Model-based analyses: Promises, pitfalls, and example applications to the study of cognitive control

Rogier B. Mars[1,2], Nicholas J. Shea[3], Nils Kolling[4], and Matthew F. S. Rushworth[5,6]

[1]Department of Experimental Psychology, University of Oxford, Oxford, UK
[2]Centre for Functional Magnetic Resonance Imaging of the Brain (FMRIB), John Radcliffe Hospital, University of Oxford, Oxford, UK
[3]Department of Philosophy, University of Oxford, Oxford, UK
[4]Department of Experimental Psychology, University of Oxford, Oxford, UK
[5]Department of Experimental Psychology, University of Oxford, Oxford, UK
[6]Centre for Functional Magnetic Resonance Imaging of the Brain (FMRIB), John Radcliffe Hospital, University of Oxford, Oxford, UK

We discuss a recent approach to investigating cognitive control, which has the potential to deal with some of the challenges inherent in this endeavour. In a model-based approach, the researcher defines a formal, computational model that performs the task at hand and whose performance matches that of a research participant. The internal variables in such a model might then be taken as proxies for latent variables computed in the brain. We discuss the potential advantages of such an approach for the study of the neural underpinnings of cognitive control and its pitfalls, and we make explicit the assumptions underlying the interpretation of data obtained using this approach.

Keywords: Cognitive control; Model-based approach; Model; Reinforcement learning; Action regulation.

We humans can engage in a complex repertoire of behaviours geared towards often far-removed goals. We have to override reflexive and habitual reactions in order to orchestrate behaviour in accordance with our intentions. These mechanisms are commonly referred to as "cognitive control processes", and their function is to control lower level sensory, memory, and motor operations for a common purpose (Miller, 2000). Processes associated with cognitive control are often highly

dynamic and context dependent. They rely not just on the presented stimuli, but also on factors that are difficult to control and to observe for the experimenter, such as the participant's experience, trial history, motivation, and individual differences. These factors can vary on a trial-by-trial basis. This poses a number of problems for experimental scientists, who often rely on averaging data recorded over a substantial number of trials in the same circumstances to achieve a reasonable power for statistical analysis. Even though these variables are difficult to observe, it is widely accepted that the brain does use these types of variable. Thus, traditional experimental designs often allow only a limited view on the computational processes that underlie our behaviour (Corrado & Doya, 2007).

The goal of the current paper is to introduce the model-based technique for studying cognitive control as it is recently being employed in neuroimaging to the wider audience of experimental psychologists. This technique allows the researcher to circumvent some of the problems mentioned above. The paper is divided up into two parts. In the first part we discuss the model-based approach in detail and focus on general methodological and interpretational issues. In the second part, we discuss some examples of applications of the model-based approach to problems of cognitive control. We focus specifically on data obtained from experiments with human participants, using as dependent variables behaviour and measures of brain activity associated with behaviour, such as event-related brain potentials (ERPs) and the blood-oxygen-level-dependent (BOLD) signal that can be recorded using functional magnetic resonance imaging (fMRI).

The model-based approach

The solution adopted by the model-based approach is to construct an explicit computational model of the task the participant has to solve. This model should describe the transformation of stimuli to the observable behavioural responses and should contain the unobservable (i.e. latent) variables that affect this transformation. Variations in the estimated levels of the latent variables on each trial are then correlated with behaviour or neural activity (Figure 1a; see also Corrado & Doya, 2007; Corrado, Sugrue, Brown, & Newsome, 2009). The most well-known application of this type of model-based analysis is in studies of reward-based decision making, and we look at this application to describe the approach in more detail.

Behavioural studies have shown that reward-related learning depends on the predictability of the reward (Rescorla & Wagner, 1972). Studies of the monkey dopamine system have shown that activity of midbrain dopamine neurons, which project to the ventral striatum, does not simply differentiate rewarded from unrewarded events, but codes the difference between expectations of reward and the actual reward that is received (Schultz, Dayan, & Montague, 1997). This *reward prediction error* thus has a strong correlate in neural activity but its variations are difficult to observe in a standard "subtraction" design, simply comparing activity in two experimental conditions. It relies on the animal having a prediction of how rewarding a future event (e.g., action or stimulus) is likely to be, and this, in turn, is dependent on the animal's past reward history.

In the 1980s the insight that behaviour is determined by expectations about reward found its way into neural network models (Sutton & Barto, 1981). Since then, the relationship between the predictions of these models and neurophysiological activity has become a viable topic of research. These models emphasize that when behaviour is determined by expectations of reward it is crucial to ensure that the expectations are revised appropriately when a previous prediction turned out to be wrong, as evidenced by the reward prediction error. The variation of the reward prediction error on a trial-by-trial basis can be estimated in a simple reinforcement learning model (Sutton & Barto, 1998). In this class of models, the computational goal is to maximize the obtained reward. On each trial, the model makes a prediction about the value of the reward V associated with
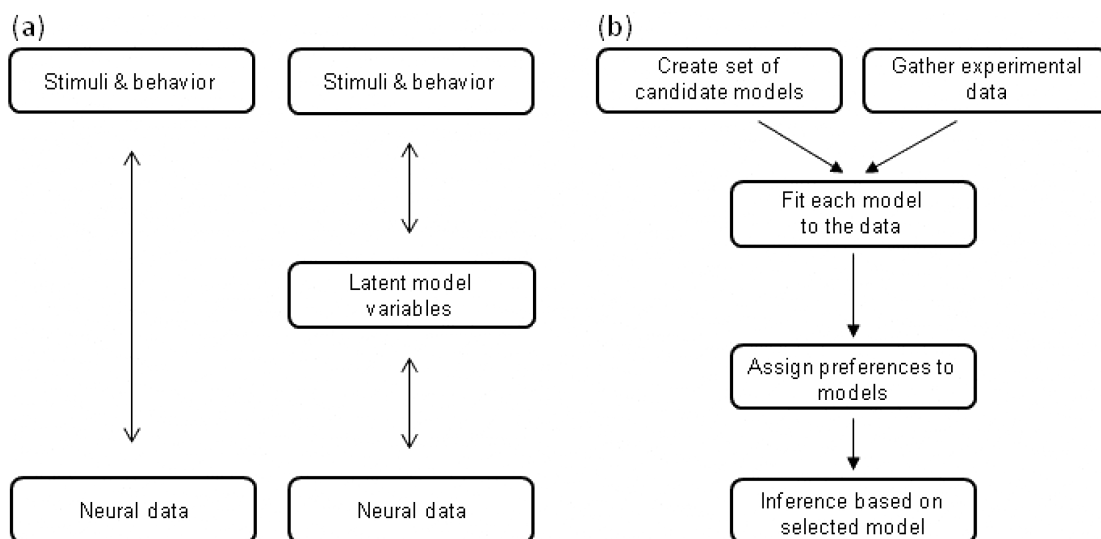
(a)

(b)

Figure 1. *(a) Approaches to data analysis. The traditional approach (left) tries to directly correlate variations in stimuli and observable behaviour to variations in neural data, while the model-based approach infers latent variables in an explicit computational model based on the observable stimuli and behaviour and, in turn, correlates these variables to neural data. From "Understanding Neural Coding through the Model-Based analysis of Decision Making", by G. Corrado and K. Doya, 2007, Journal of Neuroscience, 27, pp. 88–180. Copyright 2007 by the Society for Neuroscience. Adapted with permission. (b) Processing pipeline for model-based analysis of neuroimaging data. The experimenter formulates a set of candidate models and gathers the experimental data (e.g., behaviour; blood-oxygen-level-dependent, BOLD signals; event-related potentials, ERPs; or motor-evoked potentials). Each model is than fitted to the data, and the models are compared using some model comparison technique, such as Akaike's information criterion (AIC) or Bayes factors. Inference is then based on the model that best explains the data. Adapted from MacKay (1992).*

the state s it is in at time t, $V(s_t)$, and updates this estimate of value for the next trial, $V(s_{t+1})$, on the basis of the prediction error at time t, δ_t:

$$V(s_{t+1}) = V(s_t) + \alpha \times \delta_t \qquad (1)$$

where α is parameter determining the rate of learning, and the prediction error at t, δ_t, is defined as the difference between the actual reward R_t and the expected reward $V(s_t)$:

$$\delta_t = R_t - V(s_t) \qquad (2)$$

The model can then be used to generate values of $V(s_t)$ and δ_t for each trial. $V(s_t)$ and δ_t can then be regressed against the behavioural data or to neuroimaging data to discover areas of the brain in which the BOLD signal varies parametrically with these quantities, trial-by-trial. This approach has been used extensively, and activity that

correlates with δ_t has been found in the striatum and prefrontal and anterior cingulate cortex (see O'Doherty, Hampton, & Kim, 2007; Rushworth & Behrens, 2008, for reviews).

Fit each model to the data

Models often have a number of free parameters. Free parameters are variables that have to be set to a certain value in order for the model to be able to make predictions. According to the procedure advocated in Figure 1b (MacKay, 1992), these parameters are fitted to the data before the model is used to predict trial-by-trial fluctuations in latent variables. In the case of the reinforcement learning model described above α can be regarded as a free parameter. This learning rate can differ between individuals and different learning environments (Behrens, Woolrich, Walton, & Rushworth, 2007; M. X. Cohen, 2007). In most

studies, the learning rate is kept constant over the course of the experiment.

Parameters can be fitted to both behavioural and neural data. However, a more common strategy for the analysis of neural data, such as fMRI or ERP data, is to fit the parameter to the behavioural data and use the resulting parameters for the model, which is then fitted to the neural data. Behavioural data might be less noisy, and thus fitting the parameters to these data might be preferable. This procedure of course relies on the assumption that all neural data and the behavioural data are fitted optimally by the same parameters, an assumption that might not always hold.

Comparing models: Which model accounts for the data best?

An important limitation of the model-based approach is that it does not allow inference beyond the model tested. Thus, it is possible to have a model that describes a significant amount of variance in the neuroimaging data even though this model is not the best description of the algorithm employed by the brain. Indeed, when fitting a single model, it is only possible to find evidence in favour of this model or not. This runs against the normal practice in science, to try to disprove one's hypothesis, rather than to simply try to find evidence in favour of it. To overcome this problem, rather than just fitting one potential model, there should be a set of different candidate models of how the brain solves a particular problem. Each of these models is than fitted to the data, and a model comparison technique is used to determine with model is best supported by the data. These techniques should not simply test which model explains the most experimental variance, but also take into account the complexity of the model. Models with more free parameters are penalized, since adding an extra parameter will normally explain more variance, even if the parameter is not plausible. Inference is then based on this "best" model or a combination of the "best" models (Burnham & Anderson, 2002).

Two popular model comparison techniques are information-theoretic selection based on Kullback–Leibler information loss and model selection based on Bayes factors (see Burnham & Anderson, 2004). The first class is often represented by Akaike's information criterion (AIC; Akaike, 1973), which is an approximation of the log-evidence for a model. Bayes factors are computationally quite complex to calculate, but can be approximated by Bayesian information criterion (BIC). Both AIC and BIC are thus approximations for the true evidence in favour of a model. Based on empirical evidence, Kass and Raftery (1993) suggest that BIC is biased towards simple models and AIC to complex models. One strategy to circumvent these problems is to only consider one model in favour of another if both AIC and BIC agree. Model selection techniques have recently found their way into the analysis of neuroimaging data (e.g., Kiebel, Garrido, Moran, Chen, & Friston, 2009; Rosa, Bestmann, Harrison, & Penny, 2010).

Just as parameter fitting can be done with respect to either behavioural or neural data so can model comparison be done on the basis of just the behavioural data or the neural data. As an example of the former, in a recent study Lau and Glimcher (2005) were interested in the type of information monkeys can use to guide their decisions in a probabilistic reinforcement learning task. They formulated a number of models to predict a monkey's choice on each trial based on the past history of reinforcements and choices. A family of candidate models varied the length of reinforcement and choice histories that influenced the choice on a given trial. The authors then fitted each model separately to each monkey's data using maximum likelihood and used AIC to compare the different models. A number of recent neuroimaging studies have used a variant of the former approach by fitting different models to the behavioural data and regressing to the neuroimaging data only the parameters of the model that comes out "best" by reference to the behavioural data (e.g., Forstmann et al., 2008). Other studies directly compare models fitted to the neural data (Bestmann et al., 2008; Hampton, Bossaerts, & O'Doherty, 2006). Importantly, these studies do

not just compare a strong candidate model to an obviously implausible model, such as using a model with and without learning parameters to assess neural contributions to a learning process, but test equally valid models that could all plausibly underlie the process of interest.

Although formally comparing different models is preferable to only fitting a single model, this still leaves open the possibility that the "best" model is not part of the candidate set. Indeed, it has been argued that the number of factors influencing any type of biological data is far too great to ever allow the specification of the "true" underlying model that generates the data (Burnham & Anderson, 2002). It follows that we can only ever reach a comparative conclusion. It might be argued that this puts the model-based analysis of neuroimaging data at a significant disadvantage to standard approaches, which might be more explorative in nature. However, more explorative studies can never hope to draw strong conclusions about the nature of the computation used by a neural system.

Which models to compare?

We saw in the previous section that, when using models to analyse behavioural or neuroimaging data, instead of just fitting a single model it is preferable to compare the performance of several models with one another. However, this does not mean that the model that is supported best by the data is always the best model from the researcher's point of view. The aim of fitting models is ultimately to discover how the brain carries out the computations that lead to participants' behavioural performance. For that to succeed, quantities in the model (parameters and components) must correspond in some way to quantities being computed over in algorithms implemented in neural circuits. Some elements of the model need not map in any obvious way onto neural circuitry; indeed the model may only capture some aspects of the algorithms that are actually being computed, but when a correlation *is* found between a model component and a neural signal, that is taken as evidence that the brain implements an algorithm that involves

calculating over that component. For example, the correlation between the trial-by-trial prediction error δ in a simple reinforcement learning model and the BOLD signal in the striatum is taken as evidence that the relevant neural circuit implements an algorithm that calculates over trial-by-trial prediction errors (or something like them), amongst other quantities. In this section, we make a number of points that need to be considered for a model-based analysis and accompanying model comparison to successfully lead to insight into neural processing.

First, although model comparison techniques like AIC do penalize models for having more free parameters, they take no account of how likely it is that components of the model reflect aspects of the mechanism being studied (they are not designed to). Given free choice from the wide range of models that are a priori possible, the one that comes out best on the basis of model comparison might nevertheless have little empirical plausibility. The point is that model-based analysis should not be driven entirely by the specific data set in this way but instead it should reflect background knowledge about such factors as the model's anatomical plausibility and its success in predicting performance on related tasks. For example, Friston (2003) has emphasized the relationship between aspects of anatomy and certain computational models, and Rushworth and colleagues (Rushworth, Mars, & Summerfield, 2009) argued for particular computational perspectives into visual and social learning partly because similar concepts have already proved useful in another domain—that is, that of reinforcement learning. Apart from using prior information in the formulation of models, it is possible to bias the model comparison process by means of formulating priors in order to give more weight to evidence in favour of certain models. Notice that this makes the model-based analysis approach more hypothesis driven, and less prone to overfitting of the data, than is sometimes assumed. We have already alluded to the importance of choosing plausible models for the comparison. Comparing a model only with

respect to a number of neutrally implausible models does not lead to novel insight.

A problem arises when several plausible candidate models cannot be effectively distinguished by standard procedures for model comparison. For instance, the prediction error discussed above is under some circumstances similar to the "surprise" or Shannon's information (Mars et al., 2008; Shannon, 1948) of a stimulus. If the models are very similar to one another—that is, when two different algorithms make very similar predications regarding the neural data—then there is a danger that the result of a model comparison will be driven by peculiarities of a particular data set, rather than facts about which model best captures the algorithms being computed. So the general conclusion can be made much more confidently than the specific one. Once compared to the range of alternative models that seem plausible given background knowledge, it is reasonable to conclude that the brain implements an algorithm that computes over something like surprise or prediction error—that some quantity of this general sort is one of the decision variables deployed in whatever algorithm is in play. A more specific conclusion, for example preferring a prediction error model over a surprise model, is necessarily much more tentative. One can argue that the problem of highly correlated predictions is precisely why the model-based approach is employed in the first place, since standard subtraction designs often predict the same pattern of behaviour or neural activity in a number of different comparisons. However, this does not mean that the model-based approach is immune to the problem of correlated predictions, and researchers should be careful in the conclusion they draw.

A third issue that arises is whether each component of a model should be expected to be found in neuroimaging data. For instance, although most studies on reinforcement learning focus on the prediction error, Behrens et al. (2007) focused on the learning rate α, and some authors have focused on the representation of the value weights (see M. X. Cohen, 2008). However, typically not all components of a model will be found in the neuroimaging data. In most reinforcement learning studies, for example, neural signals related directly to the representation of the probability of each stimulus are not reported—indeed, they are not searched for. Rather, these studies only search for signals correlating with a computation performed on these estimated probabilities: the prediction error. We do not feel that the fact that not all of a model's components are represented in observable neural signals is a weakness of the model-based approach. Indeed, one of the advantages of a model-based approach is that models can abstract away from details to capture a general insight. In this respect, there is a disadvantage to making the model more complicated, which is that it may deliver less insight about which processes are most important. One aspect of this trade-off is reflected in model comparison tests like the AIC, which penalize free parameters; but such tests do not directly reflect the value of having a model that fails to fully describe the data, but which seems to capture an important underlying feature of the phenomenon. Asking which model best fits the data can obscure the more important point, which is that they both predict a lot of the variance in the data set, they both have good background plausibility, and they share a key structural feature in the similarity between surprise and prediction error.

Recent work in philosophy of science reinforces this point (Sterrett, 2002; Weisberg, 2007). One of the merits of model-based science is that practitioners can assess how well a model fits reality and can work on the internal coherence of the model itself, without having very much idea how some aspects of the model map onto features of the phenomenon being modelled. Contrast a theory or other direct representation of a phenomenon. Various properties are known about and measured, and their relationships are investigated. Boyle's law is a theory about how the pressure and volume of a gas are related. It gains empirical support by measuring pressure, measuring volume, and observing the relation between these quantities. By contrast, the reinforcement learning model gained some support from its fit to behavioural data, even before anyone had any idea how

components like prediction error and learning rate mapped onto the neural mechanisms carrying out actual computations. The fact that the modeller can remain neutral about how features of a model are reflected in properties of the target system can be seen as an advantage of the strategy of model-based science (Godfrey-Smith, 2006), although the model's neural plausibility is important, as emphasized above. That advantage is undercut by a premature attempt to achieve a best fit between neuroimaging data and all the components of a very specific model.

Corrado et al. (2009) reflect the same motivation in their agenda-setting paper on the model-based analysis of decision making. Rejecting the idea that models should be taken as literal hypotheses about neural computations and embracing the notion that the computations carried out in neural circuits may differ from the details of the model, they say:

> This conservative stance frees us from the necessity of demonstrating that all elements of the model are plausibly implemented in the brain, and instead allows us to focus on our primary objective—identifying neural correlates of the key decision variables. (p. 467)

The remark is puzzling, since the strategy would only work in finding neural correlates if the key decision variables over which the model quantifies, or something similar, are indeed being computed in neural circuits. Rather than disclaiming any mapping relation between the model and the neural system that is its target, their remark is perhaps reflecting the fruitful neutrality that modelling permits about the relation between model and target. That allows Corrado et al. to remain neutral about how, or even whether, various aspects of a model will map onto computations performed in neural circuits, while still discovering how other aspects of the model (e.g., prediction errors) map onto the target system.

Interpretational issues: Neural algorithms and neural signals

Informed by background knowledge about likely mechanisms, a comparative model-based analysis can yield conclusions about the class of algorithms that it is likely that the brain uses in performing a given task, including identifying neural structures or circuits that are involved in representing some of the quantities over which the algorithms compute. Model-based analysis can remain neutral about how, or even whether, some components of the model are realized in neural algorithms, while gathering evidence about the neural implementation of others. In this section we draw out and make explicit the assumption about how such algorithms are implemented in the brain, which underpins this inference.

Marr (1982) distinguished between computational, algorithmic, and implementational levels of analysis of a system. The computational level sets out some goal or function that is to be performed and outlines the logic or strategy by which it is carried out. For example, in reward-guided decision making the problem is to take as input a series of stimuli and outcomes and to produce as output the series of actions that maximize rewarding outcomes. A computational theory might further specify that the system chooses its actions based on the history of reinforcement of those actions. At the algorithmic level, a particular way of performing this computation is specified, for example in the algorithm for simple reinforcement learning, which is given by Equations 1 and 2 above.[1] However, a problem now arises on any view, including Marr's, since the algorithm itself is multiply realizable in physical structures. Put in another way, there are many ways of implementing the same algorithm in neural circuits. An interesting example comes from the cognitive control literature, where a popular model of anterior cingulate cortex function models the activity in this brain region as the product of activity in neurons representing difference responses (Botvinick, Braver, Barch, Carter, & Cohen, 2001). Although it is unlikely that neurons literally compute a product, at the algorithmic level this model has proven extremely successful.

---

[1] Here, we consider the distinction between computation and algorithm heuristically useful, rather than suggesting that one can be drawn precisely in each case.

Multiple realizability emerges in many guises. The same algorithm could be realized in different neural circuits in different species, in differ individuals of the same species, in the same individual at different times through neural plasticity and learning, and in the same individual in different contexts. Nevertheless, we do find evidence of commonalities in the way various decision-relevant parameters are represented: The same neural circuits are often involved in a range of species (e.g., prediction errors in dopamine neuron-rich regions of the brain such as the ventral tegmental area in rats, monkeys, and humans; D'Ardenne, McClure, Nystrom, & Cohen, 2008; Roesch, Calu, & Schoenbaum, 2007; Schultz et al., 1997), in a range of individuals, and within a given individual during the course of an experiment. When a neural correlate of a model parameter or component is found, then that vindicates the assumption against multiple realizability. For example, it has been found that the covert decision to add rather than subtract a pair of numbers that are yet to be viewed is sufficiently stably realized within an individual that their covert intention can be decoded from activity in medial and lateral prefrontal cortex (Haynes et al., 2007). The correlation between a quantity in the algorithm and a neural signal is evidence that some neural algorithm is being realized that computes over that quantity (or something like it) and that each time the quantity is neurally represented, there is some detectable similarity in the specific pattern of brain activation that is produced. For instance, prediction errors could be multiply realizable within an individual in a way that would make them undetectable to fMRI. But it turns out that they are not: They are reflected in the BOLD signal. That result means that a whole series of nested assumptions can be maintained:

- Representations of prediction errors rely on a common neural circuit each time they are computed in the brain.
- The subset of neurons in that circuit involved in representing prediction errors make a detectable difference to cerebral blood flow, hence BOLD signal, against the background of other factors capable of making a difference, which are kept fixed or varied randomly between conditions.
- The BOLD signal relates quantitatively to represented prediction errors in an approximately linear relationship.

The key point for our purposes is that, as a way of discovering the algorithms computed by the brain, the sensitivity of this method is superior to its specificity. If prediction errors were represented in some other way, for example by phase coding instead of rate coding, then it would be impossible to find evidence of them by recessing a model against the BOLD signal. But that would not licence a conclusion against such features of the model being represented and computed over. It would just mean that the neural signal we are using happens not to carry evidence of them.

Given, then, that a particular algorithm is implemented the same way in each case, the question arises whether we can say anything more about this implementation. It is important to realize that the system-level neuroimaging methods discussed here do not have power to make inferences at the level of individual neurons. Although we can conclude that the algorithm is realized the same way every time around, we cannot say with any confidence whether the relevant parameter is computed at the neural level in the same way as the researcher's model does. Additionally, one has to be careful to conclude that the component of the algorithm detected is actually computed in the regions identified, as BOLD seems to correlate better with the afferent input to a brain region and interneural activity within a regions, rather than its spiking output activity (Logothetis, Pauls, Auguth, Trinath, & Oeltermann, 2001).

Applications of the model-based approach to cognitive control

Beyond neuroimaging of reward-based decision making

The model-based approach outlined above has recently started to be employed outside the context of reward-based decision making in

which it became popular; both in terms of the methodology and in the terminology (see Rushworth et al., 2009). For instance, model-based approaches have recently been applied to study trial-by-trial modulations in behaviour and neural activity during motor preparation (Bestmann et al., 2008), associative learning (Den Ouden, Friston, Daw, McIntosh, & Stephan, 2009), and social interactions (Behrens, Hunt, Woolrich, & Rushworth, 2008). Consequently, it has been suggested that neuroimaging is moving from a strict focus on localization of function ("where") to more "how" type questions (Dolan, 2008).

Researchers interested in cognitive control have often been at the forefront of employing computational models to understand brain function (e.g., Botvinick et al., 2001; Brown, Reynolds, & Braver, 2007; J. D. Cohen, Dunbar, & McClelland, 1990; Gilbert & Shallice, 2002; Holroyd & Coles, 2002; Yu & Dayan, 2005). However, although in the past investigators have tended to test whether the overall distribution of the data is as predicted by the model, they have rarely used trial-by-trial evaluations and formal model comparisons, as in some of the reward-based decision-making studies discussed above. Furthermore, although this model-based approach is becoming increasingly popular in fMRI studies, it has not yet achieved widespread application in studies employing more traditional psychophysiological methods, such as ERPs and motor-evoked potentials. As a case in point, reinforcement learning models have recently become a valuable tool for describing the behaviour of the error-related negativity, an ERP component associated with the processing of errors and subsequent behavioural adjustments (Holroyd & Coles, 2002; Holroyd, Nieuwenhuis, Mars, & Coles, 2004), but these studies mostly compare model predications and data on a qualitative basis (e.g., M. X. Cohen & Ranganath, 2007; Holroyd & Coles, 2002; Nieuwenhuis et al., 2002).

Recently, however, a more formal model-based approach has started to be applied to psychophysiological data, such as data obtained using transcranial magnetic stimulation (Bestmann et al.,

2008) and event-related brain potentials (Kiebel et al., 2009; Mars et al., 2008). In this section, we describe a number of example applications of the model-based approach to the study of cognitive control. The examples focus on behavioural, neuroimaging, and psychophysiological data and illustrate some of the issues discussed above regarding model fitting, model comparison, and interpretation.

Cognitive control and the anterior cingulate cortex
The anterior cingulate cortex (ACC) of the human brain is one of the primary foci of researchers interested in the neural correlates of cognitive control (Botvinick et al., 2001; Ridderinkhof, Ullsperger, Crone, & Nieuwenhuis, 2004; Walton & Mars, 2007). Adding to its popularity, the ACC has been suggested to be the source of two ERPs that are quite prominent in the study of cognitive control: the error-related negativity (ERN; Dehaene, Posner, & Tucker, 1994) and the N2 (Van Veen & Carter, 2002). The ACC is activated by a wide range of cognitive demands (Duncan & Owen, 2000) and is a prominent "blob" in many neuroimaging studies, which makes it difficult to ascribe any specific computational function to it. Indeed, contrasts showing activity in the ACC have been claimed to rely on processes as diverse as conflict detection (Botvinick et al., 2001), performance monitoring (Ridderinkhof et al., 2004), action selection (Holroyd et al., 2004), and autonomous functions (Critchley et al., 2003).

As described above, the ERN has been suggested to reflect the prediction error δ in reinforcement learning models (Equation 2, this study; Holroyd & Coles, 2002). Indeed, reinforcement learning models, and especially prediction errors, have been used extensively to describe activity in the ACC (Rushworth & Behrens, 2008). Taking this approach a step further, Behrens et al. (2007) recently investigated processes related to the learning rate parameter in reinforcement learning models, α (Equation 1). As discussed above, the learning rate is typically set by the experimenter or estimated from the behavioural data and then kept constant. Behrens et al. (2007) introduced a further level of

sophistication, arguing that the learning rate should be dependent on the rate with which the statistics of the environment change. The best estimate of the reward value of what is going to happen next is based on using the most relevant information available. Therefore, in a very stable environment that changes only slowly, participants should not just consider the outcomes of their most recent decisions but they should also consider historically distant information. Undue consideration should not be given to the most recent outcome, and estimates of future reward should not be dramatically changed if there is only a single surprising outcome. By contrast, in a very variable environment historically distant events are a poor guide to what will happen next, and relatively more weight should be given to information signalled by the most recent outcome. Thus, when the environment is stable it is preferable to have a lower learning rate, whereas when the environment is highly changeable or very uncertain only recent trials should be considered, and the learning rate should be higher.

The authors developed an explicit model of how participants integrated past information to select their choices on any given trial. Rather than fixing the learning rate α they constructed the model to not only update the value estimate V (Equation 1), but also its "trust" in the consistency of the environment. This model predicted behaviour better than competing reinforcement learning models that had fixed learning rates, suggesting that the brain indeed uses this parameter in its decision making. The model taking into account the consistency of the environment was even superior to models that had a separate, but fixed, α for stable and volatile periods. Subsequently, the authors showed that the volatility of the environment correlated with the BOLD signal in the anterior cingulate cortex. Although there are numerous studies that suggest a role for the anterior cingulate cortex in reward-based learning (e.g., Hester, Barre, Murphy, Silk, & Mattingley, 2008; Mars et al., 2005), this type of inference could only be made with a model-based analysis. Thus, the model-based approach is able to show that activity in the ACC covaries with the

dynamics of certain computational variables even though conventional subtraction designs fail to do so. Formal model comparison allows the research to test which model provides the best description of the data.

Task switching in probabilistic environments

As a further illustration of a potential application of the model-based approach to cognitive control, we can consider data from a task-switching paradigm. Task switching is one of the most well-established paradigms in the study of cognitive control (Monsell, 2003), requiring participants to switch from the implementation of one rule to respond to environmental stimuli to another rule. Following the results obtained by Behrens et al. (2007) and discussed above, we suggested that when participants are trying to learn which rule is appropriate in a given environment the influence of previous trials on the choice of rule might depend on the rate of change of the environment. We asked participants to perform a classification task in which stimuli had to be categorized based on either their shape or their colour (Figure 2a). The correct rule was determined probabilistically, requiring participants to switch between using the different classification rules. Critically, the probabilities changed either infrequently (stable environment) or more frequently (volatile environment). Indeed, we found that in the stable environment participants took into account trials going back further in the past, while in the volatile environment only very recent trials were taken into account (Figure 2c). By fitting a reinforcement learning model to the participants' behavioural data, we then determined the learning rate α for each participant in the different environments. The learning rate was significantly modulated by the volatility of the environment, with a larger α, indicating more influence of the most recent trials, in the highly volatile environment (Figure 2d). Additionally, a model such as the one described by Behrens et al. (2007, see above) provided a better explanation of the data than models with a stationary value of α. This example demonstrates that the model-based approach to data can be successfully
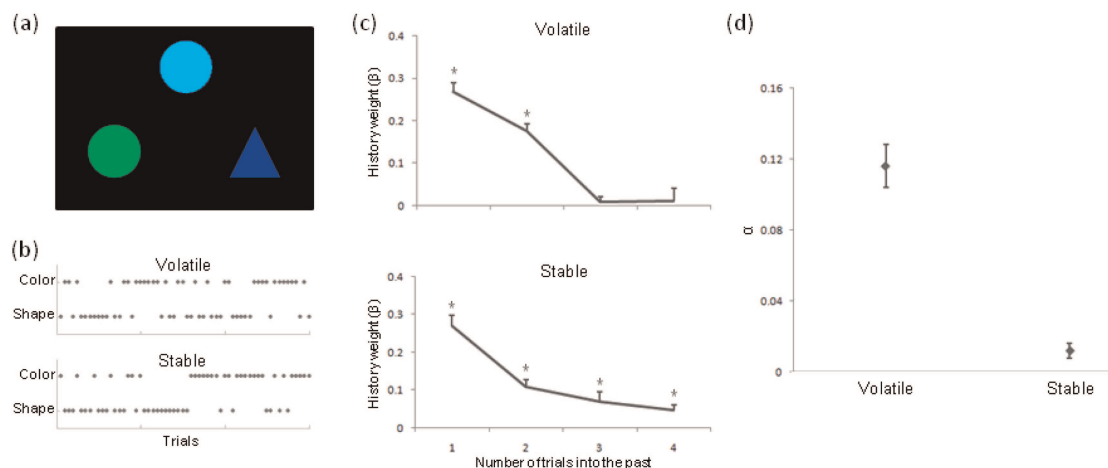
Figure 2. *Application of model-based analysis to examine the influence of the volatility of the environment on behaviour during task switching (N. Kolling, R. B. Mars, & M. F. S. Rushworth, personal communication, September 2008). (a) On each trial participants had to match a top stimulus to one of two bottom stimuli based on either the shape or the colour. (b) Which of these sorting criteria was correct was determined probabilistically. Critically, the rate at which the probabilities change differed between volatile (top) and stable (bottom) phases of the experiment. Figure indicates which rule was correct on any given trial. (c) Participants took into account a greater number of trials into the past in determining their selection criteria on any given trial in the volatile (top) than in the stable (phase) phase as determined by a regression analysis determining the influence of past trials on current trial criterion. *Indicates significant influence of this trial on the selected criterion on the current trial. (d) A reinforcement learning type model was fitted to each participant's data allowing an estimation of each participant's learning rate α in both the volatile and the stable conditions. As predicted, α was greater in the volatile than in the stable phase, indicating a greater effect of more recent trials. To view a colour version of this figure, please see the online issue of the Journal.*

applied to paradigms employed in the study of cognitive control and, moreover, that similar models might be applicable in different cognitive functions.

Balancing speed and accuracy

The speed–accuracy trade-off is one of the hallmarks of action control. It refers to the balance between the competing demands of deliberate choice and response or decision speed. The processes responsible for determining this balance have been studied extensively, resulting in a large behavioural literature and a range of formal models. Generally, these models take the form of an accumulator model, which commits to a decision when the evidence in favour of a particular response reaches a certain threshold. This type of modelling is consistent with known physiological data (Gold & Shadlen, 2007). The speed–accuracy balance is then implemented by raising or lowering the response threshold. Even though

there is a large body of work investigating the mechanisms of the speed–accuracy trade-off, little is known about its neural basis. This is partly due to the fact that simple contrasts of conservative versus fast responses are difficult to control for confounding factors, such as difficulty and attention.

In a recent study, Forstmann et al. (2008) attempted to circumvent these problems by using a model-based approach to study the neural substrates of the speed–accuracy trade-off. They used a version of the popular accumulator models, the linear ballistic accumulator (Brown & Heathcote, 2008). They fitted the model to each participant's behavioural data, estimating a parameter representing each individual's response threshold. BIC was used to determine that a model varying only this parameter allowed the best description of the data. The resulting parameter was then used as a covariate in a fMRI analysis of the same participant's BOLD data

while performing the task. They then showed that a measure related to the estimated response parameter, the so-called "response caution", correlated negatively with BOLD signal strength in the pre-SMA (pre-supplementary motor area) and striatum. This suggests that participants who decreased their response threshold more in response to demands for faster responding had more activity in these two brain areas. These results thus show that the striatum has a role in adjusting response caution. Although this had previously been assumed by a number of computational models, this has proven difficult to determine experimentally. The combined use of computational models, fitted to behavioural data, and fMRI were, however, able to address this question.

P300: A model-based analysis of the event-related brain potential

Our final example concerns a model-based analysis of the P300 component of the ERP. Although one might argue that the P300 is not a component traditionally studied in the realm of cognitive control, it has been associated with processes such as allocation of attention (Nieuwenhuis, Aston-Jones, & Cohen, 2005) and updating of the brain's internal models (Donchin & Coles, 1988), both of which are relevant to the study of cognitive control. Furthermore, the discussed study is one of only very few studies applying the model-based approach to trial-by-trial fluctuations in ERPs. Given the large application of ERPs to the study of cognitive control, we believe this makes this example particularly useful for researchers interested in cognitive control.

Mars and colleagues (Mars et al., 2008) were interested in studying the factors modulating the amplitude of the P300 component of the human ERP. The extensive literature on P300 suggests that this component is involved in the updating of a contextual representation. This has been studied in a number of tasks, but the most famous example is in the context of oddball tasks, in which participants have to respond to a number of stimuli presented in succession. In these tasks, some of the stimuli are presented only rarely. The typical result is that participants

respond slower and make more errors in response to the infrequent stimuli. Neurally, the infrequent response is associated with an enlarged P300 as compared to the frequent stimuli (Duncan-Johnson & Donchin, 1977).

To investigate the role of the P300 in this so-called context updating more closely, Mars et al. (2008) asked participants to perform a simple learning oddball task (Figure 3a). Participants were first trained on the associations between four visual stimuli and four manual responses (button presses). Following this training, participants performed blocks in which the relative probability of the stimuli was manipulated. Participants were not informed of this manipulation, but were simply instructed to respond to each stimulus with the appropriate button press as quickly as possible. Context-updating models suggest that participants maintain a representation of the probability of each stimulus that is updated on each trial based on the stimuli encountered. Accordingly, the authors constructed a simple model, which treated participants as ideal observers who tried to learn the probability distribution of the stimuli presented. They assumed that each participant started with equal prior—that is, that all stimuli were equally probable—and updated the probability of each stimulus, $p(x_i)$, on each trial:

$$p(x_i) = \frac{n_i^j + 1}{\sum_{k=1}^{4} n_k^j + 1} \quad (3)$$

where n_i^j refers to the number of occurrences of outcome i up until observation j, and the summation is over all k stimuli. From these estimated probabilities, they calculated on each trial the unpredictability or "surprise" of the presented stimuli (Figure 3b). Formally, the surprise I was quantified as the Shannon (1948) information carried by the stimulus x_i:

$$I(x_i) = -log_2 p(x_i) \quad (4)$$

Thus, if there had been very few occurrences of a stimulus x_i the information or surprise associated
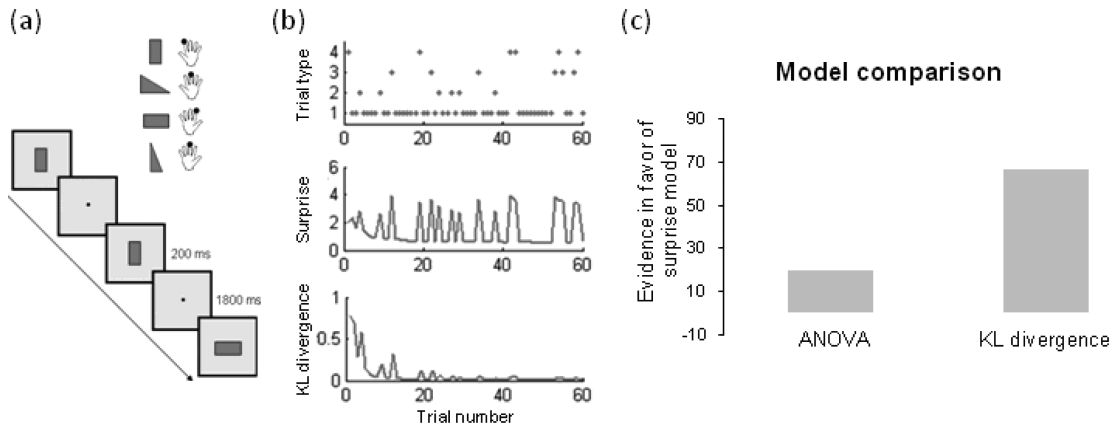
Figure 3. *Example of model-based analysis of event-related potential (ERP) data in a choice reaction time task. (a) Participants responded with button presses to visual stimuli presented every 2 seconds. (b) Manipulating the occurrence of different trials (top) allows calculation of surprise/Shannon information (middle) and Kullback−Leibler (KL) divergence (bottom) on a trial-by-trial basis. (c) Model comparison shows that surprise/Shannon information provides the best description of the data. Positive score on y-axis indicates evidence in favour of the surprise model as compared to a traditional analysis of variance (ANOVA) model and the KL divergence. From "Trial-by-Trial Fluctuations in the Event-Related Electroencephalogram Reflect Dynamic Changes in the Degree of Surprise", by R. B. Mars et al., 2008,* Journal of Neuroscience, 28, *pp. 12539−12545. Copyright 2008 by the Name of Copyright Holder. Adapted with permission.*

with an occurrence of that stimulus x_i would be high. Following the suggestion that it is these "surprising" stimuli that require additional attentional processing and updating of ongoing task processing it was hypothesized that this surprise would predict trial-by-trial changes in P300 amplitude. Rather than sorting trials by a prior frequency, as determined by the experimenter, this approach thus predicts trial-by-trial fluctuations in P300 amplitude based on the participant's internal, unobservable estimation of the stimulus probabilities. This simple model captures a number of features of the P300. For instance, the P300 to the second of two successive oddball stimuli is smaller than the P300 to the first oddball stimulus because the participant has updated the estimated probability of the occurrence of this stimulus following the presentation of the first oddball.

In the example study discussed above, Mars et al. (2008) tested not only surprise as defined in Equation 4, but also an alternative measure of context updating, the Kullback−Leibler (KL) divergence, which is a measure of the difference between the participant's estimated probability

distribution before and after trial j:

$$KL = \sum_{i=1}^{4} p_i^j \ln \frac{p_i^j}{p_i^{j-1}} \qquad (5)$$

Bayesian model comparison was then used to compare these two candidate models and a third model, a traditional analysis of variance (ANOVA) model in which the trials were sorted by a priori stimulus category by the experimenter. As shown in Figure 3c, the surprise model provided the best description of the ERP data, better than the alternative KL divergence model and better than a traditional ANOVA ("subtraction design") model.

This example shows the feasibility of applying the model-based analysis advocated in this paper to more traditional psychophysiological measures, such as the event-related brain potential. Additionally, it provides a possible route to making the rather general theories associated with some ERP components, in this case the P300, more explicit and to formally test them.

CONCLUSION

We have discussed the use of model-based analyses of behavioural and neuroimaging data. These types of analyses are becoming more and more widespread, finding their way into not only the study of reward-based decision making, but also visual perception, motor preparation, and cognitive control. They offer the exciting possibility of investigating aspects of neural processing that are not directly observable using standard "subtraction" type designs. This approach might be particularly useful in the study of the neural substrates of cognitive control given the dependence of these processes on context, participants' expectancy, and trial-by-trial dependence. Moreover, they are being applied to an ever-extending range of neuroimaging techniques, using ever more sophisticated techniques.

However, this technique relies on a number of underlying assumptions. We have tried to make these assumptions more explicit and provide some pointers on the types of interference that can be reliably drawn from model-based analyses. We have illustrated a number of issues associated with these types of analyses. We have advocated a model comparison approach, in which a number of plausible models are tested against one another. The candidate models should be plausible, for example based on the underlying anatomy or proven ability of the model in other domains. However, even given these precautions, the inference is limited to the set of candidate models. An important problem is the level of description of this approach. Although this technique can differentiate between different algorithms that might be computed by the brain, it cannot tell us directly how an algorithm is implemented.

REFERENCES

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest, Hungary: Akademiai Kiado.

Behrens, T. E. J., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. S. (2008). Associative learning of social value. *Nature, 456,* 245–250.

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience, 10,* 1214–1221.

Bestmann, S., Harrison, L. M., Blankenburg, F., Mars, R. B., Haggard, P., Friston, K. J., et al. (2008). Influences of contextual uncertainty and surprise on human corticospinal excitability during preparation for action. *Current Biology, 18,* 775–780.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review, 108,* 624–652.

Brown, S. D., & Heathcote, A. J. (2008). The simplest complete model of choice reaction time: Linear ballistic accumulation. *Cognitive Psychology, 57,* 153–178.

Brown, J. W., Reynolds, J. R., & Braver, T. S. (2007). A computational model of fractionated conflict-control mechanisms in task-switching. *Cognitive Psychology, 55,* 37–85.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multi-model inference. A practical information-theoretic approach* (2nd ed.). New York: Springer.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods Research, 33,* 261–304.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing account of the Stroop effect. *Psychological Review, 97,* 332–361.

Cohen, M. X. (2007). Individual differences and the neural representations of reward expectation and reward prediction errors. *Social Cognitive and Affective Neuroscience, 2,* 20–30.

Cohen, M. X. (2008). Neurocomputational mechanisms of reinforcement-guided learning in humans: A review. *Cognitive, Affective, and Behavioral Neuroscience, 8,* 113–125.

Cohen, M. X., & Ranganath, C. (2007). Reinforcement learning signals predict future decisions. *Journal of Neuroscience, 27,* 371–378.

Corrado, G., & Doya, K. (2007). Understanding neural coding through the model-based analysis of decision making. *Journal of Neuroscience, 27,* 8178–8180.

Corrado, G. S., Sugrue, L. P., Brown, J. R., & Newsome, W. T. (2009). The trouble with choice: Studying decision variables in the brain. In P. W. Glimcher, C. F. Camerer, E. Fehr, & R. A. Poldrack (Eds.), *Neuroeconomics: Decision making and the brain* (pp. 463–480). Amsterdam: Elsevier.

Critchley, H. D., Mathias, C. J., Josephs, O., O'Doherty, J., Zanini, S., Dewar, B. K., et al. (2003). Human cingulate cortex and autonomic control: Converging neuroimaging and clinical evidence. *Brain*, 126, 2139–2152.

D'Ardenne, K., McClure, S. M., Nystrom, L. E., & Cohen, J. D. (2008). BOLD responses reflecting dopaminergic signals in the human ventral tegmental area. *Science*, 319, 1264–1267.

Dehaene, S., Posner, M. I., & Tucker, D. M. (1994). Localization of a neural system for error detection and compensation. *Psychological Science*, 5, 303–305.

Den Ouden, H. E. M., Friston, K. J., Daw, N. D., McIntosh, A. R., & Stephan, K. E. (2009). A dual role for prediction error in associative learning. *Cerebral Cortex*, 19, 1175–1185.

Dolan, R. J. (2008). Neuroimaging of cognition: Past, present, and future. *Neuron*, 60, 496–502.

Donchin, E., & Coles, M. G. H. (1988). Is the P300 component a manifestation of context updating? *Behavioral and Brain Science*, 11, 357–374.

Duncan, J., & Owen, A. M. (2000). Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends in Neurosciences*, 23, 475–483.

Duncan-Johnson, C., & Donchin, E. (1977). On quantifying surprise: The variation of event-related brain potentials with subjective probability. *Psychophysiology*, 14, 456–467.

Forstmann, B. U., Dutilh, G., Brown, S., Neumann, J., Von Cramon, D. Y., Ridderinkhof, K. R., et al. (2008). Striatum and pre-SMA facilitate decision-making under time pressure. *Proceedings of the National Academy of Sciences USA*, 105, 17538–17542.

Friston, K. (2003). Learning and inference in the brain. *Neural Networks*, 16, 1325–1352.

Gilbert, S. J., & Shallice, T. (2002). Task switching: A PDP model. *Cognitive Psychology*, 44, 297–337.

Godfrey-Smith, P. (2006). The strategy of model-based science. *Biology and Philosophy*, 21, 725–740.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Reviews of Neuroscience*, 30, 535–574.

Hampton, A. N., Bossaerts, P., & O'Doherty, J. P. (2006). The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *Journal of Neuroscience*, 26, 8360–8367.

Haynes, J. D., Sakai, K., Rees, G., Gilbert, S., Frith, C., & Passingham, R. E. (2007). Reading hidden intentions in the human brain. *Current Biology*, 17, 323–328.

Hester, R., Barre, N., Murphy, K., Silk, T. J., & Mattingley, J. B. (2008). Human medial frontal cortex activity predicts learning from errors. *Cerebral Cortex*, 18, 1933–1940.

Holroyd, C. B., & Coles, M. G. H. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, 109, 679–709.

Holroyd, C. B., Nieuwenhuis, S., Mars, R. B., & Coles, M.G. H. (2004). Anterior cingulate cortex, selection for action, and error processing. In M.I. Posner (Ed.), *Cognitive neuroscience of attention* (pp. 219–231). New York: Guilford Press.

Kass, R. E., & Raftery, A. E. (1993). *Bayes factors and model uncertainty*, (Tech. Rep. No. 254). University of Washington, Seattle, WA.

Kiebel, S. J., Garrido, M. I., Moran, R., Chen, C. C., & Friston, K. J. (2009). Dynamic causal modelling for EEG and MEG. *Human Brain Mapping*, 20, 1866–1876.

Lau, B., & Glimcher, P. W. (2005). Dynamic response-by-response models of matching behaviour in rhesus monkeys. *Journal of the Experimental Analysis of Behavior*, 84, 555–579.

Logothetis, N. K., Pauls, J., Auguth, M., Trinath, T., & Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. *Nature*, 412, 150–152.

MacKay, D. J. C. (1992). Bayesian interpolation. *Neural Computation*, 4, 415–447.

Marr, D. (1982). *Vision. A computational investigation into the human representation and processing of visual information*. New York: W.H. Freeman and Company.

Mars, R. B., Coles, M. G. H., Grol, M. J., Holroyd, C. B., Nieuwenhuis, S., Hulstijn, W., et al. (2005). Neural dynamics of error processing in human medial frontal cortex. *NeuroImage*, 28, 1007–1013.

Mars, R. B., Debener, S., Gladwin, T. E., Harrison, L. M., Haggard, P., Rothwell, J. C., et al. (2008). Trial-by-trial fluctuations in the event-related electro-encephalogram reflect dynamic changes in the degree of surprise. *Journal of Neuroscience*, 28, 12539–12545.

Miller, E. K. (2000). The prefrontal cortex and cognitive control. *Nature Reviews Neuroscience*, *1,* 59–65.

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, *7,* 134–140.

Nieuwenhuis, S., Aston-Jones, G., & Cohen, J. D. (2005). Decision making, the P3, and the locus coeruleus-norepinephrine system. *Psychological Bulletin*, *131,* 510–532.

Nieuwenhuis, S., Ridderinkhof, K. R., Talsma, D., Coles, M. G. H., Holroyd, C. B., Kok, A., et al. (2002). A computational account of altered error processing in older age: Dopamine and the error-related negativity. *Cognitive, Affective, and Behavioral Neuroscience*, *2,* 19–36.

O'Doherty, J. P., Hampton, A., & Kim, H. (2007). Model-based fMRI and its application to reward learning and decision making. *Annals of the New York Academy of Sciences*, *1104,* 35–53.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning. II: Current research and theory* (pp. 64–99). New York: Appleton-Century-Crofts.

Ridderinkhof, K. R., Ullsperger, M., Crone, E. A., & Nieuwenhuis, S. (2004). The role of the medial frontal cortex in cognitive control. *Science*, *306,* 443–447.

Roesch, M. R., Calu, D. J., & Schoenbaum, G. (2007). Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. *Nature Neuroscience*, *10,* 1615–1624.

Rosa, M. J., Bestmann, S., Harrison, L., & Penny, W. (2010). Bayesian model selection maps for group studies. *NeuroImage*, *49,* 217–224.

Rushworth, M. F. S., & Behrens, T. E. J. (2008). Choice, uncertainty and value in prefrontal and cingulate cortex. *Nature Neuroscience*, *11,* 389–397.

Rushworth, M. F. S., Mars, R. B., & Summerfield, C. (2009). General mechanisms for making decisions? *Current Opinion in Neurobiology*, *19,* 75–83.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275,* 1593–1599.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Systems Technical Journal*, *27,* 379–423.

Sterret, S. G. (2002). Physical models and fundamental laws: Using one piece of the world to tell about another. *Mind & Society*, *5,* 51–66.

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, *88,* 135–170.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Van Veen, V., & Carter, C. S. (2002). The timing of action-related processes in the anterior cingulate cortex. *Journal of Cognitive Neuroscience*, *14,* 593–602.

Walton, M. E., & Mars, R. B. (2007). Probing human and monkey anterior cingulate cortex in variable environments. *Cognitive, Affective, and Behavioral Neuroscience*, *7,* 413–422.

Weisberg, M. (2007). Who is a modeler? *British Journal for Philosophy of Science*, *58,* 207–233.

Yu, A., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, *46,* 681–692.