Model-Based fMRI and Its Application to Reward Learning and Decision Making

JOHN P. O'DOHERTY,[a,b] ALAN HAMPTON,[a] AND HACKJIN KIM[b]

[a] *Computation and Neural Systems Program, California Institute of Technology, Pasadena, California, USA*

[b] *Division of Humanities and Social Sciences, California Institute of Technology, Pasadena, California, USA*

ABSTRACT: In model-based functional magnetic resonance imaging (fMRI), signals derived from a computational model for a specific cognitive process are correlated against fMRI data from subjects performing a relevant task to determine brain regions showing a response profile consistent with that model. A key advantage of this technique over more conventional neuroimaging approaches is that model-based fMRI can provide insights into how a particular cognitive process is implemented in a specific brain area as opposed to merely identifying where a particular process is located. This review will briefly summarize the approach of model-based fMRI, with reference to the field of reward learning and decision making, where computational models have been used to probe the neural mechanisms underlying learning of reward associations, modifying action choice to obtain reward, as well as in encoding expected value signals that reflect the abstract structure of a decision problem. Finally, some of the limitations of this approach will be discussed.

KEYWORDS: computational models; neuroimaging; prediction error; expected value; conditioning; striatum; ventromedial prefrontal cortex

INTRODUCTION

In this review, we will highlight a recent development in neuroimaging research, the "model-based" approach, which involves the application of computational models in the design and analysis of neuroimaging experiments, particularly those involving functional magnetic resonance imaging (fMRI). We will outline the advantages of this method over more traditional neuroimaging approaches as a means of advancing psychological or computational theories

Address for correspondence: John P. O'Doherty, California Institute of Technology, M/C 228-77, 1200 E. California Boulevard, Pasadena, CA 91125, USA. Voice: +1-626-395-5981; fax: +1-626-793-8580.
jdoherty@hss.caltech.edu

Ann. N.Y. Acad. Sci. 1104: 35–53 (2007).
doi: 10.1196/annals.1390.022

of brain function. It will be argued that the model-based approach provides a means of using neuroimaging data to discriminate between existing theories of brain function as well as to advance new novel theories, in a way that could not be achieved with more conventional neuroimaging techniques or in traditional behavioral studies. This argument will be supported with reference to the field of reward learning and decision making, where the model-based approach has up to now been most profitably employed.

MODEL-BASED fMRI

The central approach behind model-based imaging is that one needs to start out with a quantitative computational model that describes a mapping or transformation between a set of stimulus inputs, and a set of behavioral responses. The specific "internal" operations required to effect such a transformation are the variables of interest in the neuroimaging study, as it is these variables that will ultimately be correlated with the neuroimaging data. Perhaps one of the simplest examples of quantitative models in psychology, are those developed to account for the acquisition of conditioned responses during Pavlovian conditioning, the most widely cited example being the Rescorla–Wagner (RW) model.[1] In the RW model, the process by which a conditioned stimulus (CS) comes to produce a conditioned response, is represented by two scalar variables: v, which is the strength of the conditioned response elicited, or more abstractly the value of the CS, and u, which is the value of the unconditioned stimulus (UCS). For conditioning to occur, through repeated contingent presentations of the CS and US, the variable v (which may be initially zero) should converge toward the value of u. At the core of the RW model is that this convergence is accomplished by means of a prediction error δ, which is the difference between the current value of u and v, on each conditioning trial. The value of v is then updated in proportion to δ, so that over the course of learning, v eventually converges to u, δ tends to zero, and learning is complete. Here, in this simple model there are three variables of interest, v, δ, and u. In traditional behavioral psychology, the only directly observable variable here is v, because it can be measured by assessing the strength of the conditioned response in a human or an animal undergoing conditioning. The other variables, such as δ, while not directly observable can be inferred indirectly from the behavioral data, by for example, analyzing the behavioral learning curves.[2] However, the ability to directly measure such variables as they are implemented in the brain, is arguably a much more compelling approach in establishing the validity of such a model. In other words, the neural data become a rich source of evidence, which in addition to the behavior, can be used to constrain the computational models.[3]

Once one moves beyond simple models, such as the RW model, to models accounting for more complex cognitive or behavioral phenomena with many

internal variables, then the ability to validate such models on the basis of behavioral observation alone becomes an ever more difficult proposition as the more complex the model (and hence the more associated free parameters), the more unconstrained the behavioral fitting becomes. In this event, the use of neurophysiology or neuroimaging techniques to measure such internal variables and thus impose additional constraints, becomes even more critical. In the latter part of this review, we will give a specific example of how neuroimaging data can be used to impose just such constraints and permit discrimination between competing models.

A RECIPE FOR MODEL-BASED fMRI

The standard approach in model-based MRI is to first fit the computational model to subjects' actual behavior to find specific values for the free parameters in the model, which minimize the difference between the model predictions and the behavioral data. In the case of the RW model example outlined previously, this would involve finding a value for the one free parameter in the simplest variant of this model, the learning rate α, which is used to scale the updates to v in proportion to the prediction error δ on each trial t: ($v_t = v_{t-1} + \alpha\delta$), such that the trial by trial values for v are as close as possible to the actual behavioral conditioned responses elicited in a particular subject or group of subjects. Once the best-fitting model parameters have been found, then the different model components can be regressed against the fMRI data. For the RW model, there are two variables that change on a trial by trial basis as a function of learning: v and δ. Both of these variables can be entered into the fMRI analysis as a time series and convolved with a canonical hemodynamic response function or other basis function(s), to capture the effects of hemodynamic lag, as is standard in many fMRI analysis approaches. These signals are then regressed against the fMRI data using a general linear model,[4] to identify areas where the model-predicted time series show significant correlations with the actual changes in blood oxygenation level-dependent (BOLD) signal over time. While the RW model only accounts for signal changes on a trial by trial basis, model time series predictions can be much more fine-grained, capturing dynamic changes in activity within as well as between trials, on a second by second basis. A schematic of the model-based fMRI procedure is provided in FIGURE 1.

MODEL VALIDATION AND COMPARISON

Establishing correlations between a given model and a time series of fMRI activity in a given region does not of course demonstrate conclusively that this region is implementing that specific model. It is possible that some aspects

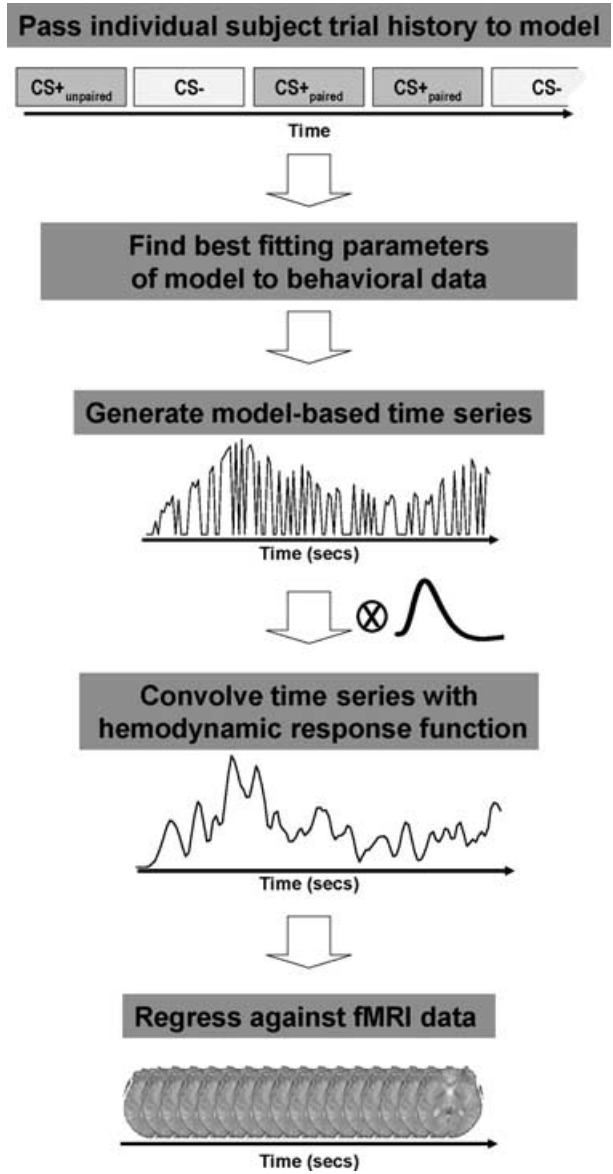

FIGURE 1. Illustration of model-based fMRI approach. Each individual subject's trial history is passed to the model, and the parameters of the model are fit so as to minimize the difference between the model predictions and an external behavioral measure, which in the conditioning example could be an external measure of conditioning, such as galvanic skin conductance responses or pupil dilation. Next, the best model-fitting parameters are used to generate a time series for each trial in the fMRI, which are then convolved with basis function(s) to account for the effects of hemodynamic lag, such as the canonical hemodynamic response function, and then regressed against the fMRI data.

of the time series data correlate well with the model while other parts of the data do not, yet a significant correlation still prevails when the model is fitted to the fMRI data overall. To address this possibility it is useful to plot the model predictions against the observed data to evaluate whether there are any systematic deviations of the model against the data. If evidence is found for systematic deviations of the data from the model predictions (i.e., structure in the residuals), this could indicate that the computational model needs to be modified to account better for the neural activity in a specific area.

A more principled approach is to compare the explanatory power of different competing models against both the behavioral and fMRI data. Usually this is done first for the behavioral data, such that the fit of a variety of models may be compared against the behavioral data to select the subset of models that provide the best account of behavior. An important consideration when performing model comparison is that in the trivial case, models with more free parameters will tend to provide a better fit than models with fewer parameters, thus the addition of extra free parameters needs to be penalized appropriately during the fitting procedure to take this into account.[5] However, in the event that two or more competing models might provide qualitatively similar predictions about behavior, such that there is no way to use the behavioral data alone to discriminate them, it is necessary to turn to the fMRI data to impose additional constraints. In this event such models can be entered into a regression analysis against the fMRI data and allowed to compete against each to determine which model provides a better fit to the fMRI data, for a given brain region.

INFERENCE IN MODEL-BASED fMRI

Model-based fMRI exploits the spatial and temporal resolution of event-related fMRI to characterize dynamic changes in neural activity over time in terms of a particular computational process. Moreover, this technique can be used as a means of discriminating between competing computational models of cognitive and neural function. Thus, model-based fMRI provides insight into "how" a particular cognitive function might be implemented in the brain, not only "where" it is implemented. The model-based approach described here can be contrasted with the approach taken in many conventional neuroimaging studies in which a set of brain areas are merely reported to be "activated" in a particular task or condition, which contribute knowledge about spatial localization of specific cognitive processes, but arguably provide little or no insight into how these processes are being implemented in the activated areas.

MODEL-BASED fMRI OF REWARD LEARNING AND DECISION MAKING

Model-based fMRI will now be illustrated with reference to the problem of how humans can learn to make adaptive decisions under uncertainty to

maximize rewards. Decision making in this context can usefully be fractionated into two distinct components, a reward prediction component in which the expected future reward associated with a particular set of stimuli or actions is learned through experience, and an action selection component in which over the course of learning an agent learns to bias its action choice to favor those action(s) that yield the greatest reward.

LEARNING REWARD PREDICTIONS

Reward prediction in its simplest form can be studied through the phenomenon of classical or Pavlovian conditioning, whereby an arbitrary initially affectively neutral or CS takes on reward value through repeated contingent presentation with an appetitive UCS, such as a food reward. Error correcting learning rules, such as the RW model described earlier, have proved to be successful in accounting for much though by no means all behavioral phenomena in classical conditioning. The first evidence for reward prediction error signals in the brain emerged from the work of Wolfram Schultz and colleagues who observed such signals by recording from the phasic activity of dopamine neurons in awake, behaving, nonhuman primates undergoing simple instrumental or classical conditioning tasks.[6–8] The response profile of these neurons does not correspond to a simple RW rule but rather a real-time extension of this rule called temporal difference learning in which predictions of future reward are computed at each discrete time interval t within a trial, such that the error signal is generated by computing the difference in successive predictions.[9,10] This specific model provides a good approximation to the temporal profile of activity of these neurons during classical conditioning in which the dopamine neurons first respond at the time of the UCS before learning is established but shift back in time within a trial to respond instead at the time of presentation of the cue once learning is established. To test for evidence of a temporal difference prediction error signal in the human brain, O'Doherty and colleagues[11] scanned human subjects while they underwent a classical conditioning paradigm in which associations were learned between arbitrary visual fractal stimuli and a pleasant sweet taste reward (glucose). One cue was followed most of the time by the taste reward, whereas another cue was followed most of the time by no reward. However, in addition, subjects were exposed to low frequency "error" trials in which the cue associated with reward was presented but the reward was omitted, and the cue associated with no reward was presented but a reward was unexpectedly delivered. The specific trial history that each subject experienced was next fed into a temporal difference model to generate a time series that specified the model-predicted prediction error signal at three different time points in a trial from the time at which the CS is presented until the time at which the reward is delivered (3 sec later) (FIG. 2A, B).

This time series was then convolved with a canonical hemodynamic response function and regressed against the fMRI data for each individual subject, to identify brain regions correlating with the model-predicted time series. This analysis revealed significant correlations with the model-based predictions in a number of brain regions, most notably the ventral striatum (ventral putamen bilaterally) (FIG. 2C) and orbitofrontal cortex, both prominent target regions of dopamine neurons.[12] These results suggest that prediction error signals are present in the human brain during reward learning, and that these signals conform to a response profile consistent with a specific computational model: temporal difference learning. Another study by McClure and colleagues also revealed activity in ventral striatum consistent with a reward prediction error signal using an event-related trial-based analysis.[13]

ACTION SELECTION FOR REWARD: THE ACTOR/CRITIC IN THE HUMAN BRAIN

While classical conditioning is a useful paradigm for studying the passive learning of reward predictions to gain insight into the process by which humans can learn to perform actions to obtain reward, it is necessary to turn to instrumental conditioning, which involves learning of stimulus-response or stimulus-response-outcome associations. Insight into how humans or other animals might implement instrumental conditioning has come from the application of a family of models collectively known as reinforcement learning, originally developed in computer science.[14] In one such model, the actor/critic,[15,16] action selection is conceived as involving two distinct components: a critic, which learns to predict future reward associated with particular states in the environment, and an actor, which chooses specific actions to move the agent from state to state according to a learned policy. The critic encodes the value of particular states in the world and as such has the characteristics of a Pavlovian reward prediction signal described above. The actor stores a set of probabilities for each action in each state of the world, and chooses actions according to those probabilities. The goal of the model is to modify the policy stored in the actor such that over time, those actions associated with the highest predicted reward are selected more often. This is accomplished by means of a prediction error signal, which computes the difference in predicted reward as the agent moves from state to state. This signal is then used to update value predictions stored in the critic for each state, but also to update action probabilities stored in the actor such that if the agent moves to a state associated with greater reward (and thus generates a positive prediction error), then the probability of choosing that action in future is increased. Conversely, if the agent moves to a state associated with less reward, this generates a negative prediction error and the probability of choosing that action again is decreased.

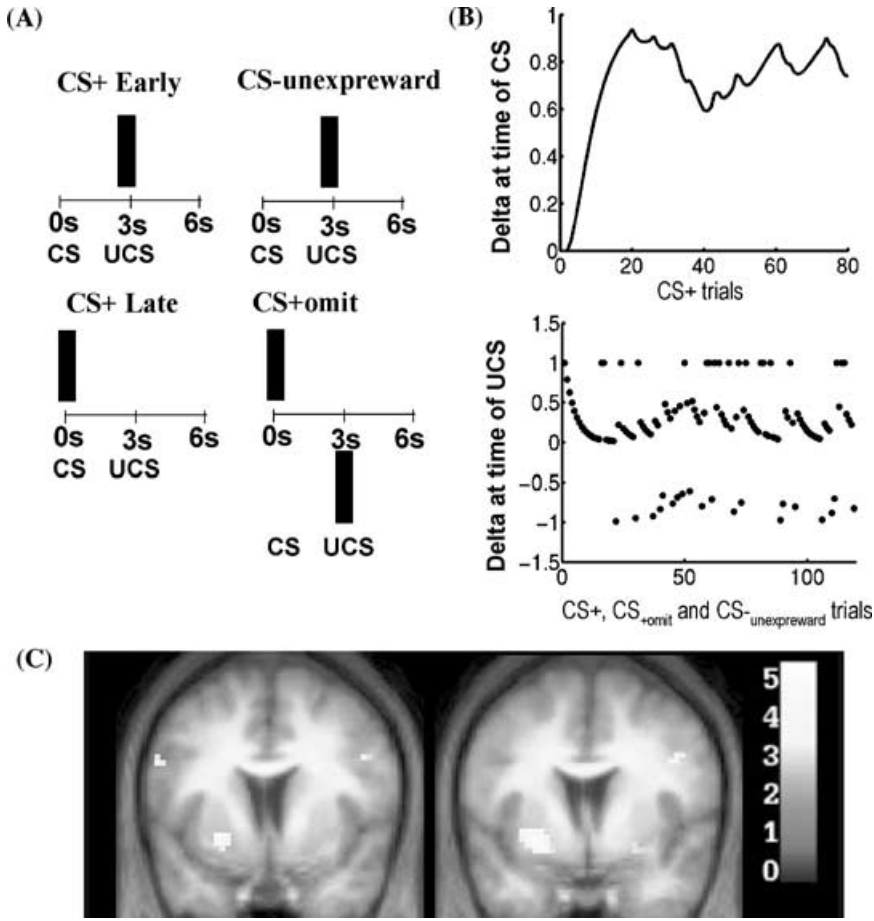

FIGURE 2. Model-based fMRI of stimulus-reward learning. (**A**) Properties of the temporal difference prediction error signal during reward learning in which a cue (CS+) is paired repeatedly with a reward (UCS) presented 3 sec later. During the initial stages of learning (CS + early trials), the error signal responds at the time of presentation of the UCS, but over the course of learning transfers back to the time of presentation of the CS (CS + late trials). On trials in which the CS+ is not presented but the reward is delivered anyway (CS–unexp. reward), the signal shows a positive response at the time the reward is delivered, whereas on trials in which the CS is presented but the reward is unexpectedly omitted the signals show a negative response at the time of outcome. (**B**) Plot of model-generated prediction error signals at the time of presentation of the CS, and the time of presentation of the UCS, over the course of the experiment for a typical subject. (**C**) Area of bilateral ventral striatum (ventral putamen bilaterally) showing significant correlations with the temporal difference prediction error signal while subjects underwent classical conditioning with sweet taste reward (1M glucose). Data from O'Doherty *et al.*[11]

Some computational neuroscientists have drawn analogies between the anatomy and connections of the basal ganglia, and possible neural architectures for implementing reinforcement learning models including the actor/critic. Houk and colleagues[17] have proposed that the actor and critic could be implemented within patch/striosome and matrix compartments distributed throughout the striatum. Montague and colleagues[10] proposed that the ventral and dorsal striatum implemented the critic and actor respectively, on the grounds of extant knowledge of the putative functions of these structures at the time, derived primarily from animal lesion studies. To test these hypotheses, O'Doherty and colleagues[18] scanned hungry human subjects with fMRI while they performed a simple instrumental conditioning task in which they were required to choose one of two actions leading to juice reward with either a high or low probability (FIG. 3A). Neural responses corresponding to the generation of prediction error signals during performance of the instrumental task were compared to that elicited during a control Pavlovian task in which subjects experienced the same stimulus-reward contingencies but did not actively choose which action to select. This comparison was designed to isolate the actor, which was hypothesized to be engaged only in the instrumental task, from the critic, which was hypothesized to be engaged in both the instrumental and Pavlovian control tasks. Consistent with the proposal of a dorsal versus ventral actor/critic architecture, activity in dorsal striatum was found to be specifically correlated with prediction error signals when subjects were actively performing instrumental responses to obtain reward. In contrast, ventral striatum was found to be active in both the instrumental and Pavlovian tasks (FIG. 3B).

These results suggest a dorsal–ventral distinction within the striatum whereby ventral striatum is more concerned with Pavlovian or stimulus-outcome learning, while the dorsal striatum is more engaged during learning of stimulus-response or stimulus-response-outcome associations. The suggestion that human dorsal striatum is specifically involved under situations when subjects need to select actions to obtain reward has received support from a number of other fMRI studies, both model based[19] and trial based.[20]

EXPECTED VALUE

This raises the question as to where in the brain expected values are represented. A number of model-based fMRI studies have consistently implicated ventromedial prefrontal cortex in encoding the value of chosen actions. Kim and colleagues,[21] used a variant of the actor/critic algorithm to generate expected value signals as subjects made decisions between which of two possible actions to chose to obtain monetary reward, as well as in a different condition, to avoid losing money. In this study, different available actions were associated

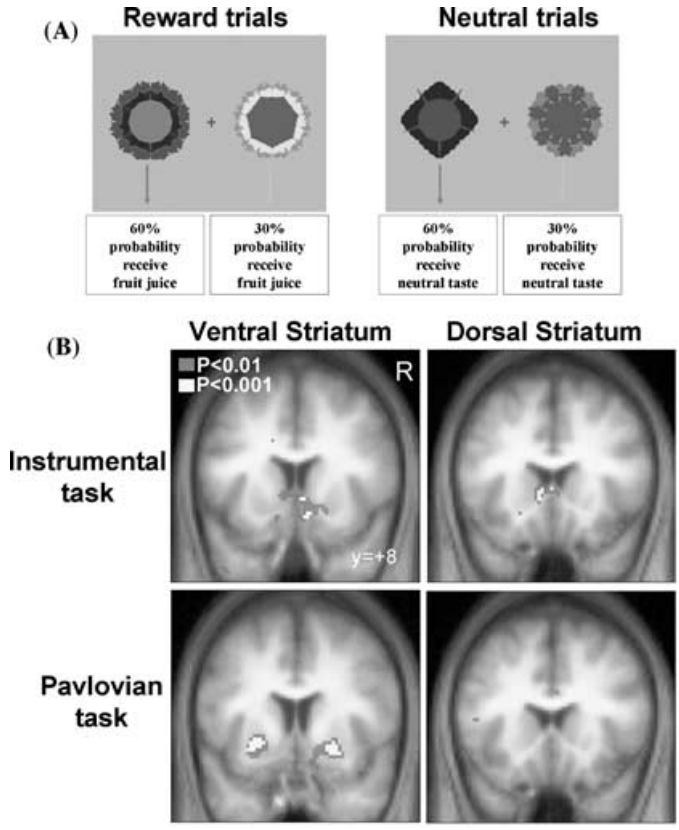

FIGURE 3. Model-based fMRI of action selection for reward. (**A**) Schematic of instrumental choice task used by O'Doherty *et al.*[18] On each trial of the reward condition subject chooses between two possible actions, one associated with a high probability of obtaining juice reward (60%), the other a low probability (30%). In a neutral condition subjects also choose between actions with similar probabilities but in this case they receive an affectively neutral outcome (tasteless solution). Prediction error responses during the reward condition of the instrumental choice task were compared to prediction error signals during a yoked Pavlovian control task. (**B**) Significant correlations with the reward prediction error signal generated by an actor/critic model were found in ventral striatum (ventral putamen extending into nucleus accumbens proper) in both the Pavlovian and instrumental tasks, suggesting that this region is involved in stimulus-outcome learning. In contrast, a region of dorsal striatum (anteromedial caudate nucleus) was found to be correlated with prediction error signals only during the instrumental task, suggesting that this area is involved in stimulus-response or stimulus-response-outcome learning. Data from O'Doherty *et al.*[18]

with distinct probabilities of either winning or losing money, such that in the reward condition one action was associated with a 60% probability of winning money, and the other action with only a 30% probability of winning. To maximize their cumulative reward, subjects should learn to choose the 60%

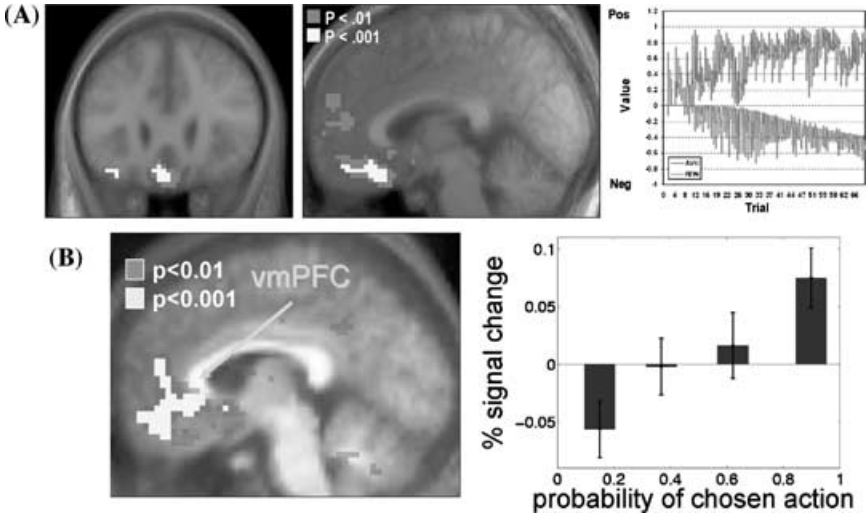

FIGURE 4. Expected value signals in ventromedial prefrontal cortex. (**A**) Regions of ventromedial prefrontal cortex (medial and central orbitofrontal cortex extending into medial prefrontal cortex) correlating with expected value signals generated by a variant of the actor/critic model during an fMRI study of instrumental choice of reward and avoidance (left and middle panels). The model-predicted expected value signals are shown for one subject in the right panel for both the reward (top line) and avoidance (bottom line) conditions. Data from Kim *et al.*[21] (**B**) Similar regions of ventromedial prefrontal cortex correlating with model-predicted expected value signals during performance of a four-armed bandit task with nonstationary reward distributions (left panel). BOLD signal changes in this region are shown plotted against model predictions (right panel), revealing an approximately linear relationship between expected value and BOLD signal changes in this region. Data from Daw *et al.*[22] The figure appears in color online.

probability action. In the avoidance condition, subjects were presented with a choice between the same probabilities, except in this context, 60% of the time after choosing one action they avoided losing money, whereas this only occurred 30% of the time after choosing the alternate action. To minimize their losses, subjects should learn to choose the action associated with the 60% probability of loss avoidance. Model-generated expected value signals for the action chosen were found to be correlated on a trial-by-trial basis with BOLD responses in bilateral orbitofrontal cortex and adjacent medial prefrontal cortex in both the reward and avoidance conditions, such that activity in these areas increased in proportion to the expected future reward associated with the specific action chosen (FIG. 4A).

Similar results were obtained by Daw and colleagues,[22] who used a four-armed bandit task in which "points" (that would later be converted into money) were paid out on each bandit. However, unlike the studies described previously,

in this case, the mean payoff available on each bandit drifted over time, such that no one bandit was consistently paying out more than the others, but at any one time some bandits paid out more than the others. As a consequence, subjects had to keep track of the mean payoffs available on each bandit over the course of the experiment, to work out which one paid out the most at any one moment. Model-predicted expected value signals specific to the action chosen were found to be correlated on a trial by trial basis with fMRI activity in ventromedial prefrontal cortex (FIG. 4B). Tanaka and colleagues also showed significant correlations between value signals derived from reinforcement learning models and activity in medial prefrontal cortex during an fMRI study where subjects performed a Markov decision task involving choices to obtain immediate or delayed rewards.[23]

Taken together, these findings suggest that orbital and medial prefrontal cortex are involved in keeping track of the expected future reward associated with chosen actions, and that these areas show a response profile consistent with an expected value signal generated by reinforcement learning models. The fact that these regions were found to be correlated with the "chosen" action, raises the question as to how the chosen action comes to be chosen in the first place. Presumably somewhere in the brain it is necessary to encode the value of each available action before a specific action is chosen to compare between them and ultimately select a particular action. In future studies it will be important to establish the presence of prechoice action values to distinguish between regions actively computing the decision itself from those merely reporting its consequences. Another open question is whether state values, corresponding to the average reward available in a particular state (usually signaled by a specific combination of stimuli) irrespective of the actions subsequently chosen, are represented separately from action values, which correspond to the future reward associated with selection of a specific action. So far, neuroimaging studies have failed to distinguish between state and action values, most probably because such signals are usually highly correlated with each other in the experimental designs that have been used to date.

ENCODING ABSTRACT RULES IN A DECISION PROBLEM

Although standard RL models can account for a wide range of human and animal choice behavior, these models do have important limitations. One such limitation is a failure to account for higher order structure in a decision problem, such as rules describing interdependencies between actions, rewards, and other exogenous variables, such as time. Yet, many real-life decision problems do incorporate such structure.[24–26] Simple RL models assume the actions available in the world and the rewards that can be obtained from choice of those actions are independent from each other, such that in a given state, information gained about the rewards available from choice of one action

provides no information about the rewards available from choice of another action. However, in many situations rules describing interdependencies between different actions do exist, and if subjects can exploit these rules, this will lead to greater reward than otherwise. One of the simplest examples of a decision task with such an abstract rule is probabilistic reversal learning.[25,26] In this task, subjects can choose between one of two actions, which give out monetary rewards or losses on a probabilistic basis, similar to the instrumental choice tasks outlined previously. However, in this case, at any one time, one of the actions pays out more reward than the other, such that if the subject continues to choose the high paying action they will obtain the greatest reward. After a time the contingencies reverse, and the subject has to switch their choice of action to continue to maximize their reward. The structure in this task is the anticorrelation between the distribution of rewards available on the two actions: when one action is "good" the other is "bad" and *vice versa*, as well as the rule that after a time the contingencies will reverse. A standard RL model could be used to learn such a task, but in this case the model would not incorporate the abstract rules in the reversal task but would instead simply learn about the values of the two actions independently.

However, what if a model were to incorporate such rules? To establish whether human subjects do use an abstract representation of the task structure to guide their choices, Hampton and colleagues[27] scanned subjects with fMRI while they underwent probabilistic reversal learning. A computational model that incorporated the structure of the task was then constructed and fitted to both the behavioral and fMRI data. This structure-based model was implemented as a hidden Markov model (HMM), where the hidden state to be estimated was the probability that on a given trial the "correct" (or currently high reward value action) is being chosen. This model also incorporated the fact that the identity of the correct action would reverse from time to time. The structure-based model was found to provide a good fit to subjects' behavior, and the signal derived from the model representing the probability that subjects were choosing the correct action (prior correct) was found to be significantly correlated at the time of action choice with fMRI activity in ventromedial prefrontal cortex (FIG. 5A). The areas thus identified overlap markedly with those regions found to correlate with expected value derived from standard RL models outlined previously. This is perhaps not surprising as the prior correct signal from the HMM model is strongly colinear with expected value signals derived from RL.

MODEL COMPARISON: STRUCTURE-BASED INFERENCE VERSUS STANDARD RL

However, the key question is whether the model that incorporates the rules of the decision task can account *better* for subjects' behavioral and fMRI data

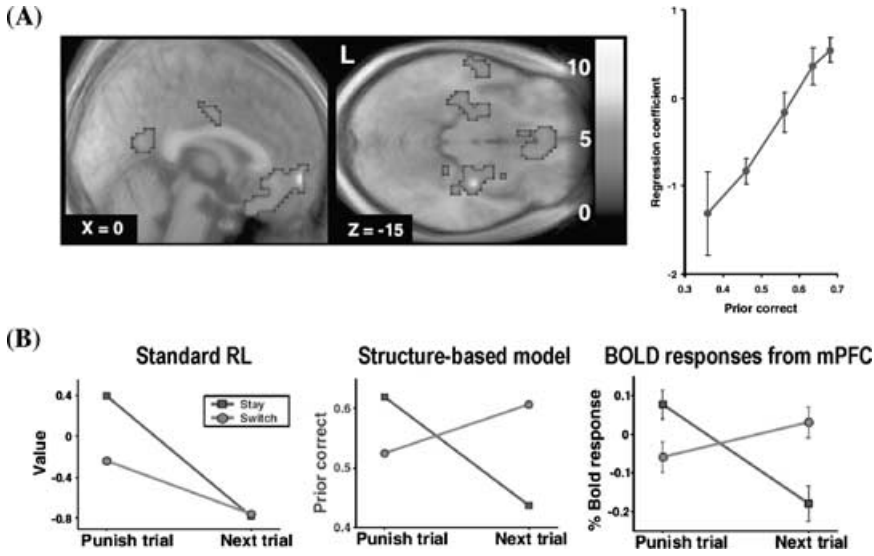

FIGURE 5. Expected value signals in ventromedial prefrontal cortex reflect abstract task structure. (**A**) Regions of ventromedial prefrontal cortex correlating with expected value signals while subjects performed a decision task with abstract structure: probabilistic reversal learning (left panel). Expected signals were generated by a model that incorporates the structure of the task. Plot of responses in this region against expected value signals reveal a strongly linear relationship (similar to that shown in FIG. 4B). (**B**) Dissociating a model that incorporates the structure of the task (structure-based model) from standard RL models that do not incorporate the rules of the task. The predictions from standard RL and the structure-based model are plotted for expected value signals before and after switching action choice following a reversal, where the subject switches back to an action that was chosen previously (shown in left and middle panels as the circled points). The structure-based model predicts that expected value signals jump up following a reversal, whereas the RL models predict no such increase in reward expectation. The actual BOLD signal in medial PFC shows a response profile consistent with the structure-based model and not simple RL (right panel). Data from Hampton *et al.*[27] Copyright 2006 Society for Neuroscience. The figure appears in color online.

than does standard RL, as this would provide evidence that human subjects *do* use knowledge of task structure to guide their choices as opposed to merely learning action values independently. In this study we therefore compared the goodness-of-fit of the HMM model to subjects' behavioral data to the fit achieved by a family of different RL models. The structure-based model was found to provide a significantly better fit to the behavioral data than did the best-fitting RL model, even after adjusting for the number of free parameters in the model. Thus, even at the behavioral level, there is evidence to suggest that subjects are using knowledge of the structure of the reversal task to guide their choices.

To determine whether neural coding of expected values also reflected knowledge of this structure, we looked specifically at those times in the experiment when the predictions of the structure-based model and of the standard RL model would be maximally divergent. This happens to be immediately after subjects switch their choice of action. According to both RL and the structure-based model, subjects should switch their choice of action once they have a low expected value for that action, presumably after having received a string of nonrewarding outcomes after selecting that action on previous trials. Where the models diverge, is when subjects switch back to an action that they previously switched away from after a prior contingency reversal. According to standard RL, once subjects switch back to an action they should still have a low expected reward for that action because the last time they chose that action it had a low expected value (hence they switched away from it). However, according to the structure-based model, once subjects switch back to an action they previously switched away from, this time they should have a high expected reward value, because they understand that the contingencies have reversed, and that therefore this action must now have a high reward value. FIGURE 5B shows the predictions of both the structure-based model and standard RL models alongside the actual signal at the time of choice extracted from the medial prefrontal cortex. As can be seen from this figure the actual fMRI data mirror the predictions of the structure-based model, by showing that subjects' expectations of reward jumps up once they switch back to a previously chosen action following a reversal. In a further analysis, model predictions from the structure-based model and standard RL model were both entered into a regression analysis against the fMRI data and the fit of both models to the fMRI data was directly compared. Once again, the structure-based model was found to provide a better fit to the fMRI data than the best-fitting RL model.

These results suggest that subjects can take into account the abstract rules of a decision problem and that knowledge of such rules can modulate expected value signals in prefrontal cortex that in turn may be used to guide behavioral choice. These results do not invalidate RL models *per se*, but rather highlight the need to incorporate additional inference mechanisms that may exist alongside standard RL either cooperatively or competitively.[28] Future studies will be needed to establish whether both structure-based and standard RL signals are present in the brain at the same time, and if so, whether different regions of the brain are involved in implementing these different learning processes. Alternatively, known rules of the task could be incorporated into the structure of the state space in a given task, and then standard RL mechanisms could be used to learn the values in this state space from then on. More generally, this particular study provides an example of how fMRI data can be used to inform about the specific computational functions being implemented in a particular brain region, in this case ventromedial prefrontal cortex. By comparing the degree to which activity in this region can be accounted for by different competing computational models, it was possible to show that signals in this region are

accounted better by one model than another. Thus, fMRI has provided a means to discriminate between different computational models of brain function.

LIMITATIONS OF MODEL-BASED IMAGING

In this review we have argued that model-based fMRI provides a powerful means with which to test computational models of brain function, and argue that this approach overcomes many of the limitations of more conventional approaches to neuroimaging in which a set of regions are merely identified as being "active" in a given task or condition. Unlike these previous approaches that permit inferences about where a given cognitive function is implemented, model-based fMRI provides insights not only into where but also as to how a specific function might be carried out.

However, it should be noted that this technique does have many limitations. Use of the model-based approach does entail testing for highly specific signals in the brain, which permits testing of specific computational hypotheses, but at the same time this approach may limit the possibility of detecting unexpected findings that would not conform to a specific *a priori* hypothesis. Consequently, it is probably a reasonable policy to continue to employ more conventional trial-based analyses alongside the model-based approach. A complementary model-free approach that has been used to explore relationships between objective stimulus events and behavioral and neural data is to examine correlations between current behavior or neural activity on a given trial against previous experience, such as, for example, regressing current choice against the history of rewarding outcomes received. By analyzing how correlations between these variables change over time, it is in principle possible to extract information about the type of computational process being used by the subject to drive behavior as a function of experience, without imposing a specific computational model at the outset. For example, by using this approach Sugrue *et al.*, were able to gain insight into the manner by which reward information was integrated over time to drive performance on a matching task.[29] A similar approach could in principle be adopted in human imaging or behavioral studies, with the caveat that while this approach has proven useful in monkey neurophysiology studies where thousands of trials are typically available for analysis, it is less clear whether this technique will have sufficient statistical power in human studies that usually involve far fewer trials.

Furthermore, model-based fMRI just as with any other direct or indirect technique for recording neural activity, can only provide insight into correlations between neural activity and behavior but cannot establish causal links between activity in a particular brain area and subsequent behavior. For this, in humans at least, it is necessary to study the effects of disruption to a region either by means of assessing the effects of a permanent lesion to this area, or by inducing temporary inactivation in an area through transcranial magnetic stimulation (TMS). The model-based approach can of course be applied to

lesion and TMS data just as it can be applied to fMRI data. Combining the results of model-based analyses of fMRI data with data derived from these other approaches will likely be an important direction for future research.

Another limitation of model-based fMRI as with any other use of fMRI data, is the poor spatiotemporal resolution of this technique compared to single or multiunit neurophysiology recording techniques available in rats and nonhuman primates. Computational signals observed with fMRI are likely to reflect cases where a large population of neurons essentially conveys the same signal, as is thought to be the case, for example, in the phasic activity of dopamine neurons. Naturally, model-based fMRI cannot provide insight into more fine-grained computational signals conveyed at the single neuron level, especially where different interleaved populations of neurons within a region may be conveying distinct computational signals. Model-based fMRI is only useful under situations where computational signals are conveyed by means of a change in mean firing rate in a population of neurons such that variation in the computational signal is reflected in variation in metabolic demand and hence blood oxygenation within a particular region. For some types of computational processes, changes in computations within a region may not result in a change in overall activity level, but rather may reflect a change in a distributed pattern of neural activity without any overall change in mean firing rates. Such signals would not be detectable with the model-based fMRI approach.

On account of these limitations, more so than ever perhaps, it will be necessary to combine the results from model-based fMRI studies in humans with other techniques, such as single- or multiunit neurophysiology in other animals, as well as with imaging methods in humans that afford greater temporal resolution at the expense of spatial precision, such as MEG or EEG. Just as the case has been made here for model-based approaches to fMRI, the case can be made for model-based approaches in these other methodologies. Indeed, it is important to note that the model-based approach to fMRI data parallels a similar approach that is increasingly being adopted in the animal neurophysiology literature, whereby the activity of single neurons are correlated against specific computational models.[24,30–33] As a consequence, the tendency toward incorporating computational models into fMRI studies of neural activity may be seen as part of a wider trend in the field toward a more quantitative and theory-driven approach to experimental neuroscience.

ACKNOWLEDGMENTS

This work was funded by grants from the Gimbel Discovery Fund for Neuroscience, the Gordon and Betty Moore Foundation, and a Searle Scholarship to JOD. We would like to thank Nathaniel Daw, Peter Dayan, Ray Dolan, Karl Friston, and Ben Seymour at UCL, and Peter Bossaerts and Shin Shimojo at Caltech, who were major collaborators on some of the research studies described here.

REFERENCES

1. RESCORLA, R.A. & A.R. WAGNER. 1972. A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. *In* Classical Conditioning II: Current Research and Theory. A.H. Black & W.F. Prokasy, Eds.: 64–99. Appleton Crofts. New York.
2. RESCORLA, R.A. 2002. Comparison of the rates of associative change during acquisition and extinction. J. Exp. Psychol. Anim. Behav. Process. **28:** 406–415.
3. HENSON, R. 2005. What can functional neuroimaging tell the experimental psychologist? Q. J. Exp. Psychol. A **58:** 193–233.
4. FRISTON, K.J. *et al*. 1995. Statistical parametric maps in functional imaging: a general linear approach. Hum. Brain Map. **2:** 189–210.
5. SCHWARZ, G. 1978. Estimating the dimension of a model. Ann. Stat. **6:** 461–464.
6. SCHULTZ, W. 1998. Predictive reward signal of dopamine neurons. J. Neurophysiol. **80:** 1–27.
7. MIRENOWICZ, J. & W. SCHULTZ. 1994. Importance of unpredictability for reward responses in primate dopamine neurons. J. Neurophysiol. **72:** 1024–1027.
8. HOLLERMAN, J.R. & W. SCHULTZ. 1998. Dopamine neurons report an error in the temporal prediction of reward during learning. Nat. Neurosci. **1:** 304–309.
9. SCHULTZ, W., P. DAYAN & P.R. MONTAGUE. 1997. A neural substrate of prediction and reward. Science **275:** 1593–1599.
10. MONTAGUE, P.R., P. DAYAN & T.J. SEJNOWSKI. 1996. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J. Neurosci. **16:** 1936–1947.
11. O'DOHERTY, J.P. *et al*. 2003. Temporal difference models and reward-related learning in the human brain. Neuron **38:** 329–337.
12. JOEL, D. & I. WEINER. 2000. The connections of the dopaminergic system with the striatum in rats and primates: an analysis with respect to the functional and compartmental organization of the striatum. Neuroscience **96:** 451–474.
13. MCCLURE, S.M., G.S. BERNS & P.R. MONTAGUE. 2003. Temporal prediction errors in a passive learning task activate human striatum. Neuron **38:** 339–346.
14. SUTTON, R.S. & A.G. BARTO. 1998. Reinforcement Learning. MIT Press. Cambridge, MA.
15. BARTO, A.G. 1992. Reinforcement learning and adaptive critic methods. *In* Handbook of Intelligent Control: Neural, Fuzzy, and Adaptive Approaches. D.A. White & D.A. Sofge, Eds.: 469–491. Van Norstrand Reinhold. New York.
16. BARTO, A.G. 1995. Adaptive critics and the basal ganglia. *In* Models of Information Processing in the Basal Ganglia. J.C. Houk, J.L. Davis & B.G. Beiser, Eds.: 215–232. MIT Press. Cambridge, MA.
17. HOUK, J.C., J.L. ADAMS & A.G. BARTO. 1995. A model of how the basal ganglia generate and use neural signals that predict reinforcement. *In* Models of Information Processing in the Basal Ganglia. J.C. Houk, J.L. Davis & B.G. Beiser, Eds.: 249–270. MIT Press. Cambridge.
18. O'DOHERTY, J. *et al*. 2004. Dissociable roles of ventral and dorsal striatum in instrumental conditioning. Science **304:** 452–454.
19. HARUNO, M. *et al*. 2004. A neural correlate of reward-based behavioral learning in caudate nucleus: a functional magnetic resonance imaging study of a stochastic decision task. J. Neurosci. **24:** 1660–1665.
20. TRICOMI, E.M., M.R. DELGADO & J.A. FIEZ. 2004. Modulation of caudate activity by action contingency. Neuron **41:** 281–292.

21. KIM, H., S. SHIMOJO & J.P. O'DOHERTY. 2006. Is avoiding an aversive outcome rewarding? Neural substrates of avoidance learning in the human brain. PLoS Biol. **4:** e233.
22. DAW, N.D. *et al*. 2006. Cortical substrates for exploratory decisions in humans. Nature **441:** 876–879.
23. TANAKA, S.C. *et al*. 2004. Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. Nat. Neurosci. **7:** 887–893.
24. SUGRUE, L.P., G.S. CORRADO & W.T. NEWSOME. 2004. Matching behavior and the representation of value in the parietal cortex. Science **304:** 1782–1787.
25. O'DOHERTY, J. *et al*. 2003. Dissociating valence of outcome from behavioral control in human orbital and ventral prefrontal cortices. J. Neurosci. **23:** 7931–7939.
26. COOLS, R. *et al*. 2002. Defining the neural mechanisms of probabilistic reversal learning using event-related functional magnetic resonance imaging. J. Neurosci. **22:** 4563–4567.
27. HAMPTON, A.N., P. BOSSAERTS & J.P. O'DOHERTY. 2006. The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. J. Neurosci. **26:** 8360–8367.
28. DAW, N.D., Y. NIV & P. DAYAN. 2005. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat. Neurosci. **8:** 1704–1711.
29. SUGRUE, L.P., G.S. CORRADO & W.T. NEWSOME. 2005. Choosing the greater of two goods: neural currencies for valuation and decision making. Nat. Rev. Neurosci. **6:** 363–375.
30. PLATT, M.L. & P.W. GLIMCHER. 1999. Neural correlates of decision variables in parietal cortex. Nature **400:** 233–238.
31. BARRACLOUGH, D.J., M.L. CONROY & D. LEE. 2004. Prefrontal cortex and decision making in a mixed-strategy game. Nat. Neurosci. **7:** 404–410.
32. SAMEJIMA, K. *et al*. 2005. Representation of action-specific reward values in the striatum. Science **310:** 1337–1340.
33. DAW, N.D. & K. DOYA. 2006. The computational neurobiology of learning and reward. Curr. Opin. Neurobiol. **16:** 199–204.