-energy formulation. *Biological Cybernetics* 102(3):227–60. [aAC]

Friston, K. & Kiebel, S. (2009) Cortical circuits for perceptual inference. *Neural Networks* 22:1093–104. [arAC]

Friston, K., Mattout, J. & Kilner, J. (2011) Action understanding and active inference. *Biological Cybernetics* 104:137–60. [aAC]

Friston, K. & Stephan, K. (2007) Free energy and the brain. *Synthese* 159(3):417–58. [aAC, TF, DYL]

Frith, C. D. (2007) *Making up the mind: How the brain creates our mental world*. Blackwell. [BP]

Frith, C. D. (2012) Explaining delusions of control: The comparator model 20 years on. *Consciousness and Cognition* 21(1):52–54. [AKS]

Frith, C. D. & Wentzer, T. S. (in press) Neural hermeneutics. In: *Encyclopedia of philosophy and the social sciences, vol. 1*, ed. B. Kaldis. Sage. [rAC, BP]

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

Froese, T. & Di Paolo, E. A. (2011) The enactive approach: Theoretical sketches from cell to society. *Pragmatics and Cognition* 19(1):1–36. [TF]

Froese, T. & Stewart, J. (2010) Life after Ashby: Ultrastability and the autopoietic foundations of biological individuality. *Cybernetics and Human Knowing* 17 (4):83–106. [TF]

Froese, T. & Ziemke, T. (2009) Enactive artificial intelligence: Investigating the systemic organization of life and mind. *Artificial Intelligence* 173(3–4):366–500. [TF]

Fuster, J. M. (2001) The prefrontal cortex – an update: time is of the essence. *Neuron* 30:319–33. [KF]

Gallagher, S. (2004) Neurocognitive models of schizophrenia: a neurophenomenological critique. *Psychopathology* 37(1):8–19. [PG]

Galton, F. (1883) *Inquires into human faculty and its development*. MacMillan. [LH]

Geertz, C. (1966) Religion as a cultural system. In: *The interpretation of cultures*, pp. 87–125. Basic Books. [AR]

Geisler, W. S. & Kersten, D. (2002) Illusions, perception and Bayes. *Nature Neuroscience* 5(6):508–10. doi:10.1038/nn0602-508. [GB]

Geissler, H.-G. (1983) The inferential basis of classification: From perceptual to memory code systems. Part 1: Theory. In: *Modern issues in perception*, ed. H.-G. Geissler, H. Buffart, E. Leeuwenberg & V. Sarris, pp. 87–105. North-Holland. [aAC]

Geissler, H.-G. (1991) Constraints of mental self-organization: The indirect validation approach toward perception. *Estratto da Comunicazioni Scientifiche di Psicologia Generale* 5:47–69. [aAC]

Geldmacher, D. S. (2003) Visuospatial dysfunction in the neurodegenerative diseases. *Frontiers in Bioscience* 8:e428–36. [SMS]

Gerrans, P. (2007) Mechanisms of madness. Evolutionary psychiatry without evolutionary psychology. *Biology and Philosophy* 22:35–56. [aAC]

Gershman, S. J. & Daw, N. D. (2012) Perception, action and utility: The tangled skein. In: *Principles of brain dynamics: Global state interactions*, ed. M. I. Rabinovich, K. J. Friston & P. Varona, pp. 293–312. MIT Press. [aAC, TF]

Gibson, J. J. (1966) *The senses considered as perceptual systems*. Houghton Mifflin. [AS]

Gibson, J. J. (1979) *The ecological approach to visual perception*. Houghton Mifflin. [DCD]

Gilbert, D. T. & Wilson, T. D. (2009) Why the brain talks to itself: Sources of error in emotional prediction. *Philosophical Transactions of the Royal Society of London B Biological Science* 364(1521):1335–41. [AKS]

Glimcher, P. (2003) *Decisions, uncertainty and the brain*. MIT Press. [DRo]

Glimcher, P. (2010) *Foundations of neuroeconomic analysis*. Oxford University Press. [rAC, DRo]

Gold, J. N. & Shadlen, M. N. (2001) Neural computations that underlie decisions about sensory stimuli. *Trends in Cognitive Sciences* 5(10):16 238–55. [aAC]

Gooch, C. M., Wiener, M., Hamilton, C. A. & Coslett, B. H. (2001) Temporal discrimination of sub- and suprasecond time intervals: A voxel-based lesion mapping analysis. *Frontiers in Integrative Neuroscience* 5:1–10. [LH]

Gowaty, P. A. & Hubbell, S. P. (2009) Reproductive decisions under ecological constraints: It's about time. *Proceedings of the National Academy of Sciences USA* 106:10017–24. [PAG]

Gowaty, P. A. & Hubbell, S. P. (2005) Chance, time allocation, and the evolution of adaptively flexible sex role behavior. *Integrative and Comparative Biology* 45 (5):931–44. [PAG]

Graesser, A. C., Millis, K. K. & Zwaan, R. A. (1997) Discourse comprehension. *Annual Review of Psychology* 48(1):163–89. [JBH]

Grahn J. A. & Brett, M. (2007) Rhythm and beat perception in motor areas of the brain. *Journal of Cognitive Neuroscience* 19(5):893–906. [RSS]

Grahn, J. A. & McAuley, J. D. (2009) Neural bases of individual differences in beat perception. *NeuroImage* 47:1894–1903. [LH]

Gregory, R. (1998) Brainy mind. *British Medical Journal* 317(7174):1693–95. [GB]

Gregory, R. L. (1980) Perceptions as hypotheses. *Philosophical Transactions of the Royal Society of London B* 290(1038):181–97. [aAC, KF]

Griffiths, T., Chater, N., Kemp, C., Perfors, A. & Tenenbaum, J. B. (2010) Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences* 14(8):357–64. [aAC]

Griffiths, P. E. & Gray, R. D. (2001) Darwinism and developmental systems. In: *Cycles of contingency: Developmental systems and evolution*, eds. S. Oyama, P. E. Griffiths & R. D. Gray, pp. 195–218. MIT Press. [aAC]

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

Grodner, D. & Sedivy, J. (2011) The effect of speaker-specific information on pragmatic inferences. In: *The processing and acquisition of reference, vol. 2327*, eds. E. Gibson & N. Pearlmutter, pp. 239–72. MIT Press. [TAF]

Grossberg, S. (2013) Adaptive Resonance Theory: How a brain learns to consciously attend, learn, and recognize a changing world. *Neural Networks* 37:1–47. [LM]

Grush, R. (2004) The emulation theory of representation: Motor control, imagery, and perception. *Behavioral and Brain Sciences* 27:377–442. [aAC]

Hajcak, G. & Foti, D. (2008) Errors are aversive. *Psychological Science* 19(2):103–108. [JBH]

Harman, K., Humphrey K. G. & Goodale, M. A. (1999) Active manual control of object views facilitates visual recognition. *Current Biology* 9:1315–18. [LH]

Harnad, S. (1990) The symbol grounding problem. *Physica D* 42:335–46. [aAC]

Harrison, L. M., Bestmann, S., Rosa, M. J., Penny, W. & Green, G. G. R. (2011) Time scales of representation in the human brain: Weighing past information to predict future events. *Frontiers in Human Neuroscience* 5:1–8. [LH]

Haugeland, J. (1998) Mind embodied and embedded. In: *Having thought: Essays in the metaphysics of mind*, ed. J. Haugeland, pp. 207–40. Harvard University Press. [aAC]

Hawkins, J. & Blakeslee, S. (2004) *On intelligence*. Owl Books/Times Books. [aAC, TE]

Hay, J. & Drager, K. (2010) Stuffed toys and speech perception. *Linguistics* 48:865–92. [TAF]

Helbig, H. & Ernst, M. (2007) Optimal integration of shape information from vision and touch. *Experimental Brain Research* 179:595–605. [aAC]

Helmholtz, H. von (1860/1962) *Handbuch der physiologischen optik, vol. 3*, ed. & trans. J. P. C. Southall. Dover. (Original work published in 1860; Dover English edition in 1962). [aAC]

Helmholtz, H. von (1876) *Handbuch der physiologischen Optik*. Leopold Voss. [TE]

Hennig, H., Fleischmann, R., Fredebohm, A., Hagmayer, Y., Nagler, J., Witt, A., Theis, F. J. & Geisel, T. (2011) The nature and perception of fluctuations in human musical rhythms. *PLoS ONE* 6(10):e26457. [RSS]

Hesselmann, G., Kell, C. A. & Kleinschmidt, A. (2010) Predictive coding or evidence accumulation? False inference and neuronal fluctuations *PloS One* 5(3):9926 [LM]

Hesselmann, G., Sadaghiani, S., Friston, K. J. & Kleinschmidt, A. (2010) Predictive coding or evidence accumulation? False inference and neuronal fluctuations. *PloS (Public Library of Science) One* 5(3):e9926. [aAC]

Hinton, G. E. (2002) Training products of experts by minimizing contrastive divergence. *Neural Computation* 14(8):1711–800. [aAC]

Hinton, G. E. (2007a) Learning multiple layers of representation. *Trends in Cognitive Sciences* 11:428–34. [aAC]

Hinton, G. E. (2007b) To recognize shapes, first learn to generate images. In: *Computational neuroscience: Theoretical insights into brain function*, eds. P. Cisek, T. Drew & J. Kalaska. Elsevier. [aAC]

Hinton, G. E. (2010) Learning to represent visual input. *Philosophical Transactions of the Royal Society, B*. 365:177–84. [aAC]

Hinton, G. E., Dayan, P., Frey, B. J. & Neal, R. M. (1995) The wake-sleep algorithm for unsupervised neural networks. *Science* 268:1158–60. [aAC]

Hinton, G. E. & Ghahramani, Z. (1997) Generative models for discovering sparse distributed representations. *Philosophical Transactions of the Royal Society B* 352:1177–90. [aAC]

Hinton, G. E. & Nair, V. (2006) Inferring motor programs from images of handwritten digits. In: *Advances in neural information processing systems 18: Proceedings of the 2005 NIPS Conference*, ed. Y. Weiss, B. Scholkopf & J. Platt, pp. 515–22. MIT Press. [rAC]

Hinton, G. E., Osindero, S. & Teh, Y. (2006) A fast learning algorithm for deep belief nets. *Neural Computation* 18:1527–54. [aAC]

Hinton, G. E. & Salakhutdinov, R. R. (2006) Reducing the dimensionality of data with neural networks. *Science* 313(5786):504–507. [aAC]

Hinton, G. E. & van Camp, D. (1993) Keeping neural networks simple by minimizing the description length of weights. In: *Proceedings of COLT-93 (Sixth Annual Conference on Computational Learning Theory, Santa Cruz, CA, July 26–28, 1993)*, ed. L. Pitt, pp. 5–13. ACM Digital Library. [aAC]

Hinton, G. E. & Zemel, R. S. (1994) Autoencoders, minimum description length and Helmholtz free energy. In: *Advances in neural information processing systems 6*, eds. J. Cowan, G. Tesauro & J. Alspector. Morgan Kaufmann. [aAC]

Hirsch, H. V. B. & Spinelli, D. (1970) Visual experience modifies distribution of horizontally and vertically oriented receptive fields in cats. *Science* 168:869–71. [BB]

Hirsh, J. B., Mar, R. A. & Peterson, J. B. (2012) Psychological entropy: A framework for understanding uncertainty-related anxiety. *Psychological Review* 119 (2):304–20. [JBH]

Hochstein, S. & Ahissar, M. (2002) View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron* 36(5):791–804. [aAC]

Hogg, D. (1983) Model-based vision: A program to see a walking person. *Image and Vision Computing* 1(1):5–20. [AS]

Hohwy, J. (2007) Functional Integration and the mind. *Synthese* 159(3):315–28. [aAC]

Hohwy, J. (2012) Attention and conscious perception in the hypothesis testing brain. *Frontiers in Psychology* 3:96, 1–14. doi: 10.3389/fpsyg.2012.00096. [rAC, LM]

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

Hohwy, J., Roepstorff, A. & Friston, K. (2008) Predictive coding explains binocular rivalry: An epistemological review. *Cognition* 108(3):687–701. [aAC, MLA]

Hollensen, P. & Trappenberg, T. (2011) Learning sparse representations through learned inhibition. Poster presented at the COSYNE (Computational and Systems Neuroscience Conference) Annual Meeting, Salt Lake City, Utah, February 24, 2011. [TT]

Holleman, J. R. & Schultz, W. (1998) Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Reviews: Neuroscience* 1:304–309. [aAC]

Holm, L., Ullén, F. & Madison, G. (in press) Motor and executive control in repetitive timing of brief intervals. *Journal of Experimental Psychology: Human Perception and Performance*. doi 10.1037/a0029142. [LH]

Horstmann, G. (2002) Evidence for attentional capture by a surprising color singleton in visual search. *Psychological Science* 13(6):499–505. [HB]

Hosoya, T., Baccus, S. A. & Meister, M. (2005) Dynamic predictive coding by the retina. *Nature* 436(7):71–77. [aAC]

Howe, C. Q., Lotto, R. B. & Purves, D. (2006) Comparison of bayesian and empirical ranking approaches to visual perception. *Journal of Theoretical Biology* 241:866–75. [aAC]

Huang, Y. & Rao, R. (2011) Predictive coding. *Wiley Interdisciplinary Reviews: Cognitive Science* 2:580–93. [aAC]

Hubbell, S. P. & Johnson, L. K. (1987) Environmental variance in lifetime mating success, mate choice, and sexual selection. *American Naturalist* 130(1):91–112. [PAG]

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* 28:229–89. [TE]

Hume, D. (1739/1888/1964) *Treatise of human nature*, ed. L. A. Selby-Biggs. Oxford University Press. (Original work published in 1739; OUP edition 1888; reprint 1964 source cited.) [DCD]

Humphrey, N. (2000) How to solve the mind-body problem. *Journal of Consciousness Studies* 7:5–20. [aAC]

Hurley, M., Dennett, D. C. & Adams, R. B., Jr. (2011) *Inside jokes: Using humor to reverse-engineer the mind*. MIT Press. [DCD]

Hurley, S. (1998) *Consciousness in action*. Harvard University Press. [aAC, NS]

Huron, D. (2006) *Sweet anticipation: Music and the psychology of expectation.* MIT Press. [RSS]

Hutchins, E. (1995) *Cognition in the wild*. MIT Press. [arAC]

Ikegami, T. (2007) Simulating active perception and mental imagery with embodied chaotic itinerancy. *Journal of Consciousness Studies* 14(7):111–25. [TF]

Iriki, A. & Taoka, M. (2012) Triadic (ecological, neural, cognitive) niche construction: A scenario of human brain evolution extrapolating tool use and language from the control of reaching actions. *Philosophical Transactions of the Royal Society B* 367:10–23. [aAC]

Jaeger, H. (2011) Neural hierarchies: Singin' the blues. Oral presentation at Osnabrück Computational Cognition Alliance Meeting (OCCAM 2011), University of Osnabrück, Germany, June 22–24, 2011. Available at: http://video.virtuos.uni-osnabrueck.de:8080/engage/ui/watch.html?id=10bc55e8-8d98-40d3-bb11-17780b70c052&play=true. [TT]

Jahanshahi, M., Dirnberger, G., Fuller, R. & Frith, C. D. (2000) The role of the dorsolateral prefrontal cortex in random number generation: A study with positron emission tomography. *NeuroImage* 12:713–25. [LH]

James, W. (1890) *The principles of psychology*. Henry Holt. [AKS]

Janoff-Bulman, R. (1992) *Shattered assumptions: Towards a new psychology of trauma*. Free Press. [JBH]

Jaynes, E. T. (1957) Information theory and statistical mechanics. *Physical Review (Series II)* 106(4):620–30. [KF]

Jaynes, E. T. (2003) *Probability theory: The logic of science.* Cambridge University Press. [WAP]

Jeannerod, M. (2006) *Motor cognition: What actions tell the self*. Oxford University Press. [PG]

Jeannerod, M., Farrer, C.,Franck, N., Fourneret, P., Posada, A., Daprati, E. & Georgieff, N. (2003) Action recognition in normal and schizophrenic subjects. In: *The self in neuroscience and psychiatry*, ed. N. Kirchner & A. David, pp. 380–406. Cambridge University Press. [PG]

Jehee, J. F. M. & Ballard, D. H. (2009) Predictive feedback can account for biphasic responses in the lateral geniculate nucleus. *PLoS (Public Library of Science) Computational Biology* 5(5):e1000373. [aAC]

Jiang, J., Schmajuk, N. & Egner, T. (2012) Explaining neural signals in human visual cortex with an associative learning model. *Behavioral Neuroscience* 126(4):575–81. [TE]

Johnston, A., Arnold, D. H. & Nishida, S. (2006) Spatially localized distortions of time. *Current Biology* 16:472–79. [LH]

Kalman, R. E. (1960) A new approach to linear filtering and prediction problems. *Transactions of the ASME – Journal of Basic Engineering (Series D)* 82:35–45. [DRa]

Kant, I. (1781/1929) *Critique of pure reason,* trans. N. Kemp Smith. Macmillan. (Original work published in 1781; Kemp Smith translation 1929). [AS]

Kärcher, S. M., Fenzlaff, S., Hartmann, D., Nagel, S. K. & König, P. (2012) Sensory augmentation for the blind. *Frontiers in Human Neuroscience* 6:37. [PK]

Karmarkar, U. R & Buonomano, D. V. (2007) Timing in the absence of clocks: Encoding time in neural network states. *Neuron* 53:427–38. [LH]

Karmiloff-Smith, A. (1992) *Beyond modularity: A developmental perspective on cognitive science*. MIT Press. [AS]

Kawato, M., Hayakama, H. & Inui, T. (1993) A forward-inverse optics model of reciprocal connections between visual cortical areas. *Network* 4:415–22. [aAC]

Kay, J., Floreano, D. & Phillips, W. A. (1998) Contextually guided unsupervised learning using local multivariate binary processors. *Neural Networks* 11:117–40. [WAP]

Kay, J. & Phillips, W. A. (2010) Coherent Infomax as a computational goal for neural systems. *Bulletin of Mathematical Biology* 73:344–72. doi: 10.1007/s11538-010-9564-x. [rAC, WAP]

Keane, B. P., Silverstein, S. M., Wang, Y., Zalakostas, A., Vlajnic, V., Mikkilineni, D. & Papathomas, T. V. (in press) Reduced depth inversion illusions in schizophrenia are state-specific and occur for multiple object types and viewing conditions. *Journal of Abnormal Psychology*. [SMS]

Keller, G. B., Bonhoeffer, T. & Hubener, M. (2012) Sensorimotor mismatch signals in primary visual cortex of the behaving mouse. *Neuron* 74:809–15. [rAC, TE]

Khalil, E. L. (1989) Adam Smith and Albert Einstein: The aesthetic principle of truth. *History of Economics Society Bulletin* 11(2):222–37. [ELK]

Khalil, E. L. (2010) The Bayesian fallacy: Distinguishing internal motivations and religious beliefs from other beliefs. *Journal of Economic Behavior and Organization* 75(2):268–80. doi: 10.1016/j.jebo.2010.04.004. [ELK]

Kinoshita, M., Gilbert, C. D. & Das, A. (2009) Optical imaging of contextual interactions in V1 of the behaving monkey. *Journal of Neurophysiology* 102: 1930–44. [SMS]

Kitayama, S. & Cohen, D. (2010) *Handbook of cultural psychology*. The Guilford Press. [JBH]

Kleinschmidt, D. & Jaeger, T. F. (2011) *A Bayesian belief updating model of phonetic recalibration and selective adaptation*. Association for Computational Linguistics – Computational Modeling and Computational Linguistics. [TAF]

Knill, D. & Pouget, A. (2004) The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neuroscience* 27(12):712–19. [aAC]

Koethe, D., Kranaster, L., Hoyer, C., Gross, S., Neatby, M. A., Schultze-Lutter, F., Ruhrmann, S., Klosterkötter, J., Hellmich, M. & Leweke, F. M. (2009) Binocular depth inversion as a paradigm of reduced visual information processing in prodromal state, antipsychotic-naive and treated schizophrenia. *European Archives of Psychiatry and Clinical Neuroscience* 259:195–202. [SMS]

Kohonen, T. (1989) *Self-organization and associative memory*. Springer-Verlag. [aAC]

Kok, P., Jehee, J. F. & de Lange, F. P. (2012) Less is more: Expectation sharpens representations in the primary visual cortex. *Neuron* 75(2):265–70. [LM]

Kok, P., Rahnev, D., Jehee, J. F., Lau, H. C. & de Lange, F. P. (2011) Attention reverses the effect of prediction in silencing sensory signals. *Cerebral Cortex* 22:2197–206. [rAC, TE]

König, P. & Krüger, N. (2006) Symbols as self-emergent entities in an optimization process of feature extraction and predictions. *Biological Cybernetics* 94(4):325–34. [aAC, PK]

Körding, K. P., Kayser, C., Einhäuser, W. & König, P. (2004) How are complex cell properties adapted to the statistics of natural stimuli? *Journal of Neurophysiology* 91(1):206–12. [PK]

Körding, K. P. & König, P. (2000) Learning with two sites of synaptic integration. *Network: Computation in Neural Systems* 11:1–15. [WAP]

Körding, K. P., Tenenbaum, J. B. & Shadmehr, R. (2007) The dynamics of memory as a consequence of optimal adaptation to a changing body. *Nature Neuroscience* 10:779–86. [aAC]

Körding, K. P. & Wolpert, D. M. (2004) Bayesian integration in sensorimotor learning. *Nature* 427(6971):244–47. doi:10.1038/nature02169. [GB]

Kosslyn, S. M., Thompson, W. L., Kim, I. J. & Alpert, N. M. (1995) Topographical representations of mental images in primary visual cortex. *Nature* 378:496–98. [aAC]

Kraljic, T., Samuel, A. G. & Brennan, S. E. (2008) First impressions and last resorts: How listeners adjust to speaker variability. *Psychological Science* 19:332–38. [TAF]

Kriegstein, K. & Giraud, A. (2006) Implicit multisensory associations influence voice recognition. *PLoS (Public Library of Science) Biology* 4(10):e326. [aAC]

Kukona, A., Fang, S., Aicher, K. A., Chen, H. & Magnuson, J. S. (2011) The time course of anticipatory constraint integration. *Cognition* 119:23–42. [TAF]

Kurumada, C., Brown, M. & Tanenhaus, M. K. (2012) Pragmatic interpretation of contrastive prosody: It looks like adaptation. In: *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, ed. N. Miyake, D. Peebles, & R. P. Cooper, pp. 647–52. Cognitive Science Society. [TAF]

Kveraga, K., Ghuman, A. & Bar, M. (2007) Top-down predictions in the cognitive brain. *Brain and Cognition* 65:145–68. [aAC]

Ladinig, O., Honing, H., Haden, G. & Winkler, I. (2009) Probing attentive and preattentive emergent meter in adult listeners without extensive music training. *Music Perception* 26:377–86. [LH]

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* 104:211–40. [aAC]

Landauer, T. K., Foltz, P. W. & Laham, D. (1998) Introduction to Latent Semantic Analysis. *Discourse Processes* 25: 259–84. [aAC]

Lange, C.G. (1885/1912). The mechanisms of the emotions. In: *The Classical Psychologists*, ed. B. Rand. pp. 672–684. Houghton Mifflin. [AKS]

Langner, R., Kellermann, T., Boers, F., Sturm, W., Willmes, K. & Eickhoff, S. B. (2011) Modality-specific perceptual expectations selectively modulate baseline activity in auditory, somatosensory, and visual cortices. *Cerebral Cortex* 21 (12):2850–62. [aAC]

Large, E. W., Fink, P. & Kelso, J. A. S. (2002) Tracking simple and complex sequences. *Psychological Research* 66:3–17. [RSS]

Lee, D. & Wang, X.-J. (2009) Mechanisms for stochastic decision making in the primate frontal cortex: Single-neuron recording and circuit modeling. In: *Neuroeconomics: Decision making and the brain*, ed. P. Glimcher, C. Camerer, E. Fehr & R. Poldrack, pp. 481–501. Elsevier. [rAC, DRo]

Lee, H., Ekanadham, C. & Ng, A. (2008) Sparse deep belief net model for visual area V2. In: *Advances in Neural Information Processing Systems 20 (NIPS'07)*, ed. J. Platt, D. Koller, Y. Singer, & S. Roweis, pp. 873–80. MIT Press. [TT]

Lee, M. (2010) Emergent and structured cognition in Bayesian models: Comment on Griffiths et al. and McClelland et al. *Trends in Cognitive Sciences* 14 (8):345–46. [aAC]

Lee, S. H., Blake, R. & Heeger, D. J. (2005) Traveling waves of activity in primary visual cortex during binocular rivalry. *Nature Neuroscience* 8 (1):22–23. [aAC]

Lee, T. S. & Mumford, D. (2003) Hierarchical Bayesian inference in the visual cortex. *Journal of Optical Society of America, A* 20(7):1434–48. [aAC]

Lehnert, W. (2007) Cognition, computers, and car bombs: How Yale prepared me for the 90's. In: *Beliefs, reasoning, and decision making: Psycho-logic in honor of Bob Abelson*, ed. R. Schank & E. Langer, pp. 143–73. Erlbaum. [aAC]

Lenggenhager, B., Tadi, T., Metzinger, T. & Blanke, O. (2007) Video ergo sum: Manipulating bodily self-consciousness. *Science* 317(5841):1096. [BP]

Leopold, D. & Logothetis, N. (1999) Multistable phenomena: Changing views in perception. *Trends in Cognitive Sciences* 3:254–64. [aAC]

Levinson, S. C. (2006) On the human "Interaction Engine." In: *Roots of human sociality*, ed. N. J. Enfield & S. C. Levinson, pp. 39–69. Berg. [AR]

Lewis, P. A. & Miall, R. C. (2003) Distinct systems for automatic and cognitively controlled time measurement: Evidence from neuroimaging. *Current Opinion in Neurobiology* 13:250–55. [LH]

Lewis, P. A. & Miall, R. C. (2006) A right hemispheric prefrontal system for cognitive time measurement. *Behavioural Processes* 71:226–234. [LH]

Ling, S. & Carrasco, M. (2006) When sustained attention impairs perception. *Nature Neuroscience* 9(10):1243–45. [NB]

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

Little, D. Y. & Sommer, F. T. (2011) Learning in embodied action-perception loops through exploration. Online Publication arXive:1112.1125. [DYL]

Lochmann, T. & Deneve S. (2011) Neural processing as causal inference. *Current Opinion in Neurobiology* 21(5):774–78. [MWS]

Lochmann, T., Ernst, U. A. & Denève, S. (2012) Perceptual inference predicts contextual modulations of sensory responses. *The Journal of Neuroscience* 32 (12):4179–95. [NB]

Loui, P., Wessel, D. & Hudson Kam, C. L. (2010) Humans rapidly learn grammatical structure in a new musical scale. *Music Perception* 27:377–88. [RSS]

Luck, S. J. (2006) The operation of attention – millisecond by millisecond – over the first half second. In: *The first half second: The microgenesis and temporal dynamics of unconscious and conscious visual processing*, ed. H. Ö. B. G. Breitmeyer, pp. 187–206. MIT Press. [HB]

MacKay, D. J. C. (1995) Free-energy minimization algorithm for decoding and cryptoanalysis. *Electron Letters* 31:445–47. [aAC]

MacKay, D. M. (1956) The epistemological problem for automata. In: *Automata studies*, ed. C. E. Shannon & J. McCarthy, pp. 235–51. Princeton University Press. [aAC]

Madison, G. (2001) Variability in isochronous tapping: Higher-order dependencies as a function of inter tap interval. *Journal of Experimental Psychology: Human Perception and Performance* 27:411–22. [LH]

Madison, G., Forsman, L., Blom, Ö., Karabanov, A. & Ullén, F. (2009) Correlations between general intelligence and components of serial timing variability. *Intelligence* 37:68–75. [LH]

Maher, B. (1988) Anomalous experience and delusional thinking: The logic of explanations. In: *Delusional beliefs*, ed. T. F. Oltmanns & B. A. Maher, pp. 15–33. Wiley. [aAC]

Maloney, L. T. & Mamassian, P. (2009) Bayesian decision theory as a model of visual perception: Testing Bayesian transfer. *Visual Neuroscience* 26:147–55. [aAC]

Maloney, L. T. & Zhang, H. (2010) Decision-theoretic models of visual perception and action. *Vision Research* 50:2362–74. [aAC]

Mamassian, P., Landy, M. & Maloney, L. (2002) Bayesian modeling of visual perception. In: *Probabilistic models of the brain*, ed. R. Rao, B. Olshausen & M. Lewicki, pp. 13–36. MIT Press. [aAC]

Mandler, J. M. (1984) *Stories, scripts, and scenes: Aspects of schema theory*. Erlbaum. [JBH]

Mar, R. A. & Oatley, K. (2008) The function of fiction is the abstraction and simulation of social experience. *Perspectives on Psychological Science* 3(3):173–92. [JBH]

Marcus, G. (2008) *Kluge: The haphazard construction of the human mind*. Houghton-Mifflin. [aAC]

Mareschal, D., Johnson, M. H., Siros, S., Spratling, M. W., Thomas, M. S. C. & Westermann, G. (2007) *Neuroconstructivism – I.: How the brain constructs cognition*. Oxford University Press. [aAC, MWS]

Marr, D. (1982). *Vision: A computational approach*. Freeman. [aAC]

Matell, M. S. & Meck, W. H. (2004) Cortico-striatal circuits and interval timing: Coincidence detection of oscillatory processes. *Cognitive Brain Research* 21:139–70. [LH]

Mattusek, P. (1987) Studies in delusional perception (translated and condensed). In: *Clinical roots of the schizophrenia concept. Translations of seminal European contributions on Schizophrenia*, ed. J Cutting & M. Sheppard, pp. 87–103. Cambridge University Press. (Originally published in 1952.) [SMS]

McAdams, D. P. (1997) *The stories we live by: Personal myths and the making of the self*. The Guilford Press. [JBH]

McAdams, D. P. (2006) The problem of narrative coherence. *Journal of Constructivist Psychology* 19(2):109–25. [JBH]

McCarthy, J. (2008) The well-designed child. *Artificial Intelligence* 172(18):2003–14. [AS]

McClelland, J., Botvinick, M., Noelle, D., Plaut, D., Rogers, T., Seidenberg, M. & Smith, L. (2010) Letting structure emerge: Connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences* 14(8):348–56. [aAC]

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. [aAC]

McClelland, J., Rumelhart, D. & the PDP Research Group (1986) *Parallel distributed processing, vol. 2*. MIT Press. [aAC]

McMurray, B., Tanenhaus, M. K. & Aslin, R. N. (2009) Within-category VOT affects recovery from "lexical" garden paths: Evidence against phoneme-level inhibition. *Journal of Memory and Language* 60:65–91. [TAF]

Melloni, L., Schwiedrzik, C. M., Muller, N., Rodriguez, E. & Singer, W. (2011) Expectations change the signatures and timing of electrophysiological correlates of perceptual awareness. *Journal of Neuroscience* 31(4):1386–96. [aAC]

Menary, R. (2007) *Cognitive integration: Attacking the bounds of cognition*. Palgrave Macmillan. [aAC]

Meng, M. & Tong, F. (2004) Can attention selectively bias bistable perception? differences between binocular rivalry and ambiguous figures. *Journal of Vision* 4:539–51. [aAC]

Merker, B. (2004) Cortex, countercurrent context, and dimensional integration of lifetime memory. *Cortex* 40:559–76. [aAC]

Merker, B. H., Madison G. S. & Eckerdal, P. (2009) On the role and origin of isochrony in human rhythmic entrainment. *Cortex* 45:4–17. [LH]

Meyer, T. & Olson, C. R. (2011) Statistical learning of visual transitions in monkey inferotemporal cortex. *Proceedings of the National Academy of Sciences USA* 108:19401–406. [rAC, TE]

Meyer. L. B. (1956) *Emotion and meaning in music*. University of Chicago Press. [RSS]

Miall, R. C. (1989) The storage of time intervals using oscillating neurons. *Neural Computation* 1:359–71. [LH]

Milner, D. & Goodale, M. (2006) *The visual brain in action*, 2nd edition. Oxford University Press. [aAC]

Molnar-Szakacs, I. & Overy, K. (2006) Music and mirror neurons: From motion to 'e' motion. *Social Cognition and Affective Neuroscience* 1:235–41. [RSS]

Morrone, C. M., Ross, J. & Burr, D. (2005) Saccadic eye movements cause compression of time as well as space. *Nature Neuroscience* 8:950–54. [LH]

Muckli, L. (2010) What are we missing here? Brain imaging evidence for higher cognitive functions in primary visual cortex V1. *International Journal of Imaging Systems Technology (IJIST)* 20:131–39. [aAC]

Muckli, L., Kohler, A., Kriegeskorte, N. & Singer, W. (2005) Primary visual cortex activity along the apparent-motion trace reflects illusory perception. *PLoS (Public Library of Science) Biology* l3:e265. [aAC, LM]

Muckli, L. & Petro, L.S. (2013) Network interactions: Non-geniculate input to V1. *Current Opinion in Neurobiology* 23(2):195–201. [LM]

Mumford, D. (1992) On the computational architecture of the neocortex. II. The role of cortico-cortical loops. *Biological Cybernetics* 66(3):241–51. [arAC, TE, KF]

Mumford, D. (1994) Neuronal architectures for pattern-theoretic problems. In: *Large-scale theories of the cortex*, ed. C. Koch & J. Davis, pp. 125–52. MIT Press. [aAC]

Murray, D. J., Ellis, R. R., Bandomir, C. A. & Ross, H. E. (1999) Charpentier (1891) on the size-weight illusion. *Perception and Psychophysics* 61(8):1681–85. [GB]

Murray, S. O., Boyaci, H. & Kersten, D. (2006) The representation of perceived angular size in human primary visual cortex. *Nature Reviews: Neuroscience* 9:429–34. [aAC]

Murray, S. O., Kersten, D., Olshausen, B. A., Schrater, P. & Woods, D. L. (2002) Shape perception reduces activity in human primary visual cortex. *Proceedings of the National Academy of Sciences USA* 99(23):15164–69. [arAC, TE]

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

Musmann, H. (1979) Predictive image coding. In: *Image transmission techniques*, ed. W. K, Pratt, *Advances in Electronics and Electron Physics*, Supplement 12:73–112, Academic Press, Orlando, FL. [aAC]

Nagarajan, S. S., Blake, D. T., Wright, B. A., Byl, N. & Merzenich, M. M. (1998) Practice-related improvements are temporally specific but generalize across skin location, hemisphere and modality. *Journal of Neuroscience* 18:1559–70. [LH]

Nagel, S. K., Carl, C., Kringe, T., Märtin, R. & König, P. (2005) Beyond sensory substitution – learning the sixth sense. *Journal of Neural Engineering* 2(4):R13–R26. doi:10.1088/1741-2560/2/4/R02. [PK]

Narmour, E. (1990) *The analysis and cognition of basic melodic structures: The implication-realization model.* University of Chicago Press. [RSS]

Neal, R. M. & Hinton, G. (1998) A view of the EM algorithm that justifies incremental, sparse, and other variants. In: *Learning in graphical models*, ed. M. I. Jordan, pp. 355–68. Kluwer. [aAC]

Neisser, U. (1967) *Cognitive psychology*. Appleton-Century-Crofts. [aAC]

Nelson, K. (2003) Self and social functions: Individual autobiographical memory and collective narrative. *Memory* 11(2):125–36. [JBH]

Nelson, K. & Fivush, R. (2004) The emergence of autobiographical memory: A social cultural developmental theory. *Psychological Review* 111(2):486–511. [JBH]

Nelson, P. (2012) Towards a social theory of rhythm. In: *Topics in musical universals/ Actualités des Universaux Musicaux*, ed. J.-L. Leroy. Editions des Archives Contemporaines. [RSS]

Noë, A. (2004) *Action in perception*. MIT Press. [aAC, TF]

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/Hill and Wang. [aAC, TF]

North, A. C. & Hargreaves, D. J. (1995) Subjective complexity, familiarity, and liking for popular music. *Psychomusicology* 14:77–93. [RSS]

Oatley, K. (1992) *Best laid schemes: The psychology of emotions*. Cambridge University Press. [JBH]

Oatley, K. (1999) Why fiction may be twice as true as fact: Fiction as cognitive and emotional simulation. *Review of General Psychology* 3(2):101–17. [JBH]

Olshausen, B. A. & Field, D. J. (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381(6583):607–609. [arAC, PK, TT]

Olshausen, B. A. & Field, D. J. (2005) How close are we to understanding V1? *Neural Computation* 17:1665–99. [MWS]

Overy, K. & Molnar-Szakacs, I. (2009) Being together in time: Musical experience and the mirror neuron system. *Music Perception* 26(5):489–504. [aAC, RSS]

Owen, A. M., McMillan, K. M., Laird, A. R. & Bullmore, E. (2005) N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human Brain Mapping* 25:46–59. [LH]

Oyama, S. (1999) *Evolution's eye: Biology, culture and developmental systems*. Duke University Press. [aAC]

Pack, C. C. & Born, R. T. (2001) Temporal dynamics of a neural solution to the aperture problem in visual area MT of macaque brain. *Nature* 409:1040–42. [aAC]

Palaniyappan, L. & Liddle, P. F. (2012) Does the salience network play a cardinal role in psychosis? An emerging hypothesis of insular dysfunction. *Journal of Psychiatry and Neuroscience* 37(1):17–27. [AKS]

Pascual-Leone, A. & Hamilton, R. (2001) The metamodal organization of the brain. *Progress in Brain Research* 134:427–45. [aAC]

Paulus, M. P. & Stein, M. B. (2006) An insular view of anxiety. [Review]. *Biological Psychiatry* 60(4):383–87. [AKS]

Pearce, J. M. & Hall, G. (1980) A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review* 87:532–52. [TE]

Pecenka, N. & Keller, P. E. (2011) The role of temporal prediction abilities in interpersonal sensorimotor synchronization. *Experimental Brain Research* 211: 505–15. [RSS]

Pennebaker, J. W. & Seagal, J. D. (1999) Forming a story: The health benefits of narrative. *Journal of Clinical Psychology* 55(10):1243–54. [JBH]

Peterson, J. B. (1999) *Maps of meaning: The architecture of belief*. Routledge. [JBH]

Petkova, V. I. & Ehrsson, H. H. (2008) If I were you: Perceptual illusion of body swapping. *PLoS ONE* 3(12):e3832. [BP]

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. [arAC]

Phillips, W. A. (2012) Self-organized complexity and coherent Infomax from the viewpoint of Jaynes's probability theory. *Information* 3(1):1–15. doi:10.3390/ info3010001. [WAP, SMS]

Phillips, W. A., Chapman, K. L. S. & Berry, P. D. (2004) Size perception is less context-sensitive in males. *Perception* 33:79–86. [WAP]

Phillips, W. A., Kay, J. & Smyth, D. (1995) The discovery of structure by multistream networks of local processors with contextual guidance. *Network: Computation in Neural Systems* 6:225–46. [rAC, PK, WAP]

Phillips, W. A. & Silverstein, S. M. (2003) Convergence of biological and psychological perspectives on cognitive coordination in schizophrenia. *Behavioral and Brain Sciences* 26:65–82; discussion 82–137. [WAP, SMS]

Phillips, W. A. & Singer, W. (1997) In search of common foundations for cortical computation. *Behavioral and Brain Sciences* 20:657–722. [WAP, MWS]

Phillips, W. A., von der Malsburg, C. & Singer, W. (2010) Dynamic coordination in brain and mind. In: *Strüngmann Forum Report, vol. 5: Dynamic coordination in the brain: From neurons to mind*, ed. C. von der Malsburg, W. A. Phillips & W. Singer, Chapter 1, pp. 1–24. MIT Press. [rAC, WAP]

Phillips-Silver J. & Trainor, L. J. (2007) Hearing what the body feels: Auditory encoding of rhythmic movement. *Cognition* 105:533–46. [LH]

Phillips-Silver, J. & Trainor, L. J. (2008) Vestibular influence on auditory metrical interpretation. *Brain and Cognition* 67:94–102. [RSS]

Piaget, J. (1952) *The origins of intelligence in children*. International University Press. [PK]

Pickering, M. J. & Garrod, S. (2007) Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences* (11):105–110. [arAC]

Platt, M. & Glimcher, P. (1999) Neural correlates of decision variables in parietal cortex. *Nature* 400:233–38. [DRo]

Ploghaus, A., Tracey, I., Gati, J. S., Clare, S., Menon, R. S., Matthews, P. M. & Rawlins, J. N. (1999) Dissociating pain from its anticipation in the human brain. *Science* 284(5422):1979–81. [AKS]

Posner, M. (1980) Orienting of attention. *Quarterly Journal of Experimental Psychology* 32:33. [rAC]

Pouget, A., Dayan, P. & Zemel, R. (2003) Inference and computation with population codes. *Annual Review of Neuroscience* 26:381–410. [aAC]

Powers, W. T. (1973) *Behavior, the control of perception*. Aldine de Gruyter. [AS]

Pribram, K. H. (1971) *Languages of the brain*. Prentice-Hall. [BB]

Pribram, K. H. (1980) The orienting reaction: Key to brain representational mechanisms. In: *The orienting reflex in humans*, ed. H. D. Kimmel, pp. 3–20. Erlbaum. [aAC]

Prinz, J. J. (2005) A neurofunctional theory of consciousness. In: *Cognition and the brain: Philosophy and neuroscience movement*, ed. A. Brook & K. Akins, pp. 381–96. Cambridge University Press. [aAC]

Proulx, T., Inzlicht, M. & Harmon-Jones, E. (2012) Understanding all inconsistency compensation as a palliative response to violated expectations. *Trends in Cognitive Sciences* 16(5):285–91. [JBH]

Purves, D. & Lotto, R. B. (2003) *Why we see what we do: An empirical theory of vision*. Sinauer. [aAC]

Quine, W. V. O. (1951) Two dogmas of empiricism. *The Philosophical Review* 60:20–43. [MLA]

Rammsayer, T. (1999) Neuropharmacological evidence for different timing mechanisms in humans. *The Quarterly Journal of Experimental Psychology, Section B: Comparative and Physiological Psychology* 52:273–86. [LH]

Rao, R. P. N. & Ballard, D. H. (1999) Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience* 2(1):79–87. [aAC, TE, KF, LM, MWS]

Rao, R. P. N. & Sejnowski, T. J. (2002) Predictive coding, cortical feedback, and spike-timing dependent plasticity. In: *Probabilistic models of the brain: Perception and neural function*, ed. R. P. N. Rao, B. A. Olshausen & M. S. Lewicki, pp. 297–315. MIT Press. [aAC]

Rauss, K., Schwartz, S. & Pourtois, G. (2011) Top-down effects on early visual processing in humans: A predictive coding framework. *Neuroscience and Biobehavioral Reviews* 35(5):1237–53. [aAC]

Read, J., Perry, B. D., Moskowitz, A. & Connolly, J. (2001) The contribution of early traumatic events to schizophrenia in some patients: A traumagenic neurodevelopmental model. *Psychiatry* 64: 319–45. [SMS]

Read, J., van Os, J., Morrison, A. P. & Ross, C. A. (2005) Childhood trauma, psychosis and schizophrenia: A literature review with theoretical and clinical implications. *Acta Psychiatrica Scandinavica* 112:330–50. [SMS]

Reddy, L., Tsuchiya, N. & Serre, T. (2010) Reading the mind's eye: decoding category information during mental imagery. *NeuroImage* 50(2):818–25. [aAC]

Reich, L., Szwed, M., Cohen, L. & Amedi, A. (2011) A ventral stream reading center independent of visual experience. *Current Biology* 21:363–68. [aAC]

Reichert, D., Seriès, P. & Storkey, A. (2010) Hallucinations in Charles Bonnet Syndrome induced by homeostasis: A Deep Boltzmann Machine model. *Advances in Neural Information Processing Systems* 23:2020–28. [rAC]

Repp, B. (1999) Detecting deviations from metronomic timing in music: Effects of perceptual structure on the mental timekeeper. *Perception and Psychophysics* 61(3):529–48. [RSS]

Rescorla, M. (in press) Bayesian perceptual psychology to appear. In: *Oxford handbook of the philosophy of perception*, ed. M. Matthen. Oxford University Press. [aAC]

Ricoeur, P., Blamey, K. & Pellauer, D. (1990) *Time and narrative, vol. 3.* University of Chicago Press. [JBH]

Rieke, F. (1999) *Spikes: Exploring the neural code*, MIT Press. [aAC]

Riesenhuber, M. & Poggio, T (2000) Models of object recognition. *Nature Neuroscience* 3(Suppl.):1199–204. [TE]

Robbins, H. (1956) An empirical Bayes approach to statistics. In: *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, vol. 1: Contributions to the Theory of Statistics*, pp. 157–63. University of California Press. [aAC]

Roepstorff, A. (2008) Things to think with: Words and objects as material symbols. *Philosophical Transactions of the Royal Society, B* 363(1499):2049–54. [AR]

Roepstoff, A. & Frith, C. (2012) Neuroanthropology or simply anthropology? Going experimental as method, as object of study, and as research aesthetic. *Anthropological Theory* 12(1):101–11. [AR]

Roepstorff, A., Niewohner, J. & Beck, S. (2010) Enculturing brains through patterned practices. *Neural Networks* 23(8–9):1051–59. [arAC, AR]

Rorty, R. (1979) *Philosophy and the mirror of nature*. Princeton University Press. [MLA]

Ross, D., Sharp, C., Vuchinich, R. & Spurrett, D. (2008) *Midbrain mutiny: The picoeconomics and neuroeconomics of disordered gambling*. MIT Press. [DRo]

Ross, H. E. (1969) When is a weight not illusory? *The Quarterly Journal of Experimental Psychology* 21(4):346–55. doi:10.1080/14640746908400230. [GB]

Rowlands, M. (1999) *The body in mind: Understanding cognitive processes*. Cambridge University Press. [aAC]

Rowlands, M (2006) *Body language: Representing in action*. MIT Press. [aAC]

Rumelhart, D. E., McClelland, J. L. & the PDP Research Group (1986) *Parallel distributed processing, vol. I: Explorations in the microstructure of cognition. Foundations*. MIT Press. [aAC]

Rust, N. C., Schwartz, O., Movshon, J. A. & Simoncelli, E. P. (2005) Spatiotemporal elements of Macaque V1 receptive fields. *Neuron* 46:945–56. [TT]

Sachs, E. (1967) Dissociation of learning in rats and its similarities to dissociative states in man. In: *Comparative psychopathology: Animal and human*, ed. J. Zubin & H. Hunt, pp. 249–304. Grune and Stratton. [aAC]

Sadakata, M., Desain, P. & Honing, H. (2006) The Bayesian way to relate rhythm perception and production. *Music Perception* 23:269–88. [RSS]

Salakhutdinov, R. & Hinton, G. E. (2009) Deep Boltzmann machines. *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics* (AISTATS), *vol. 5*, ed. D. van Dyk & M. Welling, pp. 448–55. *The Journal of Machine Learning Research*, published online, at http://jmlr.csail.mit.edu/proceedings/papers/v5/ [rAC]

Sanders, L. L., Muckli, L., de Millas, W., Lautenschlager, M., Heinz, A., Kathmann, N. & Sterzer, P. (2012) Detection of visual events along the apparent motion trace in patients with paranoid schizophrenia. *Psychiatry Research*. [Epub ahead of print: April 28, 2012]. Available at: http://dx.doi.org/10.1016/j.psychres.2012.03.006. [LM]

Santhouse, A. M., Howard, R. J. & ffytche, D. H. (2000) Visual hallucinatory syndromes and the anatomy of the visual brain. *Brain* 123:2055–64. [SMS]

Sarbin, T. R. (1986) *Narrative psychology: The storied nature of human conduct*. Praeger/Greenwood. [JBH]

Sass, L. (1992) *Madness and modernism. Insanity in the light of modern art, literature and thought*. Basic Books. [SMS]

Saxe, A., Bhand, M., Mudur, R., Suresh, B. & Ng, A. (2011) Modeling cortical representational plasticity with unsupervised feature learning. Poster presented at COSYNE 2011, Salt Lake City, Utah, February 24–27, 2011. Available at: http://www.stanford.edu/~asaxe/papers. [TT]

Schachter, S. & Singer, J. E. (1962) Cognitive, social, and physiological determinants of emotional state. *Psychological Review* 69:379–99. [AKS]

Schaefer, R. S., Vlek, R. J. & Desain, P. (2011a) Decomposing rhythm processing: Electroencephalography of perceived and self-imposed rhythmic patterns. *Psychological Research* 75(2):95–106. [RSS]

Schaefer, R. S., Vlek, R. J. & Desain, P. (2011b) Music perception and imagery in EEG: Alpha band effects of task and stimulus. *International Journal for Psychophysiology* 82(3):254–59. [RSS]

Schank, R. & Abelson, R. (1977) *Scripts, plans, goals and understanding: An inquiry into human knowledge structures*. Erlbaum. [JBH]

Schenk, T. & McIntosh, R. (2010) Do we have independent visual streams for perception and action? *Cognitive Neuroscience* 1:52–78. [rAC]

Schultz, W., Dayan, P. & Montague, P. R. (1997) A neural substrate of prediction and reward. *Science* 275:1593–99. [DRo]

Schwartz, O., Hsu, A. & Dayan, P. (2007) Space and time in visual context *Nature Reviews Neuroscience* 8:522–35. [aAC]

Segall, M. H., Campbell, D. T. & Herskovits, M. J. (1963) Cultural differences in the perception of geometric illusions. *Science* 139(3556):769–71. [PK]

Sellars, W. (1962) Philosophy and the scientific image of man. In: *Frontiers of Science and Philosophy*, ed. R. G. Colodny, pp. 35–78. University of Pittsburgh Press. [Reprinted in: *Science, Perception and Reality* by W. Sellars (1963, Routledge & Kegan Paul)]. [aAC, DCD]

Sellars, W. (1963) *Science, perception, and reality*. Routledge & Kegan Paul. [JBH]

Seth, A. K., Suzuki, K. & Critchley, H. D. (2011) An interoceptive predictive coding model of conscious presence. *Frontiers in Psychology* 2:395. [rAC, AKS]

Shams, L., Ma, W. J. & Beierholm, U. (2005) Sound-induced flash illusion as an optimal percept. *NeuroReport* 16(10):1107–10. [aAC]

Shi, Yun Q. & Sun, H. (1999) *Image and video compression for multimedia engineering: Fundamentals, algorithms, and standards*. CRC Press. [aAC]

Silverstein, S. M., Berten, S., Essex, B., Kovács, I., Susmaras, T. & Little, D. M. (2009) An fMRI examination of visual integration in schizophrenia. *Journal of Integrative Neuroscience* 8:175–202. [SMS]

Silverstein, S. M. & Keane, B. P. (2011) Perceptual organization impairment in schizophrenia and associated brain mechanisms: Review of research from 2005 to 2010. *Schizophrenia Bulletin* 37:690–99. [SMS]

Simoncelli, E. P. & Olshausen, B. A. (2001) Natural image statistics and neural representation. *Annual Review of Neuroscience* 24:1193–216. [PK]

Simons, J. S., Schölvinck, M. L., Gilbert, S. J., Frith, C. D., & Burgess, P. W. (2006) Differential components of prospective memory? Evidence from fMRI. *Neuropsychologia* 4:1388–97. [LH]

Singer, T., Critchley, H. D. & Preuschoff, K. (2009) A common role of insula in feelings, empathy and uncertainty. *Trends in Cognition Science* 13(8):334–40. [AKS]

Singer, W. (1995) Development and plasticity of cortical processing architectures. *Science* 270:758–64. [SMS]

Sloman, A. (1971) Interactions between philosophy and AI: The role of intuition and non-logical reasoning in intelligence. In: *Proceedings of the 2nd IJCAI [International Joint Conference on Artificial Intelligence]*, ed. D. C. Cooper, pp. 209–26. William Kaufmann. Available at: http://www.cs.bham.ac.uk/research/cogaff/04.html#200407. [AS]

Sloman, A. (1978) *The computer revolution in philosophy*. Harvester Press/Humanities Press. [AS]

Sloman, A. (1979) The primacy of non-communicative language. In: *The analysis of meaning: Informatics 5, Proceedings ASLIB/BCS Conference, Oxford, March 1979*, ed. M. MacCafferty & K. Gray, pp. 1–15. Aslib. Available at: http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html#43. [AS]

Sloman, A. (1982) Image interpretation: The way ahead? In: *Physical and biological processing of images* (Proceedings of an International Symposium organised by The Rank Prize Funds, London, 1982), ed. O. Braddick & A. Sleigh, pp. 380–401. Springer-Verlag. Available at: http://www.cs.bham.ac.uk/research/projects/cogaff/06.html#0604. [AS]

Sloman, A. (1987) Reference without causal links. In: *Advances in artificial intelligence, vol. 2*, ed. J. B. H. du Boulay, D. Hogg, & L. Steels, pp. 369–81. North-Holland. Available at: http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html#5. [AS]

Sloman, A. (1989) On designing a visual system (towards a Gibsonian computational model of vision). *Journal of Experimental and Theoretical AI* 1(4):289–37. Available at: http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html#7. [AS]

Sloman, A. (1990) Must intelligent systems be scruffy? In: *Evolving knowledge in natural science and artificial intelligence*, ed. J. E. Tiles, G. T. McKee & G. C. Dean. Pitman. [aAC]

Sloman, A. (1993) The mind as a control system. In: *Philosophy and the cognitive sciences*, ed. C. Hookway & D. Peterson, pp. 69–110. Cambridge University Press. Available at: http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html#18. [AS]

Sloman, A. (1996) Actual possibilities. In: *Principles of knowledge representation and reasoning: Proceedings of the 5th International Conference (KR '96)*, ed. L. Aiello & S. Shapiro, pp. 627–38. Morgan Kaufmann. [AS]

Sloman, A. (2002) Diagrams in the mind. In: *Diagrammatic representation and reasoning*, ed. M. Anderson, B. Meyer & P. Olivier, pp. 7–28. Springer-Verlag. [AS]

Sloman, A. (2006) Requirements for a fully deliberative architecture (or component of an architecture). Research Note No. COSY-DP-0604, May 2006. School of Computer Science, University of Birmingham, UK. [AS]

Sloman, A. (2008) A multi-picture challenge for theories of vision. Research Note No. COSY-PR-0801. School of Computer Science, University of Birmingham, UK. [AS]

Sloman, A. (2009) Some requirements for human-like robots: Why the recent over-emphasis on embodiment has held up progress. In: *Creating brain-like intelligence*, ed. B. Sendhoff, E. Koerner, O. Sporns, H. Ritter & K. Doya, pp. 248–77. Springer-Verlag. [AS]

Sloman, A. (2010) If learning maths requires a teacher, where did the first teachers come from? In: *Proceedings of the International Symposium on Mathematical Practice and Cognition, AISB 2010 Convention, De Montfort University, Leicester*, ed. A. Pease, M. Guhe & A. Smaill, pp. 30–39. AISB (Society for the Study of Artificial Intelliegence and Simulation of Behaviour). [AS]

Sloman, A. (2011a) Varieties of meta-cognition in natural and artificial systems. In: *Metareasoning: Thinking about thinking*, ed. M. T. Cox & A. Raja, pp. 307–23. MIT Press. [AS]

Sloman, A. (2011b) What's vision for, and how does it work? From Marr (and earlier) to Gibson and beyond. (Online tutorial presentation, September 2011. Available at: http://www.slideshare.net/asloman/). [AS]

Sloman, A. (2012) Paper 4: Virtual machinery and evolution of mind (Part 3). Meta-morphogenesis: Evolution of information-processing machinery. In: *Alan Turing – his work and impact*, ed. S. B. Cooper & J. van Leeuwen. Elsevier. [AS]

Smith, F. W. & Muckli, L. (2010) Nonstimulated early visual areas carry information about surrounding context. *Proceedings of the National Academy of Sciences USA* 16:20099–103. [aAC, LM]

Smith, L. & Gasser, M. (2005) The development of embodied cognition: Six lessons from babies. *Artificial Life* 11(1):13–30. [aAC]

Smith, P. L. & Ratcliff, R. (2004) Psychology and neurobiology of simple decisions. *Trends in Neuroscience* 27:161–68. [aAC]

Sokolov, E. N. (1960) Neuronal models and the orienting reflex. In: *The central nervous system and behavior*, ed. M. Brazier, pp. 187–276. Josiah Macy Jr. Foundation. [BB, aAC]

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

Spratling, M. W. (2008a) Predictive coding as a model of biased competition in visual attention. *Vision Research* 48(12):1391–408. [aAC, WAP, MWS]

Spratling, M. W. (2008b) Reconciling predictive coding and biased competition models of cortical function. *Frontiers in Computational Neuroscience* 2(4):1–8. [LM, MWS]

Spratling, M. W. (2010) Predictive coding as a model of response properties in cortical area V1. *Journal of Neuroscience* 30(9):3531–543. [MWS]

Spratling, M. W. (2011) A single functional model accounts for the distinct properties of suppression in cortical area V1. *Vision Research* 51(6):563–76. [MWS]

Spratling, M. W. (2012a) Predictive coding accounts for V1 response properties recorded using reverse correlation. *Biological Cybernetics* 106(1):37–49. [MWS]

Spratling, M. W. (2012b) Unsupervised learning of generative and discriminative weights encoding elementary image components in a predictive coding model of cortical function. *Neural Computation* 24(1):60–103. [MWS]

Srinivasan, M. V., Laughlin, S. B. & Dubs A. (1982) Predictive coding: A fresh view of inhibition in the retina. *Proceedings of the Royal Society of London, B* 216:427–59. [aAC]

Staum Casasanto, L. (2008) Does social information influence sentence processing? In: *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, ed. B. C. Love, K. McRae, & V. M. Sloutsky, pp. 799–804. Cognitive Science Society. [TAF]

Stephan, K., Friston, K. & Frith, C. (2009) Dysconnection in schizophrenia: From abnormal synaptic plasticity to failures of self-monitoring. *Schizophrenia Bulletin* 35(3):509–27. [rAC]

Sterelny, K. (2003) *Thought in a hostile world: The evolution of human cognition*, Blackwell. [aAC]

Sterelny, K. (2007) Social intelligence, human intelligence and niche construction. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* 362(1480):719–30. [aAC]

Still, S. (2009) Information-theoretic approach to interactive learning. *Europhysics Letters* 85:28005. [DYL]

Störmer, V., McDonald, J. & Hillyard, S. A. (2009) Cross-modal cueing of attention alters appearance and early cortical processing of visual stimuli. *Proceedings of the National Academy of Sciences USA* 106(52):22456–61. [NB]

Stotz, K. (2010) Human nature and cognitive–developmental niche construction. *Phenomenology and the Cognitive Sciences* 9(4):483–501. [aAC]

Summerfield, C. & Egner, T (2009) Expectation (and attention) in visual cognition. *Trends in Cognition Science* 13:403–409. [aAC, TE]

Summerfield, C., Egner, T., Greene, M., Koechlin, E., Mangels, J. & Hirsch, J (2006) Predictive codes for forthcoming perception in the frontal cortex. *Science* 314:1311–14. [TE]

Summerfield, C. & Koechlin, E. (2008) A neural representation of prior information during perceptual inference. *Neuron* 59:336–47. [TE]

Summerfield, C., Trittschuh, E. H., Monti, J. M., Mesulam, M. M. & Egner, T. (2008) Neural repetition suppression reflects fulfilled perceptual expectations. *Nature Neuroscience* 11(9):1004–1006. [aAC, TE]

Summerfield, C., Wyart, V., Johnen, V. M. & De Gardelle, V (2011) Human scalp electroencephalography reveals that repetition suppression varied with expectation. *Frontiers in Human Neuroscience* 5:67. (Online publication). doi:10.3389/fnhum.2011.00067. [TE]

Switkes, E., Mayer, M. J. & Sloan, J. A. (1978) Spatial frequency analysis of the visual environment: Anisotropy and the carpentered environment hypothesis. *Vision Research* 18:1393–99. [BB]

Synofzik, M., Thier, P., Leube, D. T., Schlotterbeck, P. & Lindner, A. (2010) Misattributions of agency in schizophrenia are based on imprecise predictions about the sensory consequences of one's actions. *Brain* 133(Pt. 1):262–71. [AKS]

Tanaka, K. (1996) Inferotemporal cortex and object vision. *Annual Review of Neuroscience* 19:109–39. [PK]

Tanenhaus, M. K. & Hare, M. (2007) Phonological typicality and sentence processing. *Trends in Cognitive Science* 11:93–95. [TAF]

Temperley, D. (2004) Communicative pressure and the evolution of musical styles. *Music Perception* 21:313–37. [RSS]

Temperley, D. (2007) *Music and probability*. The MIT Press. [RSS]

Thelen, E. & Smith, L. (1994) *A dynamic systems approach to the development of cognition and action*. MIT Press. [aAC]

Thompson, E. (2007) *Mind in life: Biology, phenomenology, and the sciences of mind*. Harvard University Press. [aAC]

Tishby, N., Pereira, F. C. & Bialek, W. (1999) The information bottleneck method. In: *Proceedings of the 37th Allerton Conference on Communication, Control, and Computing,* ed. B. Hajek & R. S. Sreenivas, pp. 368–77. University of Illinois Press. [DYL]

Todorov, E. (2004) Optimality principles in sensorimotor control. *Nature Neuroscience* 7(9):907–15. [NS]

Todorov, E. (2006) Optimal control theory. In: *Bayesian brain*, ed. K. Doya, pp. 269–98. MIT Press. [NS]

Todorov, E. (2009) Parallels between sensory and motor information processing. In: *The cognitive neurosciences, 4th edition,* ed. M. Gazzaniga, pp. 613–24. MIT Press. [aAC]

Todorov, E. & Jordan, M. I. (2002) Optimal feedback control as a theory of motor coordination. *Nature Neuroscience* 5(11):1226–35. [aAC, NS]

Todorovic, A., van Ede, F., Maris, E. & de Lange, F. P. (2011) Prior expectation mediates neural adaptation to repeated sounds in the auditory cortex: An MEG study. *Journal of Neuroscience* 31:9118–23. [TE]

Tong, F., Meng, M. & Blake, R. (2006) Neural bases of binocular rivalry. *Trends in Cognitive Sciences* 10:502–11. [MLA]

Toussaint, M. (2009) Probabilistic inference as a model of planned behavior. *Künstliche Intelligenz* 3:23–29. [aAC]

Townsend, B. R., Paninski, L. & Lemon, R. N. (2006) Linear encoding of muscle activity in primary motor cortex and cerebellum. *Journal of Neurophysiology* 96 (5): 2578–92. [DRa]

Trehub, A. (1991) *The cognitive brain*. MIT Press. Available at: http://www.people.umass.edu/trehub/. [AS]

Tribus, M. (1961) *Thermodynamics and thermostatics: An introduction to energy, information and states of matter, with engineering applications*. D. Van Nostrand. [aAC, KF]

Tudusciuc, O. & Nieder, A. (2009) Contributions of primate prefrontal and posterior parietal cortices to length and numerosity representation. *Journal of Neurophysiology* 101(6):2984–94. [DRa]

Turing, A. M. (1952) The chemical basis of morphogenesis. *Philosophical Transactions of Royal Society of London B* 237:37–72. [AS]

Uhlhaas, P. J. & Mishara, A. L. (2007) Perceptual anomalies in schizophrenia: Integrating phenomenology and cognitive neuroscience. *Schizophrenia Bulletin* 33:142–56. [SMS]

Ungerleider, L. G. & Mishkin, M. (1982) Two cortical visual systems. In: *Analysis of visual behavior*, ed. D. Ingle, M. A. Goodale & R. J. Mansfield, pp. 549–86. MIT Press. [KF]

Van Essen, D. C. (2005) Corticocortical and thalamocortical information flow in the primate visual system. *Progress in Brain Research* 149:173–85. [LM]

Van Voorhis, S. & Hillyard, S. A. (1977) Visual evoked potentials and selective attention to points in space. *Perception and Psychophysics* 22(1):54–62. [HB]

Varela, F. J. (1999) The specious present: A neurophenomenology of time consciousness. In: *Naturalizing phenomenology: Issues in contemporary phenomenology and cognitive science*, ed. J. Petitot, F. J. Varela, B. Pachoud & J.-M. Roy, pp. 266–317. Stanford University Press. [TF]

Varela, F. J., Lachaux, J.-P., Rodriguez, E. & Martinerie, J. (2001) The brainweb: Phase synchronization and large-scale integration. *Nature Reviews Neuroscience* 2:229–39. [TF]

Varela, F. J., Thompson, E. & Rosch, E. (1991) *The embodied mind*. MIT Press. [aAC]

Velleman, J. D. (1989) *Practical reflection*. Princeton University Press. [aAC]

Verschure, P., Voegtlin, T. & Douglas, R. (2003) Environmentally mediated synergy between perception and behaviour in mobile robots. *Nature* 425:620–24. [aAC]

Vetter, P., Edwards, G. & Muckli, L. (2012) Transfer of predictive signals across saccades. *Frontiers in Psychology* 3 (176):1–10. [LM]

Vetter, P., Grosbras, M. H. & Muckli, L. (under revision) TMS over V5 disrupts motion predictability. [LM]

Vilares, I. & Körding, K. (2011) Bayesian models: The structure of the world, uncertainty, behavior, and the brain. *Annals of the New York Academy of Science* 1224:22–39. [aAC]

Villalon-Turrubiates, I. E., Andrade-Lucio, J. A. & Ibarra-Manzano, O. G. (2004) Multidimensional digital signal estimation using Kalman's theory for computer-aided applications. In: *Proceedings of the International Conference on Computing, Communications, and Control Technologies, Austin, Texas, August 14–17, 2004 (CCCT Proceedings, Vol. 7)*, ed. H.-W. Chu, pp. 48–53. University of Texas Press. [DRa]

Vinje, W. E. & Gallant J. L. (2000) Sparse coding and decorrelation in primary visual cortex during natural vision. *Science* 287:1273–76. [TT]

von der Malsburg, C., Phillips, W. A. & Singer, W., eds. (2010) *Strungmann Forum Report, Vol. 5. Dynamic coordination in the brain: From neurons to mind*. MIT Press. [rAC, WAP]

von Uexküll, J. (1934/1957) A stroll through the worlds of animals and men: A picture book of invisible worlds. In: *Instinctive behavior: The development of a modern concept*, ed. & trans. C. H. Schiller. International Universities Press (1957). [DCD]

Vuust P. & Frith, C. D. (2008) Anticipation is the key to understanding music and the effects of music on emotion. *Behavioral and Brain Sciences* 31:599–600. [RSS]

Wacongne, C., Changeux, J. P. & Dehaene, S. (2012) A neuronal model of predictive coding accounting for the mismatch negativity. *Journal of Neuroscience* 32:3665–78. [TE]

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

Waydo, S., Kraskov, A., Quiroga, R. Q., Fried, I. & Koch, C. (2006) Sparse representation in the human medial temporal lobe. *Journal of Neuroscience* 26:10232–34. [TT]

Weber J. (2002) *The judgement of the eye. The metamorphoses of geometry – one of the sources of visual perception and consciousness (a further development of Gestalt psychology).* Springer. [SMS]

Weiss, Y., Simoncelli, E. P. & Adelson, E. H. (2002) Motion illusions as optimal percepts. *Nature Neuroscience* 5(6):598–604. doi:10.1038/nn858. [aAC, GB]

Wells, J. B., Christiansen, M. H., Race, D. S., Acheson, D. J. & MacDonald, M. C. (2009) Experience and sentence comprehension: Statistical learning and relative clause comprehension. *Cognitive Psychology* 58:250–71. [TAF]

Wheeler, M. (2005) *Reconstructing the cognitive world*. MIT Press. [aAC]

Wheeler, M. & Clark, A. (2009) Culture, embodiment and genes: Unravelling the triple helix. *Philosophical Transactions of the Royal Society of London, B* 363 (1509):3563–75. [aAC]

Wilson, R. A. (1994) Wide computationalism. *Mind* 103:351–72. [aAC]

Wilson, R. A. (2004) *Boundaries of the mind: The individual in the fragile sciences – cognition*. Cambridge University Press. [aAC]

Womelsdorf, T., Anton-Erxleben, K., Pieper, F. & Treue, S. (2006) Dynamic shifts of visual receptive fields in cortical area MT by spatial attention. *Nature Neuroscience* 9:1156–60. [NB]

Wu, Z. (1985) Multidimensional state space model Kalman filtering with applications to image restoration. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 33:1576–92. [DRa]

Wyart, V., Nobre. A. C. & Summerfield, C. (2012) Dissociable prior influences of signal probability and relevance on visual contrast sensitivity. *Proceedings of the National Academy of Sciences USA* 109:3593–98. [rAC, TE]

Wyss, R., König, P. & Verschure, P. F. M. J. (2004) Involving the motor system in decision making. *Proceedings of the Royal Society of London, B: Biological Sciences* 271(Suppl. 3):S50–52. [PK]

Yeshurun, Y. & Carrasco, M. (1998) Attention improves or impairs visual performance by enhancing spatial resolution. *Nature* 396:72–75. [NB]

Yu, A. J. (2007) Adaptive behavior: Humans act as Bayesian learners. *Current Biology* 17:R977–80. [aAC]

Yuille, A. & Kersten, D. (2006) Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Science* 10(7):301–308. [aAC, KF]

Zahedi, K., Ay, N. & Der, R. (2010) Higher coordination with less control – a result of information maximization in the sensorimotor loop. *Adaptive Behavior* 18(3–4):338–55. [aAC]

Zhu, Q. & Bingham, G. P. (2011) Human readiness to throw: The size-weight illusion is not an illusion when picking the best objects to throw. *Evolution and Human Behavior* 32(4):288–93. [rAC]