Observing the observer

J. Daunizeau¹, K. E. Stephan¹, H. E. M. den Ouden¹, M. Pessiglione², and K. J. Friston¹

¹ Functional Imaging Laboratory, University College of London, London, UK
² INSERM U610, Université Pierre et Marie, Paris, France

This paper concerns the nature of evoked brain responses and the principles underlying their generation. We start with the premise that the sensory brain has evolved to represent or infer the causes of changes in its sensory inputs [1]. The problem of inference is well formulated in statistical terms. The statistical fundamentals of inference may therefore afford important constraints on neuronal implementation. Our focus is on plausible phenomenological models that can be inverted using behavioural/neurophysiological data; namely models that are useful operationally in an experimental setting and that embody a principled account of learning.

From a Bayesian perspective, the brain might be considered as an *observer* of its own sensory signals [2]. In other words, the brain dynamically inverts some forward model of its sensory inputs for representing the unobserved (hidden) causes that might have generated them. Additionally, we assume that human observers make decisions on the basis of their (posterior) states of belief regarding these hidden causes. We then rely on Bayesian decision-theoretic arguments for modelling action prescription as a goal-reaching problem in an uncertain environment. This boils down to defining the so-called utility/loss function for which observed decisions are optimal.

We argue that, as we *observe the observer*, we can interpret its behavioural/neurophysiological activity by inferring the structure of both its prior beliefs and its utility/loss function from psychophysical/physiological measures. As a consequence, two key concepts are required: (i) the definition of the class of priors and utility/loss function the brain might be using and (ii) the relationship between the brain’s posterior states of belief and the experimental psychophysical/physiological observations. Note that there is a crucial distinction between the brain’s (primary) generative model, which is generating sensory signals and the experimenter’s (secondary) generative model, which is describing the subject’s responses (e.g., reaction times, neuroimaging time series).

We then propose a principled and generic Bayesian hierarchical modelling framework for inferring the brain’s states of belief from behavioural/neurophysiological measurements during learning/decision making experiments. We illustrate the benefits of this methodology in the final section using reaction times in a simple cue-outcome associative learning task [3].

Acknowledgments
This work was supported by a Marie-Curie Intra-European fellowship program.

References