Distinguishing Theory from Implementation in Predictive Coding Accounts of Brain Function

M. W. Spratling (michael.spratling@kcl.ac.uk)
Department of Informatics, King's College London, University of London, UK.

Abstract

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

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

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 that is reviewed in section 3.1. However, they do suggest that different cortical circuitry may underlie these outward behaviours. This means that claims (repeated by Clark in section 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 section 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 and 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 and 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 prior, 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 and Ballard, 1999), or the range of attentional data accounted for by the PC/BC implementation (Spratling, 2008b) compared to the standard implementation (Feldman and 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 and 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, 2012a), as well as linking to alternative theories of brain function, such as divisive normalisation and biased competition (Spratling, 2012a, Spratling, 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.

References

Bechtel, W. (2006). Reducing psychology while maintaining its autonomy via mechanistic explanation. In Schouten, M. and de Jong, H. L., editors, The Matter of the Mind: Philosophical Essays on Psychology, Neuroscience and Reduction, chapter 8. Blackwell, Oxford, UK.

Carandini, M. (2012). From circuits to behavior: a bridge too far? Nature Neuroscience, 15(4): 507-509.

Carandini, M., Demb, J. B., Mante, V., Tolhurst, D. J., Dan, Y., Olshausen, B. A., Gallant, J. L., and Rust, N. C. (2005). Do We Know What the Early Visual System Does? Journal of Neuroscience, 25(46):10577–97.

Feldman, H. and Friston, K. J. (2010). Attention, uncertainty, and free-energy. Frontiers in Human Neuroscience, 4, 215-

Friston, K. (2005). A theory of cortical responses. Philosophical Transactions of the Royal Society of London: Series B, Biological Sciences, 360(1456):815-836

Lochmann, T. and Deneve S. (2011). Neural processing as causal inference. Current Opinion in Neurobiology, 21(5):774–78

Mareschal, D., Johnson, M. H., Siros, S., Spratling, M. W., Thomas, M. S. C., and Westermann, G. (2007). Neuroconstructivism: How the Brain Constructs Cognition. Oxford University Press, Oxford, UK.

Olshausen, B. A. and Field, D. J. (2005). How close are we to understanding V1? Neural Computation, 17:1665–99.

Phillips W. A. and Singer W. (1997). In search of common foundations for cortical computation. Behavioral and Brain Sciences, 20:657-722.

Rao, R. P. N. and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nature Neuroscience, 2(1):79–87.

Spratling, M. W. (2008a). Reconciling Predictive Coding and Biased Competition Models of Cortical Function. Frontiers in Computational Neuroscience 2(4).

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

Spratling, M. W. (2010). Predictive coding as a model of response properties in cortical area V1. Journal of Neuroscience, 30(9):3531-43.

Spratling, M. W. (2011). A single functional model accounts for the distinct properties of suppression in cortical area V1. Vision Research, 51(6):563-76.

Spratling, M. W. (2012a). Unsupervised learning of generative and discriminative weights encoding elementary image components in a predictive coding model of cortical function. Neural Computation, 24(1): 60-103.

Spratling, M. W. (2012b). Predictive coding accounts for V1 response properties recorded using reverse correlation. Biological Cybernetics, 106(1):37-49.