Optimal Human Reinforcement Learning in a Hierarchical Decision Task

Ulrik Beierholm[1], Klaus Wunderlich[2], Peter Bossaerts[2,3,4] and John P O'Doherty[2,3,5]

1 Gatsby Computational Neuroscience Unit, UCL, London, UK
2 Computation and Neural Systems Program, California Institute of Technology, Pasadena, CA
3 Division of Humanities and Social Sciences, California Institute of Technology, Pasadena, CA
4 Ecole Polytechnique Federale de Lausanne, Swiss Confederation
5 Trinity College Institute of Neuroscience and School of Psychology, Trinity College, Dublin, Ireland

A number of tasks have been developed for testing how human subjects learn and respond to contextual changes in their environment, e.g. versions of the Wisconsin Card Sorting Task (Drewe, 1974; Robinson, 1980) or the Stroop test (Macleod, 1991). These tasks require evaluation of information across multiple stimulus dimensions combined with learned or instructed task conditions in order to compute choice.

Physiologically, prefrontal cortex has long been known to be an essential neural structure for enabling adaptive behavioural responses to decision problems (e.g. Dias, 1996). However very little is known about how subjects solve such problems or the computational strategy the prefrontal cortex deploys.

We tested two competing computational strategies for how the subjects might solve such a problem: full "Bayesian" integration of probabilistic information across all stimulus dimensions (model averaging) versus an attentionally focused 2-layered decision strategy (model selection).

We used a task with two stimulus dimensions: colour and motion. Within each dimension there are two exemplars of each stimulus category: i.e. colour is red or green; dots moving leftward or rightward. At any given time, one dimension is "relevant", and within that dimension a particular exemplar is correct. For example, "colour" may be relevant, and within colour, "green" may be correct.

Accordingly choice of the stimulus with the relevant dimension and the correct exemplar will yield monetary rewards on a probabilistic basis (80%), whereas selection of any other stimulus will yield reward with only 20% probability.

Within a stimulus category, the correct exemplar will probabilistically switch from time to time across trials, while the relevant dimension also switches from time to time. The category switches occur on a faster time scale than the dimensional switches.

Subjects needed to establish which dimension is relevant (higher-order inference), and which exemplar within each dimension is currently rewarded (lower-order inference). Subjects were paid according to their performance and performed the task for 40 minutes while haemodynamic BOLD response was measured in a 3T MRI scanner.

A full Bayesian model for this problem is computationally intractable for the subjects and hence we approximated it using a 2-layered Reinforcement Learning model.

We found evidence that subjects' behaviour conforms better to a computational decision strategy in which subjects' use probability integration across relevant dimensions and exemplars, than to a two layered attentionally focused strategy.

Furthermore, neural activity in human prefrontal cortex was found to be better accounted for by the probability integration model than by the two-layer model. Distinct sub-regions of medial prefrontal cortex were found to correlate with the full value and the certainty within the dimension.

Our results indicate that human prefrontal cortex deploys a near optimal Reinforcement Learning ("Bayesian" like) decision strategy in which a multi-dimension decision problem is resolved by integrating optimally across dimensions to guide choice.

References:

R. Dias, T. W. Robbins, A. C. Roberts, *Nature* 380, 69 (1996).

E. A. Drewe, *Cortex* 10, 159 (1974).

C. M. MacLeod, *Psychological bulletin* 109 (2): 163–203 (1991).

A. L. Robinson, R. K. Heaton, R. A. Lehman, D. W. Stilson, *J Consult Clin Psychol* 48, 605 (1980).