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.

https://doi.org/10.1017/S0140525X12000477

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

https://doi.org/10.1017/S0140525X12000477

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

https://doi.org/10.1017/S0140525X12000477

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,

https://doi.org/10.1017/S0140525X12000477

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

https://doi.org/10.1017/S0140525X12000477

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 –

https://doi.org/10.1017/S0140525X12000477

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

https://doi.org/10.1017/S0140525X12000477

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

https://doi.org/10.1017/S0140525X12000477

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

https://doi.org/10.1017/S0140525X12000477

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

https://doi.org/10.1017/S0140525X12000477

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

https://doi.org/10.1017/S0140525X12000477

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.

https://doi.org/10.1017/S0140525X12000477

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,

https://doi.org/10.1017/S0140525X12000477

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

https://doi.org/10.1017/S0140525X12000477

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

https://doi.org/10.1017/S0140525X12000477

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)

https://doi.org/10.1017/S0140525X12000477

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