Computational neuroimaging: localising Greek letters?

Comment on Forstmann et al.

Jill X. O'Reilly[1] and Rogier B. Mars[1,2]

[1]Centre for Functional MRI of the Brain, University of Oxford, John Radcliffe Hospital, Oxford, OX3 9DU, UK; [2]Department of Experimental Psychology, University of Oxford, Oxford, OX1 3UD, UK

NOTICE: this is the author's version of a work that was accepted for publication in Trends in Cognitive Sciences. Changes resulting from the publishing process, such as peer review, editing, corrections, structural formatting, and other quality control mechanisms may not be reflected in this document. Changes may have been made to this work since it was submitted for publication. A definitive version was subsequently published in Trends in Cognitive Sciences, DOI#10.1016/j.tics.2011.07.012

In a recent issue of TiCS, Forstmann and colleagues [1] reviewed a relatively recent development in cognitive neuroscience: the application of mathematical models of behaviour to functional neuroimaging data. As computational neuroimaging begins to 'go mainstream', it is important to highlight a point concerning the interpretation of such studies that may seem obvious to modellers, but in our experience is a source of confusion for many readers of their papers.

Mathematical models are traditionally used by experimental psychologists to make quantitative predictions about the relationship between stimuli and behaviour. The application of mathematical models in neuroimaging is a substantial step forward because it allows brain activ-ity to be related to 'latent' variables captured in the math-ematical model; that is, variables that represent the state of an internal calculation that may not have a direct relationship with behavioural performance or stimuli, for example uncertainty or learning rate [2]. However, there is an important distinction to be made between activity correlating with a parameter and a brain area representing that parameter.

Correlation with a model parameter is generally only meaningful in the context of the entire cognitive process being modelled. For example, a prediction error captures the

difference between an internal model and sensory reality; a prediction error signal is therefore evidence that a certain brain region has access to the relevant internal model and sensory information.

By designing an experiment in such a way that the prediction error is specific to one cognitive goal or used in a particular way, modellers may localise the internal model of interest. Note that this is conceptually different from locating a 'prediction error module'. By designing an experiment in which the parameters of a mathematical model vary over time, modellers intend to create known trial-to-trial variation in the cognitive process being modelled; this variation may be modelled in functional imaging data, which in turn allows one to locate the cognitive process. For example, Behrens and colleagues [3] showed that participants learn faster in a more volatile environment, and that the learning rate ('alpha') correlated with activity in a part of anterior cingulate cortex. It might be tempting to infer that this region represents volatility or learning rate, but a more mechanistically meaningful interpretation is that the region is involved in learning itself; that is, updating of an internal model, which is done more vigorously in the high volatility and/or high learning rate condition.

In conclusion, the goal of modellers is not to achieve a functional localisation of alphas and other Greek letters to brain areas, but to test hypotheses about the underlying computational function of brain areas by up- or down-regulating that function. Indeed, solely localising parameters would be nonsensical as, in many cases, there is no assumption that the equations in mathematical models are literally implemented by neurons, only that they give a good quantitative approximation of a process that is calculated in the brain 'somehow'. This is in contrast to biophysical models [4], which seek primarily to describe the behaviour of neural networks and only secondarily to describe behaviour.

References

[1] Forstmann, B.U. et al. (2011) Reciprocal relations between cognitive neuroscience and cognitive models: opposites attract? Trends Cogn. Sci. 15, 272–279

[2] Corrado, G. and Doya, K. (2006) Understanding neural coding through the model-based analysis of decision making. J. Neurosci. 27, 8178–8180

[3] Behrens, T.E. et al. (2007) Learning the value of information in an uncertain world. Nat. Neurosci. 10, 1214–1221

[4] Wong, K.F. and Wang, X.J. (2006) A recurrent network mechanism of time integration in perceptual decisions. J. Neurosci. 26, 1314–1328