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

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

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

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

Sparse coding and challenges for Bayesian models of the brain

doi:10.1017/S0140525X12002300

Thomas Trappenberg and Paul Hollensen

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

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

www.cs.dal.ca/~tt

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

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

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

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

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

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

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

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

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

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

Authors' Response

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

doi:10.1017/S0140525X12002440

Andy Clark

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

andy.clark@ed.ac.uk

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

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

R1. Introduction: Combining challenge and delight

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

R2. Perils of prediction

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